Springer Optimization and Its Applications 116

Aldo Frediani
Bijan Mohammadi
Olivier Pironneau
Vittorio Cipolla *Editors*

Variational Analysis and Aerospace Engineering

Mathematical Challenges for the Aerospace of the Future

Springer

Springer Optimization and Its Applications

VOLUME 116

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics, and other sciences.

The series *Springer Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository work that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

Aldo Frediani • Bijan Mohammadi
Olivier Pironneau • Vittorio Cipolla
Editors

Variational Analysis and Aerospace Engineering

Mathematical Challenges for the Aerospace of the Future

Springer

Editors
Aldo Frediani
Department of Civil and Industrial Engineering
University of Pisa
Pisa, Italy

Olivier Pironneau
LJLL-UPMC Boite
Paris, France

Bijan Mohammadi
Montpellier Institute Alexander Grothen
University of Montpellier
Montpellier, France

Vittorio Cipolla
Department of Industrial and Civil Engineering
University of Pisa
Aerospace Section
Pisa, Italy

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-3-319-45679-9 ISBN 978-3-319-45680-5 (eBook)
DOI 10.1007/978-3-319-45680-5

Library of Congress Control Number: 2016959447

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

This volume is dedicated to the memory of Piero Villaggio on behalf of the many who have met him during the early stages of their scientific formation and wish to express the gratitude for the example and the message inspired by him.

Piero Villaggio was born in Genova on December 31, 1932. After immersion in humanistic studies at Liceo Doria, Villaggio completed his degree in Civil Engineering in 1957 with a thesis on the mechanics of fluids. His academic career moved rapidly: he became an assistant professor at the University of Genova in 1959 and was appointed as full professor of Scienza delle Costruzioni at the University of Pisa in 1966, where he became a professor emeritus in 2008. Villaggio went to become an Ordinary Member of the Italian Accademia dei Lincei.

Villaggio was a distinguished member in the international scientific landscape and collaborated with numerous eminent figures, particularly scientists from the Cllifford Truesdell School of Continuum Mechanics.

He held positions of visiting professor in prestigious institutions (Johns Hopkins University, Heriot-Watt University, University of Minnesota) and served on the editorial board of various journals. In private, he cultivated intense passions for mountains, music, history and philosophy. He was an expert climber, a member of the academic section of Club Alpino Italiano and author of technical reports which included two papers on the mechanics of climbing [122, 123]. The picture in Fig. 2 portrays Villaggio as a young and vital man. Villaggio died unexpectedly on January 4, 2014, while he was still active in the scientific community.

As many outstanding figures, Piero Villaggio's education was a complex process. Certainly, his work with Guido Stampacchia, in Genova, had a substantial impact and Villaggio had a sincere sense of esteem and gratitude toward Stampacchia. Stampacchia introduced Villaggio to the mathematical aspects of the Calculus of Variations and to the weak theory of the differential boundary value problems.

At Istituto Nazionale di Alta Matematica (INDAM) in Rome Villaggio studied under G. Krall, G. Fichera, C. Cattaneo and B. Segre, and was given the opportunity to meet A. Signorini and M. Picone. While attending INDAM, Villaggio studied the classics of early twentieth century mechanics and modeling problems, such as the theory of contact by Hertz; the plane elasticity problems through the complex variable technique developed by Muskhelishvili and the stress concentration around holes and notches problem solved by Neuber.

In Pisa, Villaggio worked with Signorini's former students T. Manacorda and G. Capriz, together they reestablished mechanics begun by Truesdell, and formed a continuum mechanics group.

Villaggio completed 149 articles, two monographs and a critique of the Johann I and Nicolaus II Bernoulli's works.

Calculus of variations and the weak formulation of the elliptic boundary value problems continued to grow rapidly at the Pisan school of mathematics. Villaggio realized the importance of foundations of these ideas and played a major role in expanding them into engineering. Villaggio used calculus of variations in his early works [1, 8, 9, 14] and in many papers such as the a priori estimates in elastodynamics [42]; the extension to finite deformations of the classical estimates of linear elasticity [32]; some maximum modulus theorems for elastic halfspaces [48]; isoperimetric distributions of loads in elastostatics [57] as well as his articles on the extension of Friedrich's method of the two-side energy bounds to the unilateral problems [39, 43].

Piero Villaggio paid special attention to optimization. Villaggio studied various optimization problems and found the optimal shape of an indenter [68, 70]; the optimal interface between an elastic and a rigid halfspace glued together [67]; the optimal shape of an elastic plate loaded on a part (known) of its boundary and free on the remaining part (unknown) [81, 88]; the problem of optimal packaging [82]; an optimal structural problem for a beam [96]. Villaggio's interest for optimization originated from works by Prager and Taylor at the end of 1960s. Toward the end of the 1970s researchers saw that existence, uniqueness and regularity issues in known solutions. Fixing the matter required looking at weak convergences, relaxation of functionals, etc. Villaggio appeared to be aware of these issues promptly [49] and in a paper with W. Velte [54] illustrated the pathologies of the optimization problems in the elementary case of a bar under axial loads.

Villiaggio also examined a long series of articles which include contact and detachment of bodies, see e.g. [60, 74] and [59, 108], along with fracture and impact. Villaggio proposed a model for a fragile fracture in compression in a joint paper with J. Dunwoody [77]; the case of fracture with curved fracture lines was studied with R. Ballarini in [117]. Impact is the topic of a few articles with R. Knops [97, 110, 140] and the historic articles [109, 139] are from a plenary lectures given in Genova (2003) and Erice (2010).

The theory of complex variables and its application to planar elasticity distinguished Villaggio's research. Villaggio dedicated a brilliant paper at the onset of his career [4] and two papers with M. Leitman [115, 142] on an erroneous application of the analytic continuation technique that is present even in works of authors like Neuber. Villaggio and Leitman illustrated the nature of the error and indicate how it can be removed. Together they provided a solution to the problem of a disk loaded at the external boundary under various loading conditions, as an application. In the application of the complex variable technique, Villagio wrote a paper with Knops [138], where Neuber's errors to the solutions in plane elasticity problems are emended.

Apart from the scientific value of his work, a major merit of Villaggio has been his work between mathematics and engineering. He made the engineering faculty aware of the technical aspects, built on the work of the classics and on their extraordinary capability of simplifying problems. In a passage from [A2] he explicitly quotes the names of Hertz and Kelvin about this matter. As a personal contribution, Villaggio called attention to qualitative mathematics focussed on the search for the general properties of solutions to boundary value problems and on the role of the a priori estimates. Which contributed to create a generation of engineers, able to read technical works and to use them to face novel problems. As a mechanician, Villaggio was open to aspects of generality and clarity that inspired the re-founded continuum mechanics.

Villaggio was an attentive observer of the scientific community. In a paper on mechanics covering the past 60 years [137] Villaggio pointed out the risks of present mechanics: the excess of specialization and the abundance of scientific journals, that deal with “unlikely problems” by an ever growing plethora of contributors. Villaggio called to simplicity which has been the distinctive trait of his legacy.

Simplicity inspired his lectures, that were sharp, concise and elegantly illustrated on the blackboard. He loved lecturing and we remember him teaching, as in Fig. 1.

Udine, Italy Cesare Davini

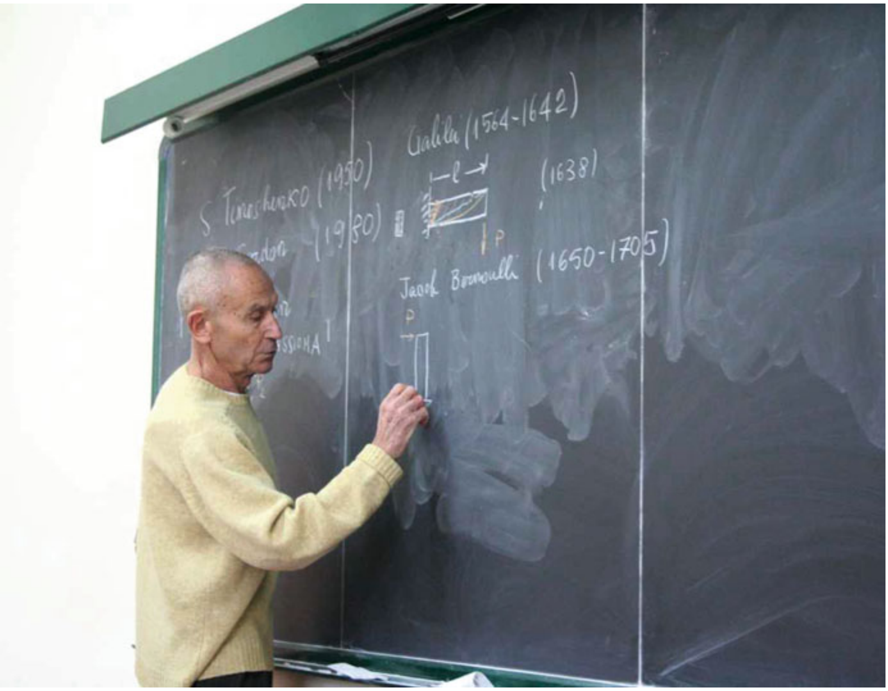


Fig. 1 Piero Villaggio, Università Mediterranea di Reggio Calabria, Aprile 2008. (by courtesy of Omi Villaggio)

Fig. 2 Giuseppe Perotti, Piero Villaggio and Mario Micoli at Rosskuppe Peak (Austrian Alps, Summer 1971)

Publications by Piero Villaggio

Articles on Journals

1. "Sul problema variazionale della lastra curva a simmetria assiale". Atti Acc. Lig. Sc. Lett. XVI, 1–11 (1959).
2. "Sulla teoria non lineare delle lastre sottili elastiche a doppia curvatura". Rend. Ist. Lomb. - Acc. Sc. Lett. 94,151–165 (1960).
3. "Su un problema non lineare relativo alla torsione di solidi prismatici". Atti Acc. Lig. Sc. Lett. XVIII, l-29 (1961).
4. "Problemi di trasmissione dell'elasticità piana". Rend. Ist. Lomb. - Acc. Sc. Lett. 95, 493–524 (1961).
5. "Su un problema elastico piano della teoria dell'attrito volvente". Rend. Ist. Lomb. - Acc. Sc. Lett. 95, 919–940 (1961).
6. "Sui problemi al contorno per sistemi di equazioni differenziali lineari del tipo di stabilità dell'equilibrio elastico". Ann. Sc. Norm. Sup. Pisa, III, vol. XV, I-II, 25–40 (1961).
7. "Su un problema misto relativo ad un semipiano elastico collegato ad una nervatura". Atti Ist. Sc. Costr. Pisa, 84, 1–16 (1961).
8. "Limiti inferiori del carico critico di lastre elastiche a doppia curvatura". Atti Ist. Sc. Costr. Pisa, 86, 1–20 (1961).
9. "Limiti inferiori del primo moltiplicatore critico di lastre sottili elastiche di rivoluzione". Atti Ist. Sc. Costr. Pisa, 89,1–19 (1961).
10. "Sull'esistenza e l'unicità di soluzioni dei problemi al contorno in elastoplasticità, Rend. Ist. Lomb. - Acc. Sc. Lett., 96, 112–127 (1962).
11. "Su un problema misto relativo alla statica delle dighe cilindriche spesse". Atti Ist. Sc. Costr. Pisa, 93, 1–18 (1962).
12. "Tensioni deformazioni in flange sottili circolari elastiche". Atti Ist. Sc. Costr. Pisa, 95, 1–34 (1962).
13. "Limiti superiori del fattore di concentrazione della tensione intorno a fori ipotrocoidali". L'Aerotecnica, vol. VI (3), 119–128 (1962).
14. "Su alcune formule di maggiorazione per l'energia di deformazione di un corpo elastico". Le Matematiche, XVII, 2, 21–135 (1962).
15. "Su un criterio di stabilità cinetica di sistemi meccanici soggetti ad azioni impulsive." Rend. Ist. Lomb. - Acc. Sc. Lett. 97, 96–100 (1963).
16. "Condizioni di sicurezza all'instabilità aerodinamica di lastre sottili in regime supersonico". Atti Ist. Sc. Costr. Pisa, 96, 1–11 (1963).
17. "Sullo scorrimento viscoso del conglomerato in flessione pura". Atti Ist. Sc. Costr. Pisa, 100, 1–14 (1963).
18. "Studio di un sistema servomotore per prove di rilassamento". Atti Ist. Sc. Costr. Pisa, 101, 1–8 (1963).
19. "Limiti di collasso plastico-rigido di unioni chiodate". Costr. Met., vol. 1, 3–11 (1964).
20. "Limiti di resistenza di pali in terreno plastico-rigido". Il Cemento, vol.10, 25–30 (1964).
21. "Analisi limite di piastre sottili appoggiate plastico-rigide". Giorn. Genio Civ., vol.3, 133–141 (1964).
22. "La nozione di efficienza teorica nei bulloni pretesi". Costr. Met., vol. 4, 284–288 (1965).
23. "Alterazione perturbativa dei carichi critici di lastre piane sottili". Giorn. Genio Civ., vol. 12, 601–616 (1965).
24. "Monodimensional solids with constrained solutions". Meccanica, vol. II, 1, 65–68 (1967).
25. "Stability conditions for Elastic-Plastic Prandtl-Reuss solids". Meccanica, vol. III, 1, 46–47 (1968).
26. "Stabilità rispetto al convesso dei domini plastici delle travature". Giorn. Genio Civ., vol. 2–3, 101–109 (1968).

27. “Proprietà di stabilitá e di monotonia nel metodo degli elementi finiti”. L’Aerotecnica. 3–6, 1–9 (1969).
28. “Strong Ellipticity conditions for the differential operator of the finite isotropic elasticity”. Meccanica, vol. V, 191–196 (1970).
29. “Theorems of convergence for minimal sequences in limit analysis”. Int. J. Solids Struct. 5, 833–841 (1969).
30. “A criterion of stability for non-linear continua”. IUTAM Symp. Herrenalb, 19–24 (1969).
31. “A criterion of classification for non-linear hypoelastic materials in simple shear”. Meccanica, vol. VI, 6, 25–29 (1971).
32. “Energetic bounds in Finite Elasticity”. Arch. Rat. Mech. AnaL vol. 45, 4, 282–293 (1972).
33. “Formulation of some homogeneous thermodynamic processes as Variational Inequalities”. In Int. Symposium on foundations of Plasticity. Warsaw, August 30 - Sept. 2, 1972.
34. “Condition of Stability for thermoelastic continua”. Meccanica, vol. VII, 1, 19–22 (1972).
35. “Formulation of some homogeneous thermodynamic processes as variational inequality”. Archiwum Mech. Stosow., vol. 25 (2),293–298 (1973).
36. “The Stability of Continuous Systems”.Meccanica.vol. X, 4, 313–314 (1975).
37. “On Stability in the Classical Linear Theory of Fluid Mixtures”. Ann. Mat. Pura e Appl. (IV), vol. CXI, 51–67 (1976) (with M. Gurtin).
38. “Condition of Stability and Wave Speeds for Fluid Mixtures”. Meccanica., vol. XI (4), 191–195 (1976). (with I. Müller).
39. “Two-Sided Estimates in Unilateral Elasticity”. Int. J. Solids Struct., vol. 13, 279–292 (1977).
40. “A model for an Elastic Plastic Body”. Arch. Rational Mech. Anal., vol. 65 (1), 25–46 (1977) (with I. Müller).
41. “Hadamard’s theorem applied to thin plates”. Journal of Elasticity, vol. 7 (4), 425–436 (1977).
42. “Concavity techniques with application to shock problems”. Journal of Elasticity, vol. 9 (1), 1–13 (1979).
43. “Buckling under unilateral constraints”. Int. J. Solids Struct., vol. 15 (3), 193–201 (1979).
44. “Comparison Properties for Solutions of Unilateral Problems”. Meccanica. vol. XIII (1), 41–47 (1978).
45. “An Elastic Theory of Coulomb Friction”. Arch. Rat. Mech. Anal. vol. 70 (2), 135–144 (1979).
46. “The thermodynamics of fatigue-sensitive materials”. Meccanica, vol. XIV (1), 48–54 (1979).
47. “A unilateral contact problem in linear elasticity”. Journal of Elasticity. vol. 10, N. 2 , 113–119 (1980).
48. “Maximum Modulus Theorems for the Elastic Half-Space”. Riv. Mat. Univ. Parma, 5 (4), 663–672(1979).
49. “Inverse Boundary Value Problems in Structural Optimization”. In Free Boundary Problems, (Proc. Sem. Pavia 1979), Ist. Naz. Alta Matem. vol. II, 587–599.
50. “The Exact Deflection of an Elastic Beam over a Soft Obstacle”. Meccanica, XIV (4), 219–223 (1979).
51. “Stress Diffusion in Non-Linear Interpenetrating Bars”. Int. J. Solids Struct., vol. 17, 411–420 (1981).
52. “The Ritz method in solving unilateral problems in elasticity”. Meccanica, vol. XVI (3),123–127 (1981).
53. “Stress Diffusion in Masonry Walls”. J. Struct. Mech., 9 (4),439–450 (1981).
54. “Are the Optimum Problems in Structural Design Well Posed?” Arch. Rational Mech. Anal. vol. 78 (3),199–211 (1982). (with W. Velte).
55. “Kelvin’s Solution and Nuclei of Strain in a Solid Mixture.” Ann. Sc. Norm. Sup. Pisa, S. IV. X, 1, 109–124 (1983).
56. “A Free Boundary Value Problem in Plate Theory.” Journ. Appl. Mech., vol. 50,297–302 (1983).
57. “Isoperimetric Distributions of Loads in Elastostatics.” Rend. Sem. Mat. Univ. Padova, vol. 68, 261–268 (1982).

58. The Motion of a Detaching Elastic Body.” Arch. Rational Mech. Anal. vol. 85 (2),161–110 (1984).
59. “Detachment Instabilities in Membranes.” Meccanica, vol. 19, 201–205 (1984).
60. “A Signorini Problem in Elasticity with Prescribed Contact Set.” Appl. Math. Optim., vol. 13, 163–774 (1985).
61. “A Correction to the Föppl-v. Karman Equations.” Bollettino U.M.I. (6) 4-B, 761–711 (1985).
62. “The Main Elastic Capacities of a Spheroid.” Arch. Rational Mech. Anal, VoI. 92 (4), 331–353 (1986).
63. “Influence of the Bending Stiffness on the Shape of a Belt in Steady Motion.” J. Appl. Mech., vol. 53 (2), 266–270 (1986). (with R. Fosdick).
64. “Self Straining at the Interface of two bonded Hyperelastic Half-Spaces Uniformly Pre-Stressed.” Quart. J. Mech. Appl. Math., vol. 41 (3), 347–361 (1988) (with J. Dunwoody).
65. “The Virtual Friction of a Sphere Vibrating in an Elastic Medium.” Arch. Rat. Mech. Anal., vol. 102 (3),193–203 (1988).
66. “ “Optimal Distributors of Loads in Plane Elastostatics.” Meccanica, vol. XXIII, 203–208 (1988).
67. “How to Model a Bonded Joint.” J. Appl. Mech., vol. 56 (3), 590–594 (1989).
68. “The Penetration of an Elastic Wedge.” Quad. Sc. Norm. Sup. Pisa, 791–797 (1989).
69. “A Note on the Motion of a String on a Unilaterally Reacting Foundation.” SIAM J. Appl. Math. 49 (4), 1223–1230 (1989). (with R.E.L. Turner).
70. “A Mathematical Theory of a Guillotine.” Arch. Rat. Mech. Anal., vol. 110 (2), 93–101 (1990).
71. “The Minimal Thickness of a Semicircular Arch with a Tension-Free Crown.” Mech. Struct. and Mach., vol. 18 (4), 515–527 (1990).
72. “The equilibrium shapes of earth masses.” Stab. Appl. Anal. Cont. Mech., vol. 1, 2, 149–164 (1991).
73. “The Rigid Inclusion with Highest Penetration.” Meccanica. vol. XXVI, 149–153 (1991).
74. “The Shape of a Free Surface of a Unilaterally Supported Elastic Body.” Ann. Sc. Norm. Sup. Pisa. S.N. V. XVIII, 525–539 (1991).
75. “On the Detachment of an Elastic Body Bonded to a Rigid Support.” Journal of Elasticity, vol. 27, 133–142 (1992). (with W. Velte).
76. “The Best Position of a Reinforcement in an Elastic Sheet.” Bollettino U.M.I (7) 6-A, 103–112 (1992).
77. “A Theory for Brittle Fracture in Compression.” Cont. Mech. Therm., vol. 5, 243–254 (1993). (with J. Dunwoody).
78. “Elastic Waves Against a Corrugated Boundary.” Meccanica, vol. XXVIII, 153–157 (1993).
79. “How to Write a Paper on a Subject in Mechanics.” Meccanica. vol. XXVIII, 163–167 (1993).
80. “The Flattening of Mountain Chains.” In Boundary value problems for partial differential equations and applications. C. Baiocchi and J.L. Lions eds., Masson (Paris), 449–454 (1993).
81. “New Results in Shape Optimization.” In Boundary Control and Boundary Variation (J.P. Zolésio, Ed.) Lecture Notes in Control and Inform. Sci. vol. 178, 356–361, Berlin Heidelberg New York: Springer (1992) (with J.P. Zolésio).
82. “The Problem of Packaging.” In Partial DifferentialEquations and Applications. (Marcellini, Talenti, and Vesentini, Eds.). Lecture Notes in Pure and Appl. Math. Series. No 177, 319–321 (1996).
83. “The Pillar of Best Efficiency.” Journal of Elasticity. vol.42, 79–89 (1996).
84. “The Added Mass of a Deformable Cylinder Moving in a Liquid.” Cont. Mech. Therm., vol. 8, 115–120 (1996).
85. “The Rebound of an Elastic Sphere Against a Rigid Wall.” J. Appl. Mech., vol. 63 (2), 259–263 (1996).
86. “The Thickness of Roman Arches.” In Contemporary Research in the Mechanics and Mathematics of Materials. R.C. Batra and M.F. Beatty (Eds.) CIMNE, Barcellona, 412–479 (1996).

87. “A Semi-Inverse Shape Optimization Problem in Linear Anti-Plane Shear.” Journal of Elasticity, vol. 45, 53–60 (1996) (with C.O. Horgan).
88. “Suboptimal Shape of a Plate Stretched by Planar Forces.” In Partial Differential Equations Methods in Control and Shape Analysis (G. Da Prato and J. P. Zolésio, Eds.). Lecture Notes in Pure Appl. Math. vol. 188, 321–331. New York: Marcel Dekker (1997) (with J.P. Zolésio).
89. “Some Extensions of Carother’s Paradox in Plane Elasticity.” Math. Mech. Solids, vol. 3, 17–28 (1998).
90. “Tlre Roots of Trees.” Cont. Mech. Therm., vol. 10, 233–240 (1998).
91. “Transverse Decay of Solutions in an Elastic Cylinder.” Meccanica, vol. XXXIII, 577–585 (1998) (with R. J. Knops).
92. “Recovery of Stresses in a Beam from those in a Cone.” Journal of Elasticity, vol. 53, 65–75 (1999) (with R. J. KnoPs).
93. “Spatial Behaviour in Plane lncompressible Elasticity on a Half-Strip.” Quart. Appl. Math., vol. LVIII, N. 2, 355–367 (2000) (with R. J. Knops).
94. “The problem of the Stiffener and Advice for the Worm.” Math. Comp. Modeling, vol. 34, 1423–1429 (2001).
95. “The Shape of Roofs.” Meccanica, vol. XXXV, 3, 215–221 (2000).
96. “The High Cantelivered Beam of Minimal Compliance.” Q. J. Mec. Appl. Math., vol. 54 (2), 329–339 (2001) (with R.J. Knops).
97. “An elementary theory of the oblique impact of rods.” Rend. Mat. Acc. Lincei, Serie 9, vol. 12, 49–56 (2001) (with R. J. Knops).
98. “Spatial behaviour in the incompressible linear elastic free cylinder.” Proc. R. Soc. Lond. A, vol. 457, 2113–2135 (2001) (with R. J. Knops).
99. “How to design a foundation.” Int. J. Solids Struct., vol. 38 (48–49), 8899–8906 (2001).
100. “How to design a shock absorber.” Struct. Multidisc. Optim., vol. 23, 88–93 (2001).
101. “’Wear of an Elastic Block.” Meccanica, vol. XXXVI, 243–249, (2007).
102. “The apparent propagation velocity of a wave.” Rend. Mat. Acc. Lincei, Serie.9, vol. 12, 191–197 (2001).
103. “Distortions in the History of Mechanics.” Meccanica, vol. XXXVI, 589–592 (2001)
104. “Newton’s aerodynamic problem in the presence of friction.” Nonlinear Differ. Equ. Appl., vol. 9 , 296–301 (2002) (with D. Horstmann and B. Kawohl).
105. “On the termodynamics of repetitive visco-plastic moulding.” Z. angew. Math. Phyis., vol. 53 , 1139–1149 (2002) (with I. Müller and H.-S. Sahota).
106. “Elastic materials with sparse, rigid reinforcement and debonding.” Cont. Mech. Therm., vol. 15, 287–294 (2003) (with J. Jenkins).
107. “The Dynamics of a Detaching Rigid Body.” Meccanica, vol. 38 (5), 595–609 (2003) (with M. Leitman).
108. “Brittle detachment of a stiffener bonded to an elastic plate.” J. Eng. Math., vol. 46 (3–4), 409–416 (2003).
109. “A Historical Survey of Impact Theories.” In Essays on the History of Mechanics, 223–234. Basel: Birkhäuser (2003).
110. “An Approximate Treatment of Blunt Body Impact.” Journal of Elasticity, vol. 72, 213–228 (2003) (with R. J. Knops).
111. “Calcolo delle variazioni e teoria delle strutture.” Boll. U.M.I. La Matematica nella Società e nella Cultura Serie VIII, vol. VII-A, 49–76 (2004).
112. “A model of seismic excitation.” Rend. Mat. Acc. Lincei 1.9, vol. 15, 119–123 (2004).
113. “Hammering of Nails and Pitons.” Math. Mech. Solids, vol. 10, 461–468 (2005).
114. “On Enriques’s Foundations of Mechanics.” In Two Cultures. Essays in Honour of D. Speiser (K. Williams Ed.) Basel: Birkhäuser, 132–138 (2005).
115. “An Extension of the Complex Variable Method in Plane Elasticity to Domains with Corners: A Notch Problem.” Journal of Elasticity, vol. 81, 206–215 (2005) (with M. Leitman).
116. “A theory of plastic splashing.” Red. Lincei Mat. Appl. 17, 89–93 (2006).
117. “Frobenius’ method for curved cracks.” Int. Journal of Fracture 139, 59–69 (2006) (with Roberto Ballarini).

118. “The Dynamics of a Membrane Shock-Absorber.” Mechanics Based Design of Structures and Machines, vol. 34, 277–292 (2006) (with M. Leitman).
119. “An optimum braking strategy.” Meccanica vol. 41, 693–696 (2006) (with R. J. Knops).
120. “Exact solutions in the reinforcement of a circular plate under concentrated loads.” Struct. Multidisc. Optim., vol. 32, 427–433 (2006) (with R. J. Knops).
121. “Problemi Variazionali Plebei.” Boll. U.M.I La Matematica nella Società e nella Cultura. Serie VIII, vol. X - A, 1–20 (2007).
122. “A mathematical theory of climbing.” IMA Joumal of Applied Mathematics, vol. 72, 570–576 (2007).
123. “How to climb the face of a mountain.” Note di Matematica, vol. 27 (2), 249–255 (2007).
124. “Small perturbations and large self-quotations.” Meccanica, vol. 43, 81–83 (2008).
125. “Karl-Eugen Kurrer, The History of the Theory of Structures. From Arc Analysis to Computational Mechanics” (Book Review). Journal of Elasticity, vol. 93, 199–202 (2008).
126. “Axial impact on a semi-infinite elastic rod.” Rend. Lincei Mat. Appl., vol. 19 , 205–210 (2008).
127. “On Saint-Venant’s Principle for Elastic-Plastic Bodies.” Math. Mech. Solids, vol. 11 (7) , 601–621 (2009) (with R. J. Knops).
128. “Some Plebeian Variational Problems.” In Variational Analysis and Aerospace Engineering. G. Buttazzo, A. Frediani (eds.). Springer Science * Business Media 509–518 (2009) .
129. “Deep Foundations.” Rend. Lincei Mat. Appl. vol. 20 , 413–420 (2009).
130. “Plastic Zone around Circular Holes” ASCE J. Eng. Mech., vol. 135 (12), 1467–1471 (2009) (with M. Leitman).
131. “Reviews of Books.” Meccanica, vol. 45, 891–899 (2010) .
132. “Optimal Parabolic Arches.” lntern. J. Eng. Science, vol. 48, 1433–1439 (2010) (with MarshalI Leitman).
133. “Artificial and Geological Isoresistant Buttresses.” In Mechanics and Architecture between Epistéme and Téchne. Ed. By Anna Sinopoli. Edizioni di Storia e Letteratura. Roma, 25–36 (2010).
134. “Elastic stress diffusion around a thin corrugated inclusion.” IMA Journal of Applied Mathematics vol.76, 633–641 (2011) (with R. Ballarini).
135. “An approximate theory of pseudo-arches.” lnt. J. Solids Struct., vol. 48, 2960–2964 (2011) (with M. Leitman).
136. “Illustrations of Zanaboni’s formulation of Saint-Venant’s principle.” Rend. Lincei Mat. Appl. vol. 22 , 347–364 (2011)(with R.J. Knops).
137. “Sixty years of solid mechanics.” Meccanica, vol. 46,1171–1189 (2011).
138. “Remarks on Neuber’s Complex Variable Procedure for Plane Elastic Solutions.” Z. Angew. Math. Mech., vol. 92 (3), 196–203 (2012) (with R.J. Knops).
139. “The Warlike Interest in Impact Theories.” In Variational Analysis and Aerospace Engineering. (G. Buttazzo and A. Frediani, Eds.) Springer, 427–434 (2012).
140. “On oblique impact of a rigid rod against a Winkler foundation.” Cont. Mech.Therm., vol. 24, 559–582 (2012) (with R. J. Knops).
141. “Zanaboni’s treatment of Saint-Venant’s principle.” Applicable Analysis, vol. 91, 345–370 (2012) (with R. J. Knops).
142. “Some Ambiguities in the Complex Variable Method in Elasticity.” Journal of Elasticity, vol. 109, 223–234 (2012) (with M. Leitman).
143. “The Dynamics of a Landslide.” Atti della Acc. Peloritana dei Pericolanti, vol. 91, Suppl. N” 1, A11, 1–9 (2013) (with M. Leitman).
144. “Crisis of mechanics literature?” Meccanica, A8, 765–767 (2013).
145. “Parabolic tunnels in a heavy elastic medium”. Rend. Lincei Mat. Appl., vol. 24, 1–10 (2013) (with M. Leitman).
146. “The shape of a glacier.” Boll. Unione Mat. Italiana, Serie IX, vol. VI, 299–317 (2013) (with R. J. Knops).
147. “The optimal shape of a plinth.” J. Eng. Sci, vol. 77, 24–29 (2014) (with R. J. Knops).

148. “Deformable stiffener welded to an elastic plate” ASCE J. Eng. Mechs, (to appear) (with R. J. Knops).
149. Transmission of concentrated forces in plane elasticity across similar bodies.” (In preparation, with M. J. Leitman).

Books

A1. P. Villaggio, Qualitative Methods in Elasticity. Nordhoff, Groningen (1977)
A2. P. Villaggio, Mathematical Models for Elastic Structures. Cambridge University Press, Cambridge (1997)
A3. P. Villaggio, Die Werke von Johann I und Nicolaus II Bernoulli. Birkhäuser, Basel (2007)

Preface

New challenges in aerospace sciences and engineering are not limited to partial improvements of the flying systems but include the design and optimization of innovative machines in order to provide a jump forward in air or space transport, to cut noxious emissions, to fly safer and quieter. The series of the workshops held at the "Ettore Majorana Foundation and Centre for Scientific Culture" of Erice brings together mathematicians and aerospace engineers from academia and industry with the aim to discuss on the new challenges in aerospace.

This volume collects most of the papers presented at a workshop held between August 28 and September 5, 2015, and other contributions from eminent authors are also presented. The editors are confident that this book will capture the interest of researchers working on the new frontiers in Aeronautics and Space Sciences and Technologies with an extensive application of Mathematics.

This volume is dedicated to Prof. Piero Villaggio, who passed away on January 4, 2014. Prof. Villaggio published "The Warlike Interest in Impact Theories" and "Some Plebeian Variational Problems" in the 2010 and 2008 volumes of this series, respectively; these contributions will remain as marvellous examples of scientific culture and elegance, as elegant was always Piero along his life.

During the preparation of this book, Prof. Angelo Miele passed away. Angelo was the director of many Erice workshops and he has been a giant in almost every field in aerospace sciences, engineering and applied mathematics.

Acknowledgements

This volume collects contributions presented in the workshop "Variational Analysis and Aerospace Engineering III: Mathematical Challenges for the Aerospace of the Future" held at the School of Mathematics of the E. Majorana Centre in Erice on August 28 to September 5, 2015.

The workshop has been possible thanks to the contributions of:

- University of Pisa, Department of Civil and Industrial Engineering (Italy)
- project HYPSTAIR ("Development and validation of hybrid propulsion system components and sub-systems for electrical aircraft"), co-funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 605305.

We gratefully acknowledge the support and organization of the E. Majorana Centre and Foundation for Scientific Culture in Erice, Sicily (Italy) and the precious help by prof. Franco Giannessi, Director of the School of Mathematics "Guido Stampacchia" of the Erice Foundation.

Pisa, Italy	Aldo Frediani
Montpellier, France	Bijan Mohammadi
Paris, France	Olivier Pironneau
Pisa, Italy	Vittorio Cipolla
July 2016	

Contents

Molding Direction Constraints in Structural Optimization via a Level-Set Method

Grégoire Allaire, François Jouve, and Georgios Michailidis

Abstract In the framework of structural optimization via a level-set method, we develop an approach to handle the directional molding constraint for cast parts. A novel molding condition is formulated and a penalization method is used to enforce the constraint. A first advantage of our new approach is that it does not require to start from a feasible initialization, but it guarantees the convergence to a castable shape. A second advantage is that our approach can incorporate thickness constraints too. We do not address the optimization of the casting system, which is considered a priori defined. We show several 3d examples of compliance minimization in linearized elasticity under molding and minimal or maximal thickness constraints. We also compare our results with formulations already existing in the literature.

1 Introduction

The increasing number of publications on industrial applications of shape and topology optimization reflects the interest of engineers to introduce these techniques in the design process of mechanical structures. Especially in case of complicated problems, where mechanical intuition is very limited, shape and topology optimization can serve as a valuable tool both in order to obtain an optimized structure and to accelerate the design process.

Among the several methods that appeared in the literature, such as Solid Isotropic Material with Penalization (SIMP) method [14, 16, 49], the homogenization method

G. Allaire (✉)
Centre de Mathématiques Appliquées (CMAP), École Polytechnique, CNRS UMR 7641, Université Paris-Saclay, 91128 Palaiseau, France
e-mail: gregoire.allaire@polytechnique.fr

F. Jouve
Laboratoire J.L. Lions (UMR 7598), University Paris Diderot, Paris, France
e-mail: jouve@ljll.univ-paris-diderot.fr

G. Michailidis
SIMaP-Université de Grenoble INPG, Grenoble Cedex 1, France
e-mail: michailidis@cmap.polytechnique.fr

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_1

[4, 5, 15], the phase-field method [17, 18, 41, 50], or the Soft Kill Option [23, 31], the level-set method for shape and topology optimization [6, 7, 35, 38, 44] seems to fulfill industrial requirements in a satisfying way. Using a level-set function to describe implicitly the boundary of a shape [36, 37] allows topological changes to appear in an easy way, while the geometric nature of the method is a benefit for the study of problems where the position of the interface plays a significant role (stress constraints, thermal problems with flux across the boundary, etc.). The method is independent of the objective function under study [3, 9, 20, 21] and the ability to adapt the mesh on the boundary [8, 12, 47] alleviates possible numerical difficulties due to the "ersatz" material or to the discontinuity of the material properties.

Moreover, industrial design introduces significant constraints according to the fabrication method, the tooling limitations, and the total cost that can be afforded. Some of them are essentially geometric constraints, related to a notion of local thickness. We have shown in our previous work [13] that thickness can be explicitly controlled using a level-set method, which constitutes a great advantage for the industrialization of the method. Such constraints are of great significance for cast parts, i.e., structures that are intended to be constructed by casting.

Casting [19] is the fabrication process where molten liquid is poured into a cavity formed by molds. The final structure is obtained after solidification of the liquid and removal of the molds. Thus, the structure should have such a shape, so that the construction and the removal of the molds is possible without destroying either the structure or the mold. This is called the "molding constraint." The casting process imposes further specifications of mechanical nature on the shape of the structure, mainly related to the solidification and filling process. In [32], we argued that such constraints can be translated, in the context of topology optimization, into geometrical constraints on the maximum and minimum allowable feature size, since the complete casting system (molding, solidification, and filling system) is usually designed after the structural form definition.

According to the choice of shape and topology optimization algorithm, different ways have been proposed to handle the molding constraint. In the framework of the SIMP method, which is a density method, Zhou et al. [51] implemented a penalization scheme that favors higher densities at the lower part of the structure. Leiva et al. [29] have chosen to introduce directly the growth direction in the parameterization of the problem, while methods of topology control, such as connectivity and growth direction control, have been applied for the Soft Kill Option [24]. A complete review of these methods and a comparison of results of topology optimization with and without manufacturing constraints can be found in [25, 26].

In the framework of the level-set method, the only works on the topic—to our knowledge—are those of Xia et al. [45, 46]. In [45] the authors have proposed a molding condition on the design velocity, i.e., a modification of the descent direction that ensures the castability of the shape at each iteration, provided that the initial shape is also castable. In this work, the molding system is a priori defined. In [46] the authors have added the optimization of the draw direction in the optimization problem. The same choice for the design velocity is done. Although the method allows those topological changes that do not come in conflict with castability, it is mentioned in [45] that the shape cannot expand orthogonally to the casting direction. This is a great disadvantage in case one wants to impose a minimum thickness constraint.

In the present paper, we introduce a new approach to handle the molding constraint in the framework of the level-set method for shape and topology optimization. A pointwise constraint is formulated using the signed distance function and a penalty functional is then constructed to turn the constraint into a global one. A shape derivative [2, 27, 33, 34, 39] is calculated for this new functional and a simple penalization method is applied which guarantees that the optimal shape is castable at convergence. A first advantage of our new approach is that it does not require to start from a feasible initialization. This is of course a key feature since, in many industrial problems, it is very hard to find out a feasible design to start with. A second advantage is that our approach can incorporate thickness constraints, contrary to the previous method in [45, 46].

The contents of our paper are as follows. Section 2 describes our model shape optimization problem. For simplicity we focus on compliance minimization with volume constraint: the main difficulty on which we shall focus is the addition of further molding and thickness constraints. Section 3 is a short review of the level-set method. Section 4 discusses the casting process while Sect. 5 introduces our new molding direction constraint. We also recall the approach of Xia et al. [45, 46], as well as the "uniform cross-section surface constraint" of Yamada et al. [48], which simplifies a lot the shape of the desired molds. Section 6 is devoted to the computation of the shape derivatives of these molding constraints. Finally Sect. 7 features our 3d numerical results which are obtained in the finite element software SYSTUS of ESI-Group [40], which is well adapted to an industrial context. Our results were partially announced in [10].

2 Setting of the Problem

Our goal is to optimize a shape $\Omega \subset \mathbb{R}^N$ ($N = 2$ or 3), a bounded domain occupied by a linear isotropic elastic material with Hooke's law A (a positive definite fourth-order tensor). Typically, the boundary of Ω is comprised of three disjoint parts, such that $\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_0$, with Dirichlet boundary conditions on Γ_D, non-homogeneous Neumann boundary conditions on Γ_N, and homogeneous Neumann boundary conditions on Γ_0. We introduce a working domain D (a bounded domain of $\mathbb{R}^N$) which contains all admissible shapes, that is, $\Omega \subset D$. The volume and surface loads are given as two vector-valued functions defined on D, $f \in L^2(D)^N$ and $g \in H^1(D)^N$. The displacement field u is the unique solution in $H^1(\Omega)^N$ of the linearized elasticity system

$$\begin{cases} -\mathrm{div}\,(A\,e(u)) = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \big(A\,e(u)\big)n = g & \text{on } \Gamma_N, \\ \big(A\,e(u)\big)n = 0 & \text{on } \Gamma_0, \end{cases} \tag{1}$$

where $e(u)$ is the strain tensor, equal to the symmetrized gradient of u. A classical choice for the objective function $J(\Omega)$ to be minimized is the compliance (the work done by the loads). It reads

$$J(\Omega) = \int_\Omega f \cdot u\,dx + \int_{\Gamma_N} g \cdot u\,ds = \int_\Omega A\,e(u) \cdot e(u)\,dx. \tag{2}$$

A typical shape optimization problem is

$$\inf_{\Omega \in \mathcal{U}_{ad}} J(\Omega)\,, \tag{3}$$

where $\mathcal{U}_{ad}$ is the set of admissible shapes. Imposing that all shapes belong to the working domain D and that they satisfy a volume constraint $0 < V < |D|$, a possible choice of admissible set is

$$\mathcal{U}_{ad} = \{\Omega \subset D \text{ such that } |\Omega| = V\}\,. \tag{4}$$

As it is well known [2, 16], problem (3) may lack an optimal solution. Numerically, the non-existence of a minimizer of (3) is reflected by the fact that approximate numerical solutions are mesh dependent (the finer the mesh, the more details or finer members in the solution). Classically, to obtain existence of optimal shapes, one needs to restrict further the admissible set $\mathcal{U}_{ad}$ by imposing additional geometrical, topological, or smoothness constraints [2, 34, 39].

In order to find a descent direction for advecting the shape, we rely on the Hadamard method of shape differentiation, following the approach of Murat and Simon [34]. Starting from a smooth reference open set Ω, we consider domains of the type

$$\Omega_\theta = \big(Id + \theta\big)(\Omega),$$

with $\theta \in W^{1,\infty}(\mathbb{R}^N, \mathbb{R}^N)$. It is well known that, for sufficiently small θ, $(Id + \theta)$ is a diffeomorphism in $\mathbb{R}^N$.

Definition 1. The shape derivative of $J(\Omega)$ at Ω is defined as the Fréchet derivative in $W^{1,\infty}(\mathbb{R}^N, \mathbb{R}^N)$ at 0 of the application $\theta \to J\big((Id + \theta)(\Omega)\big)$, i.e.,

$$J\big((Id + \theta)(\Omega)\big) = J(\Omega) + J'(\Omega)(\theta) + o(\theta) \quad \text{with} \quad \lim_{\theta \to 0} \frac{|o(\theta)|}{\|\theta\|} = 0\,,$$

where $J'(\Omega)$ is a continuous linear form on $W^{1,\infty}(\mathbb{R}^N, \mathbb{R}^N)$.

A classical result states that the shape derivative $J'(\Omega)(\theta)$ depends only on the normal trace $\theta \cdot n$ on the boundary $\partial\Omega$ [27, 34, 39]. We refer to [7] for various examples of shapes derivatives in the elasticity setting, including that for compliance.

3 Level-Set Framework

3.1 Shape Representation

We favor an Eulerian approach and use the level-set method [36] to capture the shape Ω on a fixed mesh. Then, the boundary of Ω is defined by means of a level-set function ψ (see Fig. 1) such that

$$\begin{cases} \psi(x) = 0 \Leftrightarrow x \in \partial\Omega \cap D, \\ \psi(x) < 0 \Leftrightarrow x \in \Omega, \\ \psi(x) > 0 \Leftrightarrow x \in \left(D \setminus \overline{\Omega}\right). \end{cases}$$

During the optimization process the shape is being advected with a scalar (normal) velocity $V(x)$ derived from shape differentiation, as we will see in the sequel. The advection is described in the level-set framework by introducing a pseudo-time $t \in \mathbb{R}^+$ and solving the well-known Hamilton–Jacobi equation

$$\frac{\partial\psi}{\partial t} + V|\nabla\psi| = 0. \tag{5}$$

using an explicit second-order upwind scheme [37].

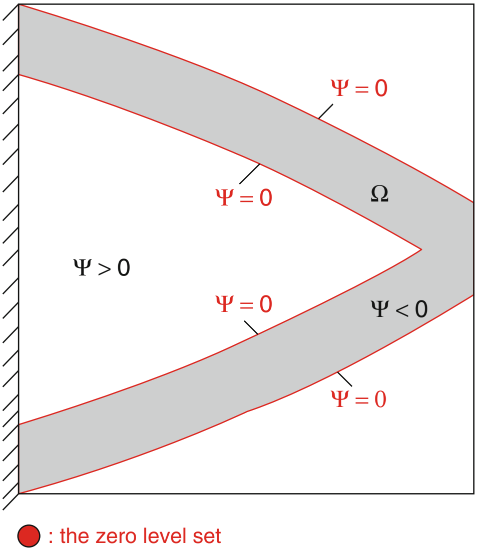


Fig. 1 Level-set representation of a shape

3.2 Signed Distance Function

We recall that if $\Omega \subset \mathbb{R}^N$ is a bounded domain, then the **signed distance function** to the boundary $\partial\Omega$ is the function $\mathbb{R}^N \ni x \mapsto d_\Omega(x)$ defined by:

$$d_\Omega(x) = \begin{cases} -d(x, \partial\Omega) & \text{if } x \in \Omega \\ 0 & \text{if } x \in \partial\Omega \\ d(x, \partial\Omega) & \text{if } x \in \left(\mathbb{R}^N \setminus \overline{\Omega}\right) \end{cases},$$

where $d(x, \partial\Omega)$ is the usual Euclidean distance from x to $\partial\Omega$.

Very often, the Hamilton–Jacobi equation (5) is initialized, or re-initialized, with the signed distance function. However, at later times t, the level-set function $\psi(t, x)$, solution of (5), is not a signed distance function. Furthermore, the functions ψ and d_Ω do not share the same boundary conditions (see [13] for details). Therefore one cannot retrieve geometrical informations on the shape $\Omega(t)$ from $\psi(t, x)$. However, at every time t it is not hard to compute the signed distance function of $\Omega(t)$. As in the case of thickness constraints [13], we shall use this signed distance function to get all necessary information for the formulation of our molding constraints.

3.3 Ersatz Material

Using the so-called ersatz material approach, we extend the state equations to the whole domain D. To do this, we fill the holes $D \setminus \Omega$ by a weak phase that mimics the void, but at the same time avoids the singularity of the rigidity matrix. More precisely, we define an elasticity tensor $A^*(x)$ which is a mixture of A in Ω and of the weak material mimicking holes in $D \setminus \Omega$

$$A^*(x) = \rho(x)A \quad \text{with} \quad \rho = \begin{cases} 1 & \text{in } \Omega, \\ 10^{-3} & \text{in } D \setminus \Omega. \end{cases} \tag{6}$$

Decomposing the boundary ∂D of the working domain in three parts

$$\partial D = \partial D_D \cup \partial D_N \cup \partial D_0,$$

and demanding that the shape boundary $\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_0$ must further satisfy

$$\Gamma_D \subset \partial D_D, \quad \Gamma_N \subset \partial D_N,$$

where ∂D_0 supports homogeneous Neumann boundary conditions, the displacement u is finally computed as the solution of

$$\begin{cases} -\text{div}\,(A^*\, e(u)) = f & \text{in } D, \\ u = 0 & \text{on } \partial D_D, \\ \big(A^*\, e(u)\big)n = g & \text{on } \partial D_N, \\ \big(A^*\, e(u)\big)n = 0 & \text{on } \partial D_0. \end{cases} \tag{7}$$

4 Casting Process

We give in this section a short description of the casting process. A simplified sequence of steps for the construction of a cast part is the following:

1. Molds are used in order to create a cavity, having the shape of the structure that we intend to construct.
2. The cavity is filled with molten liquid metal.
3. The liquid solidifies.
4. The molds are removed and the cast part is revealed.

There are many different types of casting (metal casting, sand casting, investment casting, etc.) and the choice among them depends on the type of cast part. Each type inserts different constraints on the casting process. We address the interested reader to [19] for a complete presentation of the casting process. Here we confine ourselves to **permanent mold casting**, in which the molds are removed without being destroyed. We call **parting direction** the direction along which one mold is removed and **parting surface**, the surface on which different molds come in contact [45]. Note that several molds can be used in the casting system and each one has its own parting direction (see Fig. 3). The parting surface between two molds can be predefined or it can be constructed after the optimization using suitable methods [1, 22]. In most of the industrial applications, planar parting surface is preferred because of reasons of cost and simplicity [45].

Each of the above steps introduces different constraints in the shape of the cast part. In this work we are mainly interested to ensure the feasibility of the last step, i.e., the removal of the molds. Thus, we need to impose that the cast part has such a shape, so that the molds can actually be removed after the end of the solidification process. Let us give a 2d example of the above mentioned. Suppose that for an optimization problem like the one described by Eqs. (1)–(4) we obtain the optimized shape Ω, shown in Fig. 2. In Fig. 3 we see that depending on the molding system considered, this shape can be moldable or not. In the right figure of Fig. 3, some parts of the shape oppose to the removal of the molds in their corresponding parting direction.

The construction of the casting system is usually based on the intuition of the caster. Changes on the number and on the position of the molds can turn a non-moldable shape to a moldable one. The design of the whole casting system is very difficult (if possible) to be formulated mathematically and be subjected to optimization. Works in this direction are mostly concerned with parametric or shape optimization of parts of the molds [30, 43] or of the riser [42]. In the present paper, we do not consider the optimization of the casting process, but the molding system is considered a priori defined.

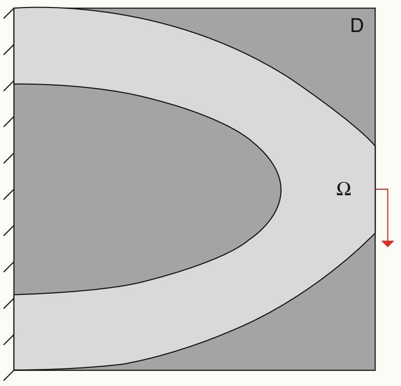


Fig. 2 Possible optimized shape of a cast part

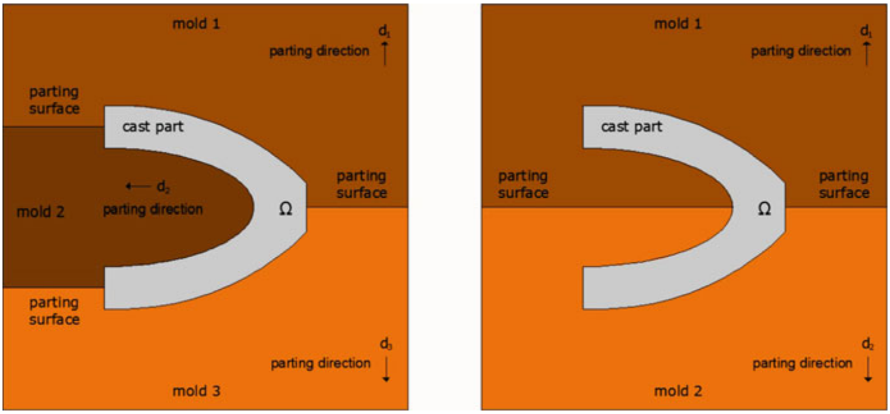


Fig. 3 *Left*: moldable shape; *right*: non-moldable shape

5 Formulation of the Molding Direction Constraint

5.1 Molding Direction Condition on Design Velocity

A molding direction condition on the design velocity was proposed by Xia et al. in [45], which is inspired by Fu et al. [22]. According to these authors, if a shape is feasible with respect to the molding direction specification for its corresponding molding system, then the boundary of the structure $\partial\Omega$ can be divided into m disjoint parts Γ_i, $i = \{1, \dots, m\}$, such that $\Gamma_i \cap \Gamma_j = \emptyset, j = \{1, \dots, m\}$, $\overline{\cup_{i=1}^{m} \Gamma_i} = \partial\Omega$ and Γ_i can be parted in the direction d_i. Thus, a molding direction condition for this shape is

$$d_i \cdot n(x) \geq 0, \quad \forall x \in \Gamma_i, \tag{8}$$

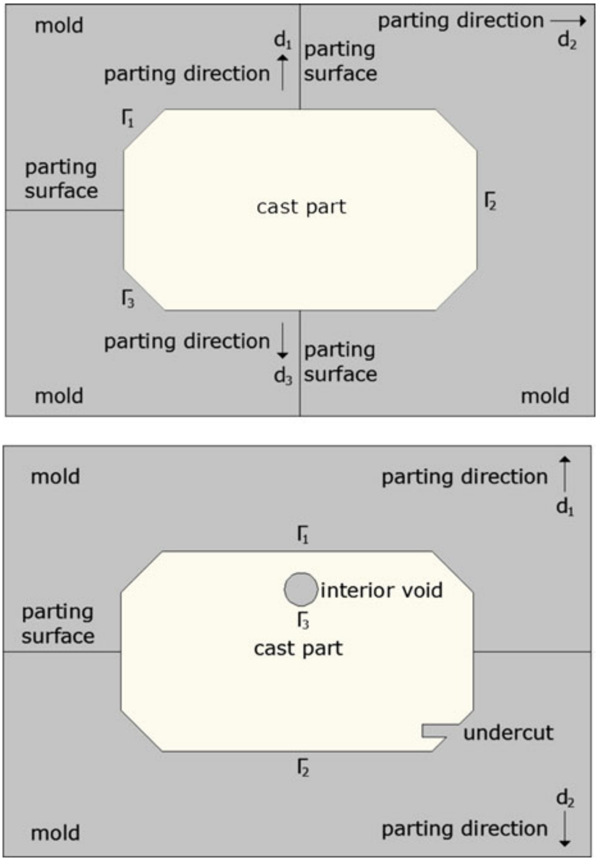


Fig. 4 *Top*: moldable shape; *bottom*: non-moldable shape

where $n(x)$ is the exterior unit normal at $x \in \Gamma_i$. The shape on the left in Fig. 4 satisfies the condition (8), while the shape on the right does not. In fact, as it is mentioned in [45], undercuts (slots that hint the removal on the mold in its parting direction) and interior voids are not allowed.

Based on the molding condition (8), Xia et al. [45] proposed the following method: starting from a shape **that satisfies the constraint (8)**, consider an advection velocity of the form

$$\theta_i(x) = \lambda(x) d_i, \quad \forall x \in \Gamma_i. \tag{9}$$

In this way, the shape remains always moldable, since no undercut can be created during the advection of the shape with this type of velocity and no interior void can be nucleated. The topological changes that can occur using this advection velocity cannot turn the shape from moldable to non-moldable [45].

This method, despite its simplicity and effectiveness, presents two major drawbacks. First, the shape must be initialized as being castable so that it can satisfy the molding constraint during the entire optimization process ; this is a severe limitation on the choice of admissible initial guess shapes. Nevertheless, if such an initialization can be found, then it turns out that the method is flexible enough, especially in 3d, in order for complicated topologies to appear from very trivial initializations. Second, and more important from our point of view, the very form (9) of the advection velocity does not allow all possible deformations of the shape, including those which are required for some other constraints. As it is stated in [45], there is no component of the advection velocity normal to the parting direction. Therefore, the shape can shrink by extinction of some part, but it cannot expand normal to its corresponding parting direction. As an example, consider the case where a minimum thickness constraint is also applied [13]. Then, if the measured thickness is in a direction orthogonal to the parting direction, the shape cannot expand in this orthogonal direction (in order to meet the constraint of minimal thickness) because it can move only parallel to its parting direction. Therefore, in such a situation, the thickness constraint will not be respected. Therefore, it is necessary to formulate a more general molding constraint, free of the above limitations.

5.2 Generalized Molding Constraint

A first idea for a generalized way to treat the molding direction constraint consists simply in regarding (8) as a pointwise constraint in our optimization problem. Then, it can be exactly penalized as we shall do in (16) to compute its shape derivative.

A second idea is to use the signed distance function to the boundary of the domain to derive all necessary information, as we have done for thickness constraints in [13]. Denoting Ω the actual shape and D the design domain, a generalized molding direction constraint can be formulated as:

$$d_{\Omega}\left(x+\xi d_i\right) \geq 0 \quad \forall x \in \Gamma_i, \forall \xi \in [0, dist(x, \partial D)]\,, \tag{10}$$

or equivalently

$$d_{\Omega}\left(x+\xi d_i\right) \geq 0 \quad \forall x \in \Gamma_i, \forall \xi \in [0, diam(D)]\,, \tag{11}$$

where we denote $diam(D) = sup_{x,y}\left\{dist(x, y),\ x, y \in D\right\}$ the diameter of the fixed domain D. We prefer to use formulation (11) instead of (10), in order to avoid the dependence of the term $dist(x, \partial D)$ on the shape Ω.

Intuitively, this formulation says that, starting from a point on the boundary, which will be casted in the direction d_i and travelling along this direction, we should not meet again some part of the structure (see Fig. 5). In case that the parting surface

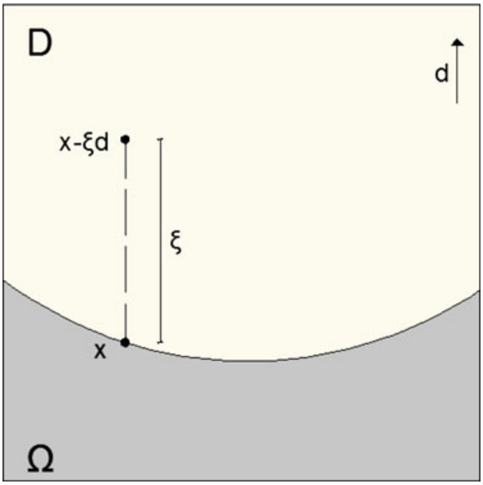


Fig. 5 Checking castability along the parting direction d at the point $x \in \partial\Omega$

is not defined a priori, but is revealed at a second step after the design has been completed and for a system of two molds (see Fig. 3, right image), the constraint (11) becomes

$$d_{\Omega}\left(x + \xi sign(n \cdot d)d\right) \geq 0 \quad \forall x \in \partial\Omega, \forall\xi \in [0, diam(D)]\,. \tag{12}$$

5.3 *Uniform Cross-section Surface Constraint*

Another useful constraint for cast parts is the so-called uniform cross-section surface constraint [48], since it simplifies a lot the shape of the desired molds. To our knowledge, Yamada et al. [48] were the first to study this type of constraint in shape and topology optimization using a combination of a phase-field and a level-set method. The constraint states that the cast part should have a uniform constant thickness along some direction d. An example of a uniform cross-section cantilever of thickness h is given in Fig. 6. The boundary conditions may not be uniform along this direction and therefore the problem cannot be reduced to a 2d problem. We can formulate this type of constraint at least in two ways. The first formulation states that the normal to the boundary cannot have a non-zero component in this direction d:

$$d \cdot n(x) = 0, \quad \forall x \in \partial\Omega \setminus \partial D. \tag{13}$$

A second way to enforce the constraint is to limit the admissible advection fields θ. Starting from an initial guess shape that has a uniform cross-section along the desired direction d and constraining the advection fields to be zero along this

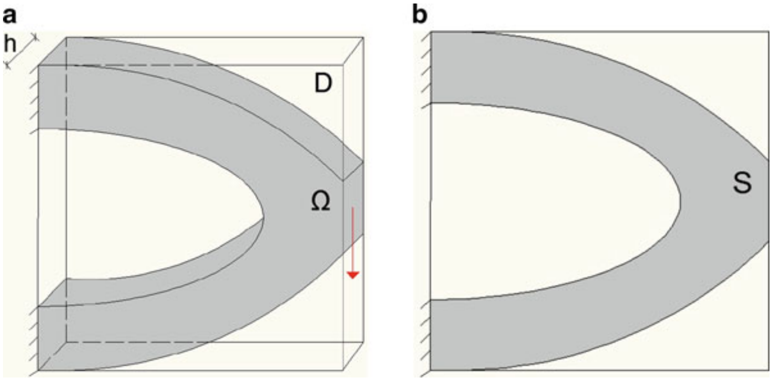


Fig. 6 (**a**) Uniform cross-section cantilever of thickness h; (**b**) cross-section S

direction, the thickness along d will not change. In fact, this is the easiest way to follow, since no mathematical constraint is imposed in the optimization process and the calculation of the velocity field is reduced to a 2d problem, as we will see in the next section.

By enforcing the constraint (13), the feasibility of the shape is guaranteed for casting along the direction d, i.e., this constraint is a sufficient but not a necessary condition.

6 Shape Derivative

6.1 Derivative of the Condition on Design Velocity

Xia et al. proposed in [45] a modification of the advection velocity according to (9) that guarantees a descent direction. Starting from the general form of the shape derivative for a functional $J(\Omega)$

$$J'(\Omega)(\theta) = \int_{\partial\Omega} \theta(s) \cdot n(s)V(s)ds = \sum_{i=1}^{m} \int_{\Gamma_i} \theta_i(s) \cdot n(s)V_i(s)ds$$

and considering admissible advection fields of the type (9), we get

$$J'(\Omega)(\theta) = \sum_{i=1}^{m} \int_{\Gamma_i} \lambda_i(s)d_i \cdot n(s)V_i(s)ds,$$

and choosing

$$\lambda_i(s) = -V_i(s) d_i \cdot n(s),\ \forall i = 1, \ldots, m$$

for each part Γ_i of the boundary $\partial\Omega$, the shape derivative becomes

$$J'(\Omega)(\theta) = -\sum_{i=1}^{m} \int_{\Gamma_i} (d_i \cdot n(s))^2 (V_i(s))^2 ds \leq 0,$$

which shows that the chosen advection velocity

$$\theta_i(s) = -V_i(s)(d_i \cdot n(s)) d_i,\ \forall i = 1, \ldots, m \tag{14}$$

is indeed a descent direction. We replace the Hamilton–Jacobi equation (5) by the linear transport equation

$$\frac{\partial \psi}{\partial t} + \theta \cdot \nabla \psi = 0, \tag{15}$$

where the vectorial velocity θ is an extension of the advection velocity (14) and the normal n is the normal associated with the initial shape.

6.2 Derivative of the Generalized Molding Constraint

We start with the derivation of constraint (8). One advantage of this constraint is that it is of local nature, i.e., it contains information only for points on the boundary without searching along rays emerging from them. On the other hand, it contains the exterior normal vector, whose derivation is more complicated than the one of the signed distance function. In a first step, a global penalty functional can be formulated as

$$P_{GMC}(\Omega) = \int_{\partial\Omega} [(d \cdot n(s))^-]^2 ds, \tag{16}$$

with the usual notations $(f)^+ = \max(f, 0)$ and $(f)^- = \min(f, 0)$.

Proposition 1. *For a smooth shape Ω, the shape derivative of (16) reads*

$$P'_{GMC}(\Omega)(\theta) = \int_{\partial\Omega} \theta(s) \cdot n(s) \Big(2d \cdot \nabla_s (d \cdot n(s))^- - H(s)[(d \cdot n(s))^-]^2 \Big) ds, \tag{17}$$

where H is the mean curvature and ∇_s the tangential gradient.

Proof. Using a classical result about shape derivation of integrals with shape-dependent integrands (see Proposition 6.28 in [2]), the shape derivative of (16) reads

$$\begin{aligned}
&P'_{GMC}(\Omega)(\theta) = \\
&\int_{\partial\Omega} \theta(s)\cdot n(s)\big[H(s)[(d\cdot n(s))^-]^2 + \frac{\partial([(d\cdot n(s))^-]^2)}{\partial n}\big]ds \\
&+\int_{\partial\Omega} \frac{\partial([(d\cdot n(s))^-]^2)}{\partial\Omega}(\theta)ds = \\
&\int_{\partial\Omega} \theta(s)\cdot n(s)\big[H(s)[(d\cdot n(s))^-]^2 + 2(d\cdot n(s))^-\frac{\partial(d\cdot n(s))}{\partial n}\big]ds \\
&+\int_{\partial\Omega} 2(d\cdot n(s))^- d\cdot n'(s)(\theta)ds = \\
&\int_{\partial\Omega} \theta(s)\cdot n(s)\big[H(s)[(d\cdot n(s))^-]^2 + 2(d\cdot n(s))^- d\cdot((\nabla n)n)\big]ds \\
&+\int_{\partial\Omega} 2(d\cdot n(s))^- d\cdot n'(s)(\theta)ds,
\end{aligned} \tag{18}$$

where $n'(s)(\theta)$ is the shape derivative of the normal. Under the smoothness assumption on the shape, there exists an extension of the unit normal in a tubular area around the boundary by $n(x) = \nabla d_\Omega(x)$. Now, the unit normal satisfies the equation $|n(x)|^2 = 1$ from which differentiating both sides, we obtain $(\nabla n)n = 0$. Thus, Eq. (18) reduces to

$$\begin{aligned}
P'_{GMC}(\Omega)(\theta) = &\int_{\partial\Omega} \theta(s)\cdot n(s)H(s)[(d\cdot n(s))^-]^2 ds \\
&+\int_{\partial\Omega} 2(d\cdot n(s))^- d\cdot n'(s)(\theta)ds.
\end{aligned} \tag{19}$$

What remains is the calculation of the shape derivative of the unit normal to the boundary. From Lemma 4.8 in [34], we have that the transported of the unit normal $n(\Omega, x)$ is

$$\begin{aligned}
n((Id+\theta)(\Omega), x+\theta(x)) &= \frac{((I+\nabla\theta)^{-1})^T n}{|((I+\nabla\theta)^{-1})^T n|} \\
&= \frac{n - (\nabla\theta)^T n + o(\theta)}{1 - (\nabla\theta)^T n\cdot n + o(\theta)} \\
&= (n - (\nabla\theta)^T n + o(\theta))(1 + (\nabla\theta)^T n\cdot n + o(\theta)) \\
&= n(\Omega, x) - (\nabla\theta)^T n + ((\nabla\theta)^T n\cdot n)n + o(\theta),
\end{aligned}$$

and so the Lagrangian shape derivative of the unit normal is

$$Y(\theta, x) = -(\nabla\theta)^T n + ((\nabla\theta)^T n\cdot n)n.$$

Since by the Hadamard structure theorem [2, 27, 34, 39], the shape derivative in the direction θ depends only on the normal component $\theta\cdot n$ on the boundary $\partial\Omega$, we can restrict our attention to a vector field $\theta(x)$ of the form $\theta(x) = w(x)n(x)$, where w is any scalar function. In such a case, we find that $(\nabla n)^T\theta = w(x)(\nabla n)^T n = 0$ and thus

$$\begin{aligned} Y(\theta, x) &= -(\nabla\theta)^T n - (\nabla n)^T\theta + ((\nabla\theta)^T n\cdot n)n + ((\nabla n)^T\theta\cdot n)n \\ &= -\nabla(\theta\cdot n) + \big[n\cdot\nabla(\theta\cdot n)\big]n \\ &= -\nabla_s(\theta\cdot n) \\ &= -\nabla_s(w(x)). \end{aligned}$$

The Eulerian shape derivative of the unit normal reads

$$n'(x)(\theta) = Y(\theta, x) - \nabla n\theta(x) = Y(\theta, x) = -\nabla_s(w(x)).$$

The same result was found in [28], using similar variational principles. In view of the above results, Eq. (19) becomes

$$P'_{GMC}(\Omega)(\theta) = \int_{\partial\Omega} w(s)H(s)[(d\cdot n(s))^-]^2 ds - \int_{\partial\Omega} 2(d\cdot n(s))^- d\cdot\nabla_s w(s)ds.$$

On the other hand, using the identity (see [28])

$$\int_{\partial\Omega} a\cdot\nabla_s b ds + \int_{\partial\Omega} (\nabla_s\cdot a)b\,ds = \int_{\partial\Omega} a\cdot n\,H\,b\,ds,$$

where a is a vector field and b is a scalar field, we deduce

$$\begin{aligned} P'_{GMC}(\Omega)(\theta) &= \int_{\partial\Omega} w(s)\big(\nabla_s\cdot(2(d\cdot n(s))^- d) - H(s)[(d\cdot n(s))^-]^2\big)ds \\ &= \int_{\partial\Omega} w(s)\big(2d\cdot\nabla_s(d\cdot n(s))^- - H(s)[(d\cdot n(s))^-]^2\big)ds, \end{aligned}$$

which completes the proof.

Lemma 1. *The shape derivative (17) can also be written in the form*

$$P'_{GMC}(\Omega)(\theta) = \int_{\partial\Omega} \theta(s)\cdot n(s)\left(\sum_{i=1}^{N-1}\kappa_i(s)(d\cdot e_i(s))^2(-sign((d\cdot n(s))^-)) - H(s)[(d\cdot n(s))^-]^2\right)ds,$$

where κ_i are the principal curvatures of $\partial\Omega$ at a point $s\in\partial\Omega$ and e_i the associated principal curvature directions ($i = 1, \dots, N-1$).

Proof. For a point $s \in \partial\Omega$ we can write $\nabla_s n(s)$ in the form (see [11]):

$$\nabla_s n(s) = \sum_{i=1}^{N-1} \kappa_i(s) e_i(s) \otimes e_i(s). \tag{20}$$

Substituting (20) in (17) yields the desired result.

We now switch to the derivation of the other constraints (11) and (12), which are pointwise constraints of the same type as the minimum thickness constraint in [13]. Therefore, the same steps need to be followed for their shape derivation and the final extraction of a descent direction. For the sake of completeness, let us mention once more the basic steps of this procedure.

For constraint (11) we formulate a penalty functional of the form

$$P_{GMC}(\Omega) = \sum_{i=1}^{m} \int_{\Gamma_i} \int_0^{diam(D)} \left[\left(d_\Omega \left(s + \xi d_i \right) \right)^- \right]^2 d\xi ds,$$

while for constraint (12), it reads

$$P_{GMC}(\Omega) = \int_{\partial\Omega} \int_0^{diam(D)} \left[\left(d_\Omega \left(s + \xi sign(n(s) \cdot d) d \right) \right)^- \right]^2 d\xi ds.$$

The two functionals are of the same type and can be written in compact notation (see Fig. 7)

$$P_{GMC}(\Omega) = \int_{\partial\Omega} \int_0^{diam(D)} \left[\left(d_\Omega \left(x_m \right) \right)^- \right]^2 d\xi dx, \tag{21}$$

where x_m denotes an offset point of the boundary which is either $x_m = x + \xi d_i$ or $x_m = x + \xi sign(n(x) \cdot d)d$.

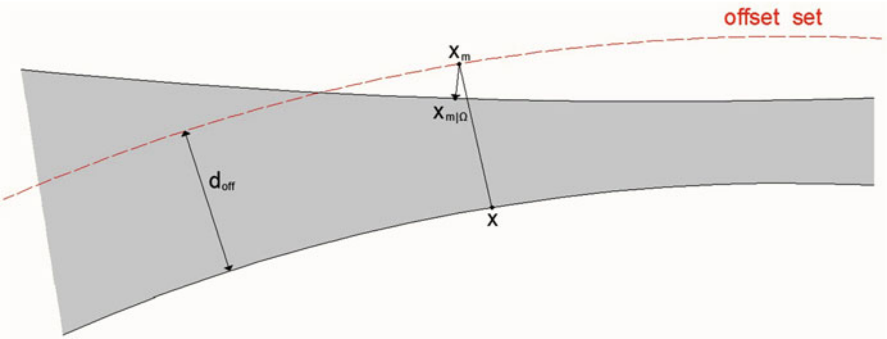


Fig. 7 Offset point x_m and its projection $x_{m|\Omega}$ on the boundary

The following result was obtained in [13] and we recall it without proof.

Proposition 2. *The shape derivative of (21) reads*

$$P'_{GMC}(\Omega)(\theta) =$$
$$\int_{\partial\Omega}\int_0^{diam(D)} \theta\,(x)\cdot n\,(x)\left[H\left((d_\Omega\,(x_m))^-\right)^2 + 2\left((d_\Omega\,(x_m))^-\right)\nabla d_\Omega\,(x_m)\cdot\nabla d_\Omega\,(x)\right] d\xi dx$$
$$-\int_{\partial\Omega}\int_0^{diam(D)} \theta\left(x_{m|\Omega}\right)\cdot n\left(x_{m|\Omega}\right) 2\,(d_\Omega\,(x_m))^-\,d\xi dx,$$

where $x_{m|\Omega}$ is the orthogonal projection of x_m on $\partial\Omega$.

The advantage of the formula of Proposition 2, compared to that of Proposition 1, is that it does not contain any tangential derivative of the normal, or equivalently principal curvatures, which are notably hard to compute with great accuracy.

Remark 1. As already noticed in [13], a descent direction can be found in a second step, after identifying the linear form of the shape derivative with another scalar product. The idea is similar to that of regularization, as described in [20]. More precisely, solving the variational formulation

$$\int_D \left(\alpha_{reg}^2\nabla Q\cdot\nabla v + Qv\right) dx = P'(\Omega)(v) \quad \forall v\in H^1(D), \tag{22}$$

where $\alpha_{reg} > 0$ is a positive number (of the order of the mesh size) which controls the regularization width, yields a solution $Q \in H^1(D)$. Then, choosing a vector field $\theta = -Q\,n$, we obtain a guaranteed descent direction for P_{GMC} since, taking $v = -Q$ in (22), we get

$$P'_{GMC}(\Omega)(-Q\,n) = -\int_D \left(\alpha_{reg}^2|\nabla Q|^2 + Q^2\right) dx.$$

6.3 Derivative of the Uniform Cross-section Constraint

For the constraint (13), a quadratic penalty functional reads

$$P_{UCS}(\Omega) = \int_{\partial\Omega\setminus\partial D} (d\cdot n(s))^2 ds, \tag{23}$$

which highly resembles (16) and thus its shape derivation is omitted here.

If we work in the feasible set of shapes which are constant in the direction d, it is even simpler to take into account this constraint. We consider shapes $\Omega = S\times[0, h]$ where S is a surface perpendicular to d (see Fig. 6). In this case, we force the advection velocity to be zero along the direction d of uniform thickness. Starting from the general formula of a shape derivative

$$J'(\Omega)(\theta) = \int_{\partial\Omega} V(s)\theta(s)\cdot n(s)ds,$$

where $\partial\Omega = \partial S \times [0, h]$, we use Fubini's theorem for the shape derivative

$$J'(\Omega)(\theta) = \int_{\partial S}\int_0^h V(\xi)\theta(s)\cdot n(s)d\xi ds = \int_{\partial S}\theta(s)\cdot n(s)\int_0^h V(\xi)d\xi ds.$$

From this, a descent direction is revealed for the uniform cross-section optimizable boundary ∂S with the choice

$$\begin{cases} \theta(s) = -n(s)\displaystyle\int_0^h V(\xi)d\xi, & \forall\, s \text{ on } \partial S, \\ \theta(s)\cdot d = 0, & \forall\, s \text{ on } \partial S, \end{cases} \tag{24}$$

where $n(s)$ is the normal to ∂S which satisfies $n\cdot d = 0$. Another, simple way to treat this constraint is through the regularization of the velocity field via Eq. (22). Choosing α_{reg} to be a tensor, instead of a positive scalar, we can smooth the advection field in an anisotropic way. Then, Eq. (22) is rewritten as

$$\int_D \left(\sum_{i=1}^N a_i^2 \frac{\partial Q}{\partial x_i}\frac{\partial v}{\partial x_i} + Qv\right) dx = J'(\Omega)(v) \quad \forall v \in H^1(D), \tag{25}$$

where $\alpha_{reg} = \sum_{i=1}^N a_i e_i \otimes e_i$ is the regularization tensor in the canonical basis $(e_i)_{i=1,\dots,N}$ of $\mathbb{R}^N$. For example, if we want a vector field θ of the type of (24) with $d = e_2$, we can set $a_2 >> a_i, i \neq 2$. Then, the solution Q of (25) will be constant in the x_2 direction and the descent direction $\theta = -Q\, n(s)$, where $n(s)$ is the normal to ∂S, satisfies $\theta\cdot d = 0$. In other words, starting from an initial shape that respects the constraint and regularizing the advection field in the way just described, we obtain a final optimized shape with a uniform cross-section.

7 Numerical Examples

We have coded all numerical examples herein in the finite element software SYSTUS of ESI-Group [40]. A quadrangular mesh has been used both for the solution of the elasticity system and for the level-set function. For the elasticity analysis, Q1 finite elements have been used, the Young modulus E is normalized to 1, and the Poisson ratio ν is set to 0.3. The "ersatz material" is considered to have the same Poisson ratio, while its Young modulus is set to 10^{-3}.

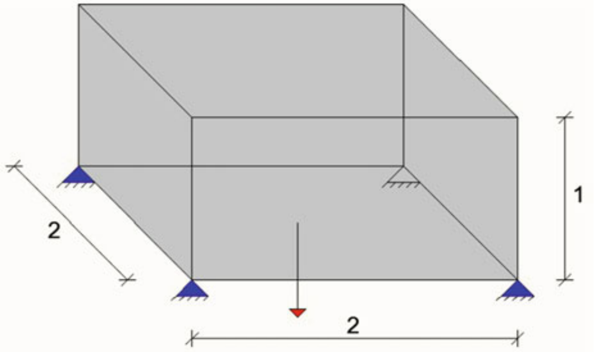


Fig. 8 Boundary and loading conditions for a 3d box

7.1 Molding Direction

The three-dimensional box-like structure of Fig. 8 is our test case to apply several molding direction constraints and compare the corresponding optimized shapes. The entire domain is used for the analysis and is discretized using $40 \times 40 \times 20$ $Q1$ elements. We minimize the compliance under an equality constraint for the volume. The optimization problem reads

$$\begin{aligned} &\min_{\Omega \in \mathscr{U}_{ad}} \int_{\partial\Omega} g \cdot u ds \\ &\text{s.t.} \int_{\Omega} dx = a_V |D|, \end{aligned} \tag{26}$$

where u is the solution of (7) and $a_V \in [0, 1]$ determines the final volume of the structure as percentage of the volume of the working domain D. An augmented Lagrangian method is used here to enforce the constraints, as in our previous work [13]. We refer to [7] for the formula of the shape derivative for the compliance.

At a first step, we impose no molding constraint and solve the optimization problem (26) for $a_V = 0.2$ using the arbitrary initialization of Fig. 9a. The optimized shape after 250 iterations is shown in Fig. 9b.

Let us now solve the same optimization problem for a cast part that must comply with a predefined molding system. For example, if we want to use one mold in the design domain D, remove it in the direction $d = (0, 0, 1)$ and impose the plane $z = 0$ to be a possible parting surface, then obviously the shape in Fig. 9b is no more feasible. Of course, we cannot hope that starting with a different, even much simpler, initialization, we can obtain a castable optimized shape without enforcing a molding direction constraint.

As we have mentioned before, in the absence of thickness constraints, we believe that the method of Xia et al. [45], as described in Sect. 5.1, gives quite satisfying results with a very simple implementation. The only restriction is that the initial

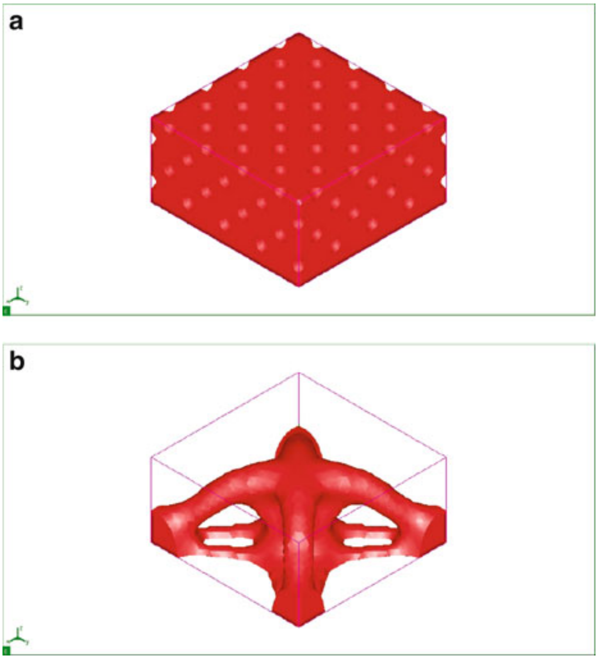


Fig. 9 Initialization (*top*, **a**) and optimized shape (*bottom*, **b**) for the optimization problem (26) without a molding constraint

design must satisfy the constraint. For later iterations it is enough to impose the molding direction condition on the design velocity. Starting with a full-domain initialization (see Fig. 10a) and taking the initial level-set function equal to the signed distance function to the upper part of the domain, we choose an advection velocity of the type (9), where $d = (0, 0, 1)$, and we obtain the optimized shape of Fig. 10b. A comparison of the performance with the previous test case is proposed in Fig. 11.

More flexibility in shape variations is given if the casting direction is set as $d = (0, 0, 1)$ and no parting surface is imposed. In this case, the design domain D can contain two molds, one removed in the direction d and the second in the opposite direction $(-d)$. The same full-domain initialization, with the signed distance function to the upper part of the domain, could be chosen, but this would unfortunately result in a final system with just one mold. Instead, it is more efficient to take the initial level-set function equal to the signed distance function both to the upper and lower part of the domain. The optimized shape is shown in Fig. 12. A comparison of the performance with a previous test case is proposed in Fig. 13.

As expected, a completely different optimized shape is obtained if we change the casting direction. Separating the molds horizontally, in the direction $d = (1, 0, 0)$ and imposing no specific parting surface, yields the optimized shape of Fig. 14. In both Figs. 12 and 14 we see that topological changes can take place by

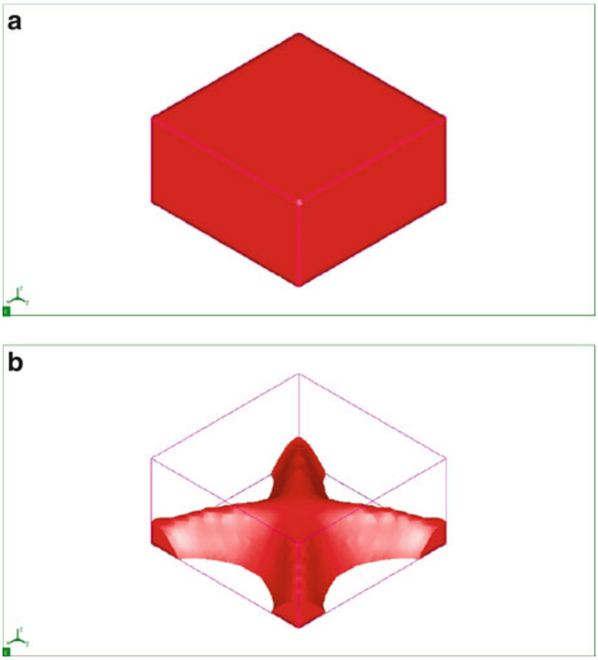


Fig. 10 Initialization (*top*, **a**) and optimized shape (*bottom*, **b**) for the optimization problem (26), setting $d = (0,0,1)$ as casting direction, $z = 0$ as a possible parting surface and using the molding direction condition (9)

"pinching a thin wall" [7], even though we started from a full-domain initialization. A comparison of the performance with a previous test case is proposed in Fig. 15.

7.1.1 Molding Direction and Maximum Thickness

A constraint on the maximum local thickness can be combined with the molding condition on the design velocity without any difficulty a priori. The reason is that the maximum thickness constraint gradient will be of uniform sign, tending always to reduce the thickness (and the volume) of the shape. As we have mentioned in Sect. 5.1, when an advection velocity of the type (9) is chosen, the shape can shrink, but not expand normal to the casting direction. Adding a maximum thickness constraint to the test case of Fig. 10, where the shape is casted along the direction $d = (0,0,1)$ and the plane $z = 0$ is chosen as a possible parting surface, we solve the optimization problem

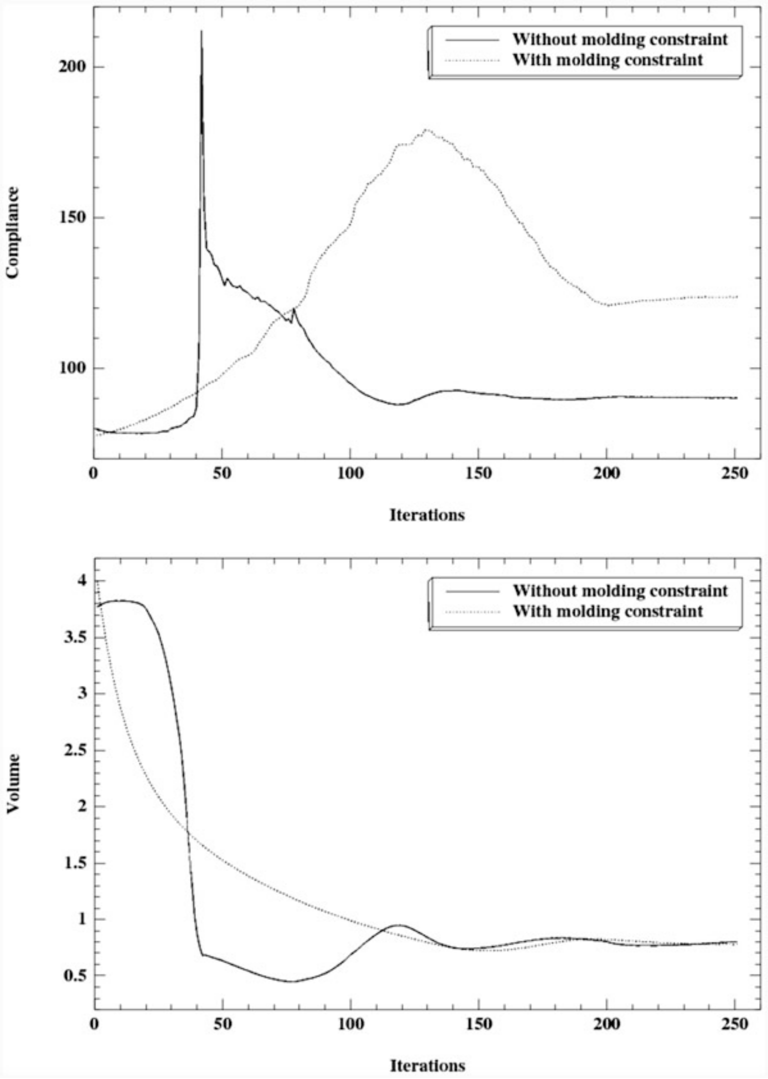


Fig. 11 Compliance (*top*) and volume (*bottom*) convergence histories for the results in Figs. 9 (bottom) and 10 (bottom)

$$
\begin{aligned}
&\min_{\Omega\in\mathcal{U}_{ad}} \int_{\partial\Omega} g\cdot u ds\\
&\quad\text{s.t.} \int_{\Omega} dx = a_V|D|,\\
&\qquad P_{MaxT}(\Omega) = \left(\frac{\displaystyle\int_{\Omega} f(d_\Omega(x))d_\Omega(x)^2 dx}{\displaystyle\int_{\Omega} f(d_\Omega(x))dx}\right)^{\frac{1}{2}} \le d_{max}/2,
\end{aligned}
\tag{27}
$$

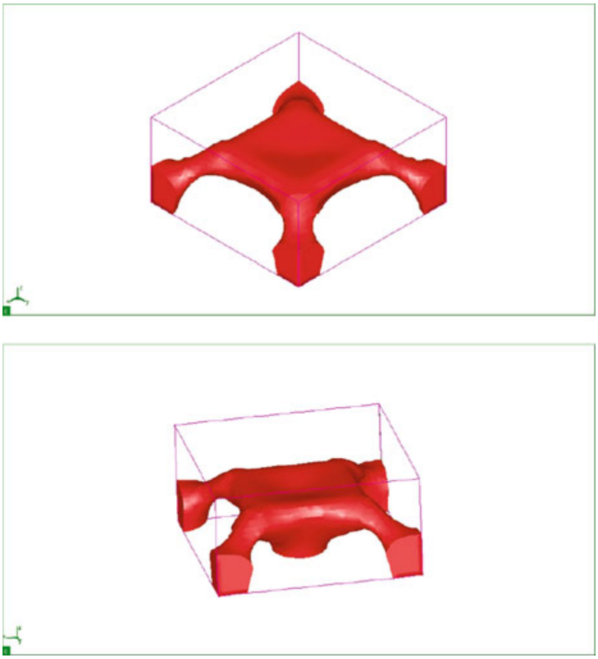


Fig. 12 Plots of the optimized shape for the optimization problem (26), setting $d = (0, 0, 1)$ as casting direction, no a priori defined parting surface and using the molding direction condition (9)

where $d_{max} = 0.2$ and f is a regularization function that reads (see [13]):

$$f(d_\Omega(x)) = 0.5\left(1 + \tanh\left(\frac{|d_\Omega(x)| - d_{max}/2}{\alpha_f d_{max}/2}\right)\right),$$

$\alpha_f > 0$ being a parameter that controls the regularization of the constraint. Using the same initialization as in Fig. 10a and enforcing $z = 0$ as the only possible parting surface, the optimized shape after 250 iterations and the convergence diagrams are shown in Figs. 16 and 17.

7.1.2 Molding Direction and Minimum Thickness

Suppose now that we want to add a minimum thickness constraint with $d_{min} = 0.4$ in the shape of Fig. 10b. The molding condition (9) is no more a suitable method to follow (see Sect. 5.1) and we shall instead combine a minimum thickness constraint with the generalized molding constraint (10). The previously optimized shape is taken as an initial guess to solve the problem

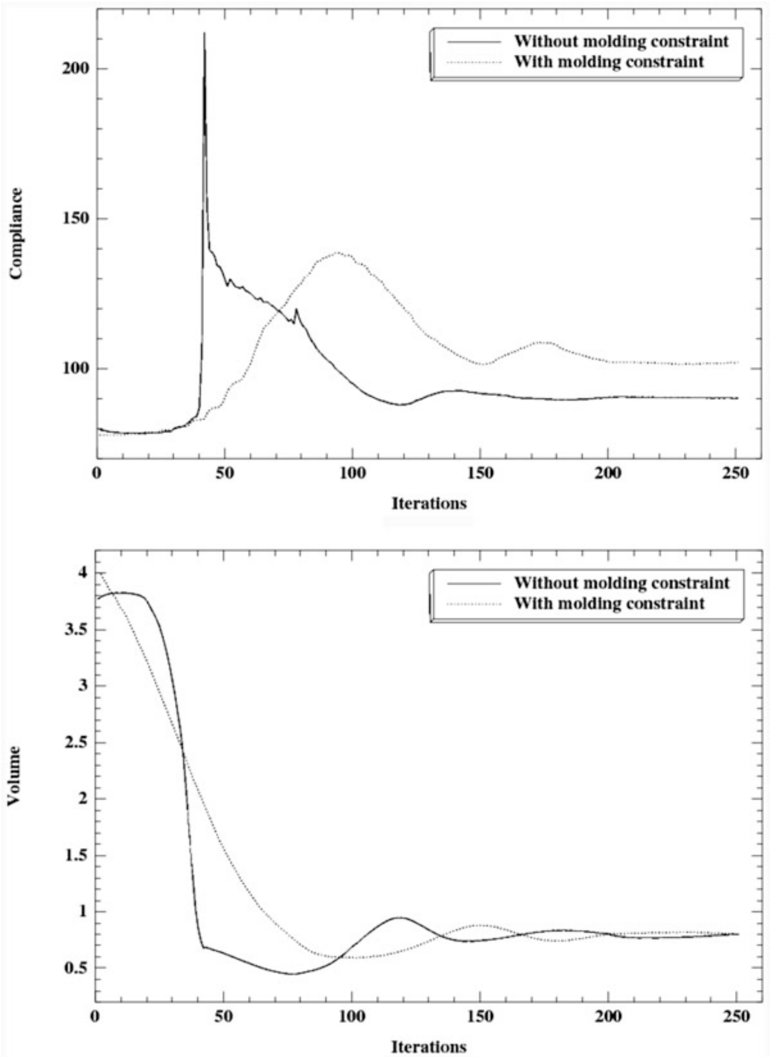


Fig. 13 Compliance (*top*) and volume (*bottom*) convergence diagrams for the results in Figs. 9 (bottom) and 12

$$
\begin{aligned}
&\min_{\Omega\in\mathscr{U}_{ad}} \int_{\partial\Omega} g\cdot u ds\\
&\quad \text{s.t.} \int_{\Omega} dx = a_V|D|,\\
&\quad P_1(\Omega) = P_{MinT}(\Omega) = \int_{\partial\Omega}\int_0^{d_{min}} \left[\left(d_\Omega\left(s-\xi n\left(s\right)\right)\right)^+\right]^2 d\xi ds = 0,\\
&\quad P_2(\Omega) = P_{GMC}(\Omega) = \int_{\partial\Omega}\int_0^{diam(D)} \left[\left(d_\Omega\left(s+\xi d\right)\right)^-\right]^2 d\xi ds = 0,
\end{aligned}
\qquad (28)
$$

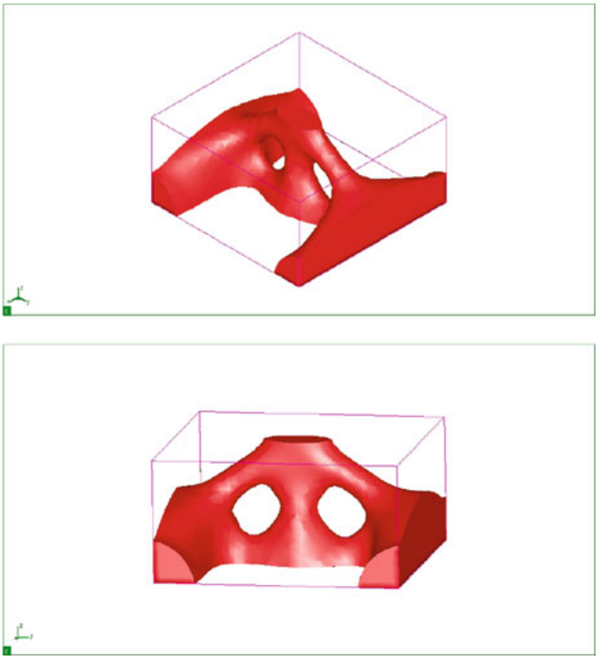


Fig. 14 Plots of the optimized shape for the optimization problem (26), setting $d = (1, 0, 0)$ as casting direction, no a priori defined parting surface and using the molding direction condition (9)

without any condition on the advection velocity. An optimized shape for the optimization problem (28) is shown in Fig. 18b. The convergence diagrams for the penalty functionals P_1 and P_2 are shown in Fig. 19.

We now switch to a minimum thickness constraint of $d_{min} = 0.3$ and to the case of two mold in the z-direction, as applied to the shape of Fig. 12. In this case, the optimization problem reads

$$\begin{aligned} \min_{\Omega \in \mathscr{U}_{ad}} & \int_{\partial\Omega} g \cdot u ds \\ \text{s.t.} & \int_{\Omega} dx = a_V |D|, \\ & P_1(\Omega) = P_{MinT}(\Omega) = \int_{\partial\Omega} \int_0^{d_{min}} \left[\left(d_\Omega \left(s - \xi n \left(s \right) \right) \right)^+ \right]^2 d\xi ds = 0, \\ & P_2(\Omega) = P_{GMC}(\Omega) = \int_{\partial\Omega} \int_0^{diam(D)} \left[\left(d_\Omega \left(s + \xi sign(n \cdot d) d \right) \right)^- \right]^2 d\xi ds = 0 \end{aligned} \tag{29}$$

Note that the constraint $P_2(\Omega)$ is different from that in (28). We obtain the optimized shape of Fig. 20. The convergence diagrams for the penalty functionals P_1 and P_2 are shown in Fig. 21. Table 1 gives the values of the optimal compliances for the test cases of Section 7.1.

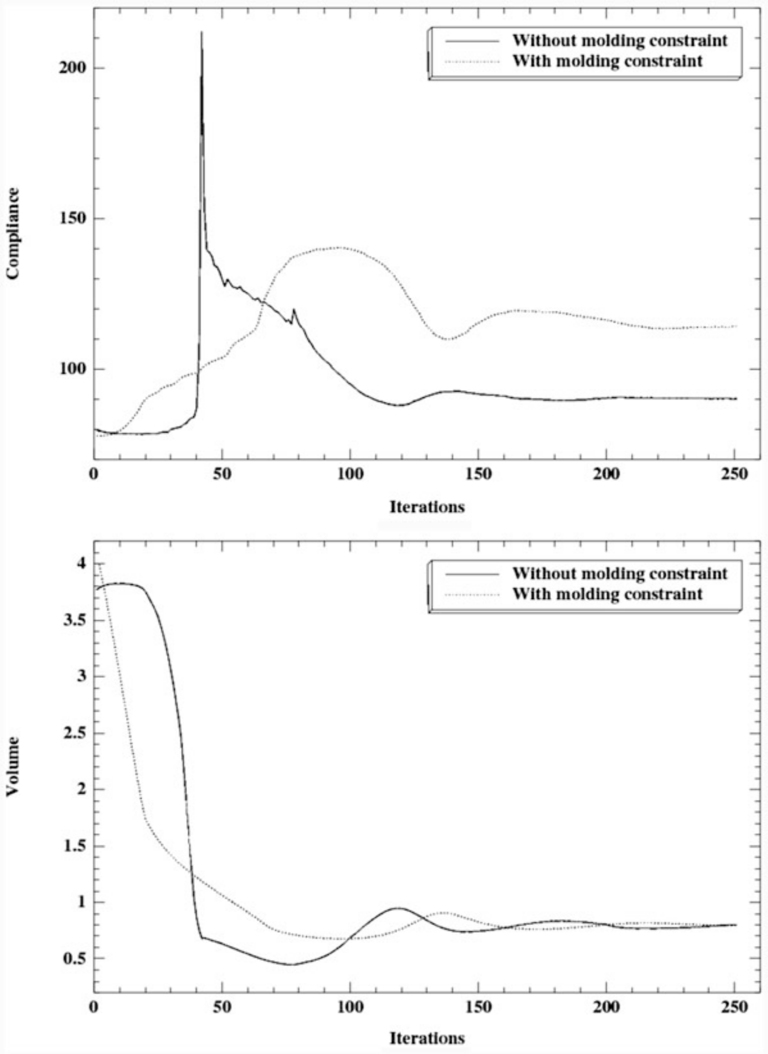


Fig. 15 Compliance (*top*) and volume (*bottom*) convergence diagrams for the results in Figs. 9 (bottom) and 14

7.2 Uniform Cross-section

The $2 \times 0.5 \times 1$ three-dimensional cantilever of Fig. 22, discretized by $40 \times 10 \times 20$ $Q1$ elements, is chosen as test case to apply the uniform cross-section surface constraint. It is clamped on one side and, at the middle of its opposite side, a unitary vertical load is applied. At a first step, problem (26) is solved for $a_V = 0.25$ without

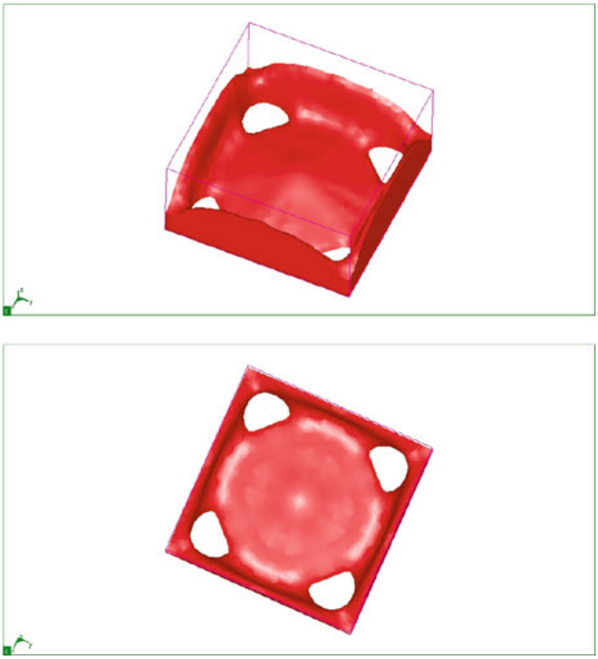


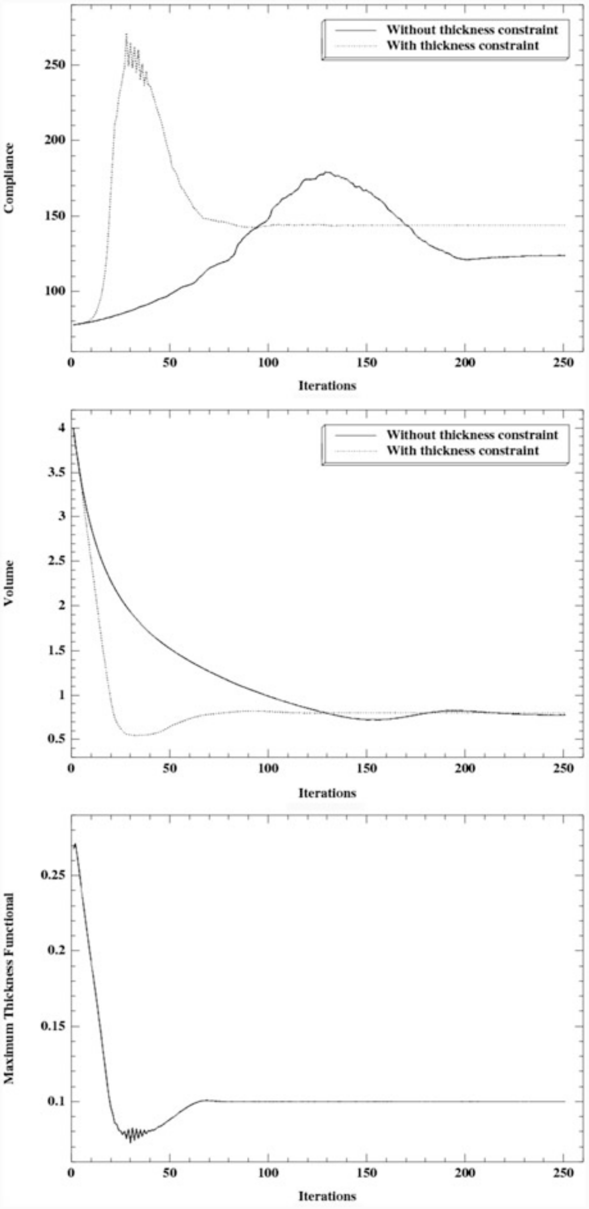
Fig. 16 Optimized shape for the optimization problem (27), with a maximum thickness constraint, setting $d = (0, 0, 1)$ as casting direction and using the molding direction condition (9)

imposing any further geometric constraint on the shape. Starting from the arbitrarily perforated shape of Fig. 23a, we obtain after 200 iterations the optimized shape of Fig. 23b.

We now look for an optimized shape with a uniform cross-section along the y-axis. Starting from the initial shape of Fig. 24a, which has five uniform holes along this direction, we regularize at each iteration the velocity field for the advection of the shape in an anisotropic way, based on Eq. (25), with a much higher regularization coefficient in the y-direction ($a_y >> a_x, a_z$). In our example, $a_x = a_z = 2\Delta x$, Δx being the uniform mesh size, has been used to regularize the advection velocity in a small region around the shape boundary in the direction of the x- and z-axis, while $a_y = \sqrt{10}$ has been set to create a uniform velocity along this direction. The optimized shape is shown in Fig. 24b. The convergence diagrams for the compliance and the volume are shown in Fig. 25.

Acknowledgements The authors acknowledge fruitful discussions and helpful remarks from Marc Albertelli (Renault) and Charles Dapogny (Laboratoire Jean Kuntzmann, CNRS, Université de Grenoble). This work has been supported by the RODIN project (FUI AAP 13). G. Allaire is a member of the DEFI project at INRIA Saclay Ile-de-France.

Fig. 17 Compliance (*top*) and volume (*middle*) convergence diagrams for the results in Figs. 10 (bottom) and 16; convergence diagram for the maximum thickness functional (*bottom*) $P_{MaxT}(\Omega)$ for the optimized shape in Fig. 16



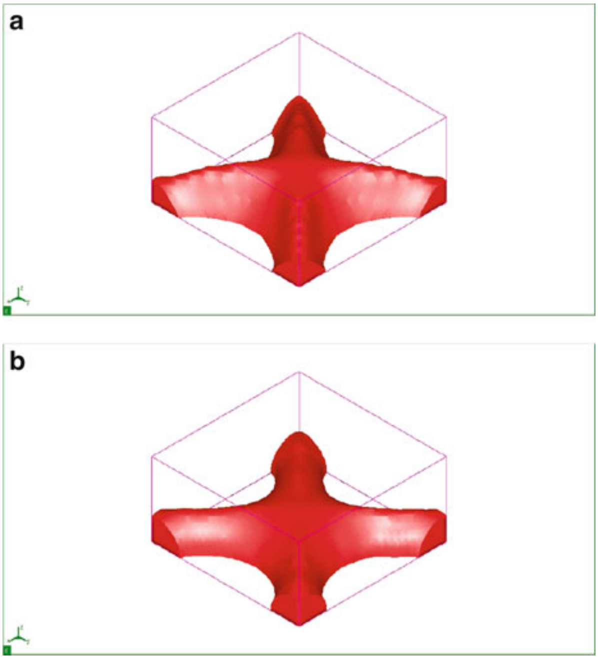


Fig. 18 Optimized shapes under a molding constraint (*top*, **a**) and a molding and minimum thickness constraint (*bottom*, **b**), with a predefined parting surface at $z = 0$

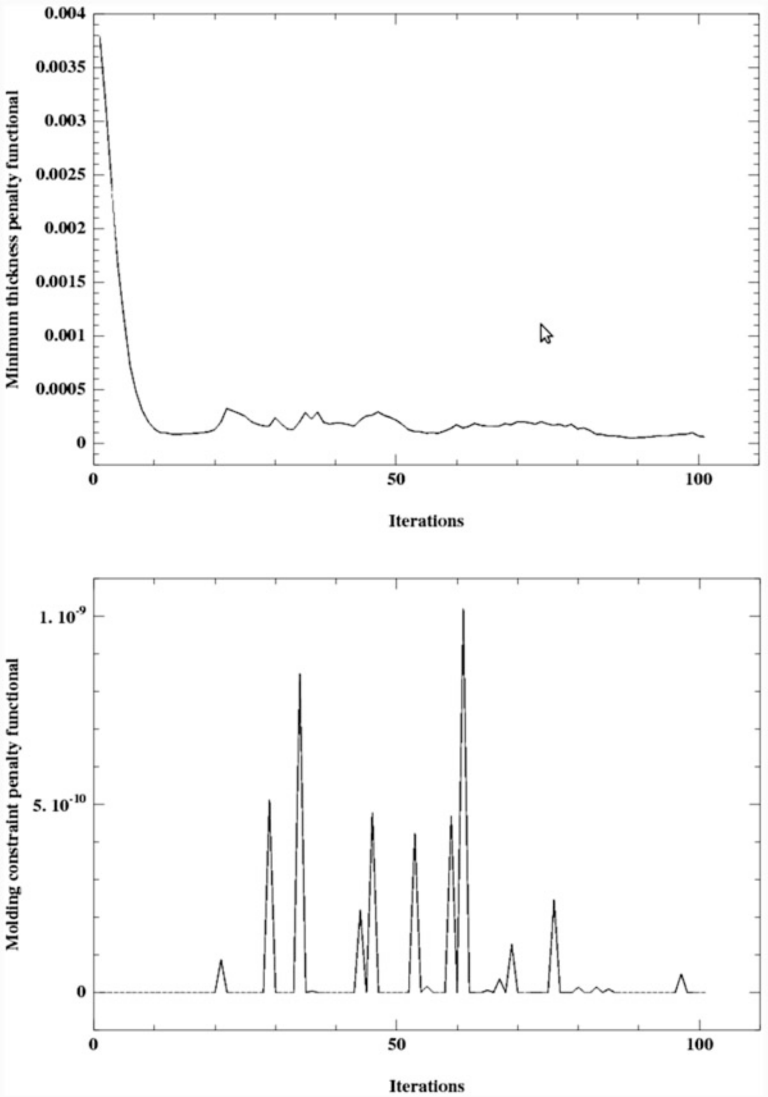


Fig. 19 Convergence diagrams for the penalty functionals: P_1 (*top*) and P_2 (*bottom*), for the results of Fig. 18 (bottom)

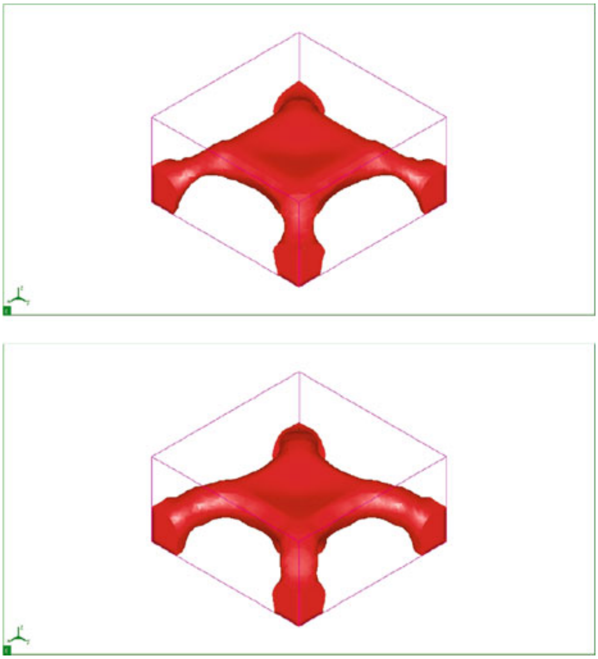


Fig. 20 Optimized shapes under a molding constraint (*top*) and a molding and minimum thickness constraint (*bottom*), without a predefined parting surface

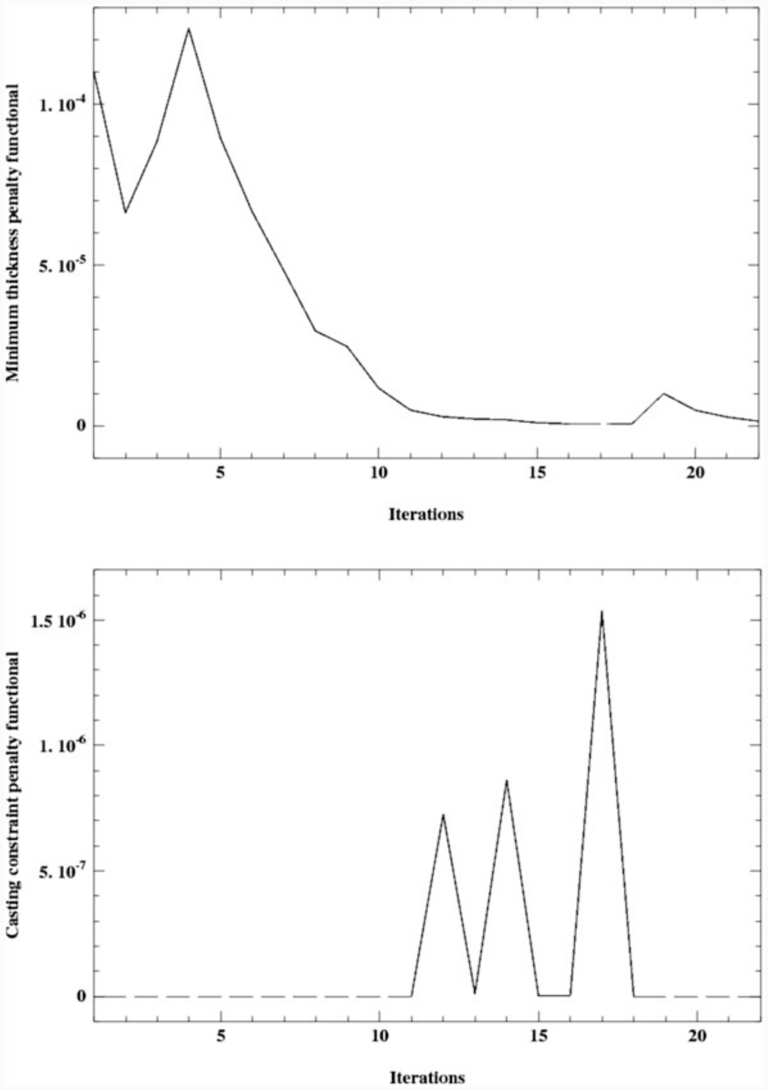


Fig. 21 Convergence diagrams for the penalty functionals: P_1 (*top*) and P_2 (*bottom*), for the results of Fig. 20 (bottom)

Table 1 Compliance of the optimized structures

	Compliance
Without molding constraint [Fig. 9 (bottom)]	90.14
With casting direction $d = (0, 0, 1)$ and no parting surface (Fig. 12)	102.07
With casting direction $d = (0, 0, 1)$, no parting surface, and minimum thickness constraint [Fig. 20 (bottom)]	105.87
With casting direction $d = (1, 0, 0)$ and no parting surface (Fig. 14)	114.13
With casting direction $d = (0, 0, 1)$ and parting surface at $z = 0$ [Fig. 10 (bottom)]	123.68
With casting direction $d = (0, 0, 1)$, parting surface at $z = 0$, and minimum thickness constraint [Fig. 18 (bottom)]	134.68
With casting direction $d = (0, 0, 1)$, parting surface at $z = 0$, and maximum thickness constraint (Fig. 16)	143.65

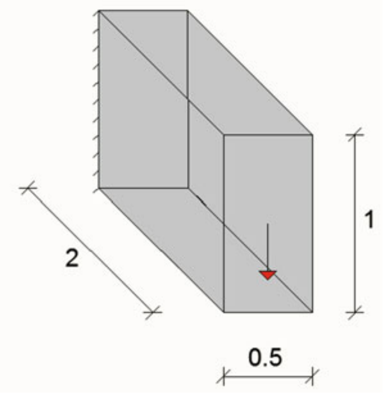


Fig. 22 Boundary conditions for the "uniform cross-section" test case

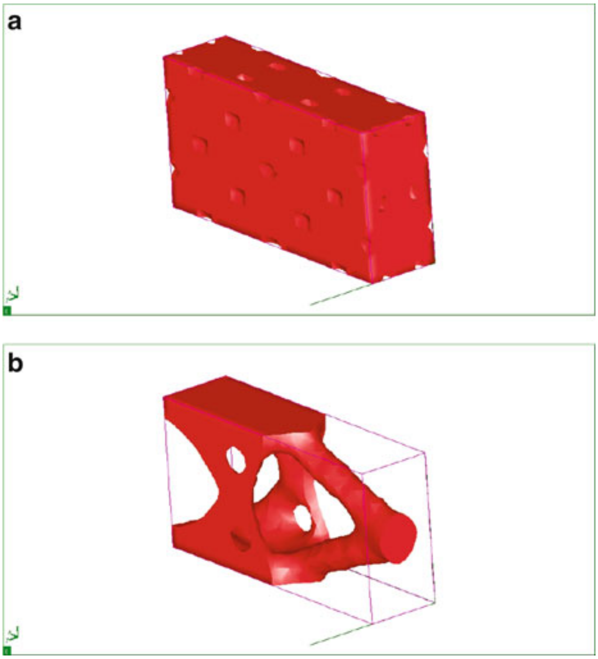


Fig. 23 Initialization (*top*, **a**) and optimized shape (*bottom*, **b**), without the “uniform cross-section” constraint

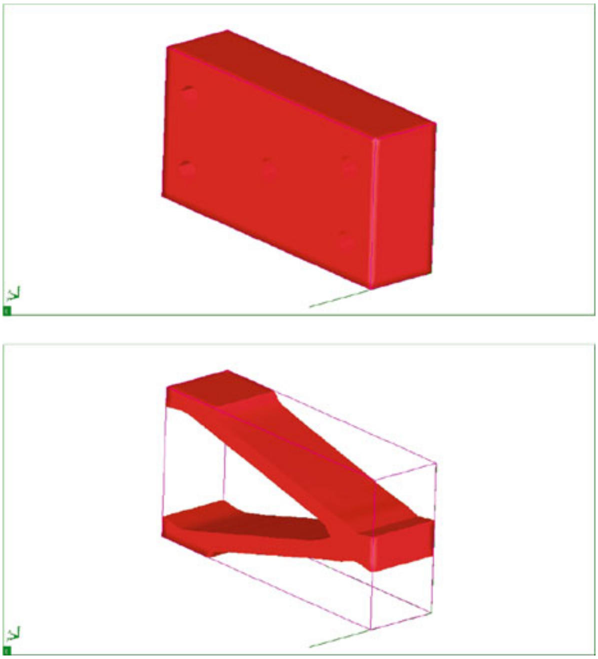


Fig. 24 Initialization (*top*) and optimized shape (*bottom*), with a “uniform cross-section” constraint

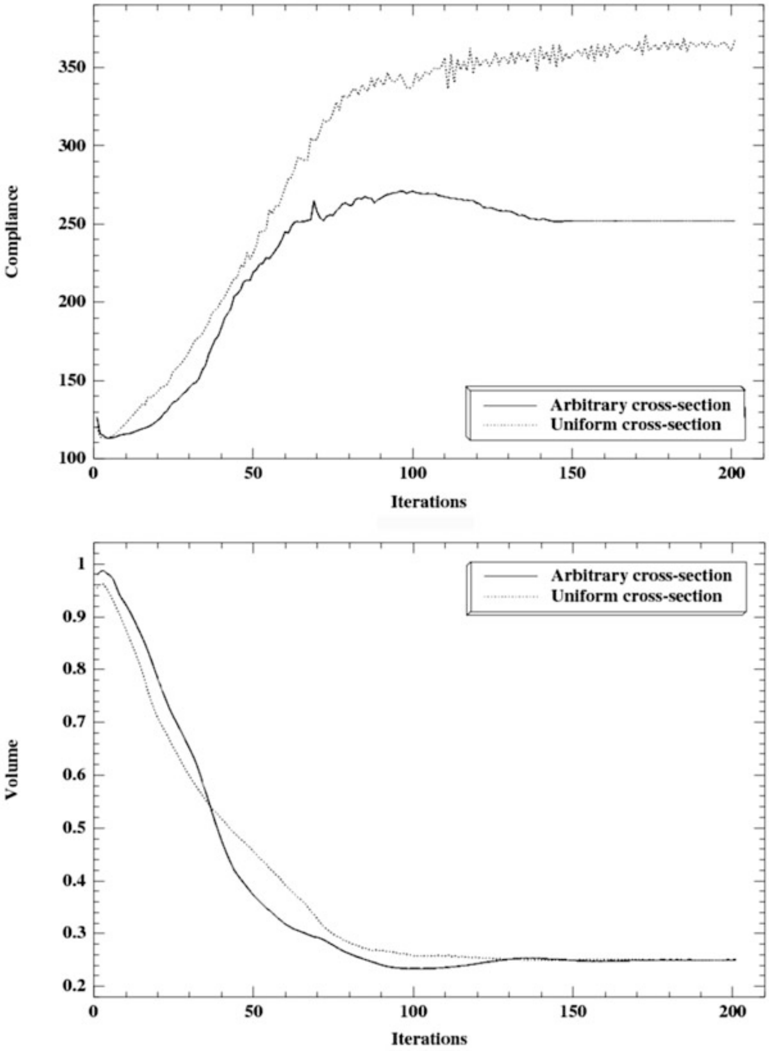


Fig. 25 Compliance (*top*) and volume (*bottom*) convergence diagrams for the results of Figs. 23 and 24

References

1. Ahn, H.K., De Berg, M., Bose, P., Cheng, S.W., Halperin, D., Matoušek, J., Schwarzkopf, O.: Separating an object from its cast. Comput. Aided Des. **34**(8), 547–559 (2002)
2. Allaire, G.: Conception Optimale de Structures. Mathématiques & Applications, vol. 58. Springer, Berlin (2007)
3. Allaire, G., Jouve, F.: A level-set method for vibration and multiple loads structural optimization. Comput. Methods Appl. Mech. Eng. **194**(30), 3269–3290 (2005)
4. Allaire, G., Kohn, R.V.: Optimal design for minimum weight and compliance in plane stress using extremal microstructures. Eur. J. Mech. A: Solids **12**(6), 839–878 (1993)
5. Allaire, G., Bonnetier, E., Francfort, G., Jouve, F.: Shape optimization by the homogenization method. Numer. Math. **76**(1), 27–68 (1997)
6. Allaire, G., Jouve, F., Toader, A.-M.: A level-set method for shape optimization. C. R. Acad. Sci. Paris, Série I **334**, 1125–1130 (2002)
7. Allaire, G., Jouve, F., Toader, A.-M.: Structural optimization using sensitivity analysis and a level-set method. J. Comput. Phys. **194**(1), 363–393 (2004)
8. Allaire, G., Dapogny, C., Frey, P.: Topology and geometry optimization of elastic structures by exact deformation of simplicial mesh. C. R. Math. **349**(17), 999–1003 (2011)
9. Allaire, G., Jouve, F., Van Goethem, N.: Damage and fracture evolution in brittle materials by shape optimization methods. J. Comput. Phys. **230**(12), 5010–5044 (2011)
10. Allaire, G., Jouve, F., Michailidis, G.: Casting constraints in structural optimization via a level-set method. In: WCSMO-10, Orlando, FL (2013)
11. Allaire, G., Dapogny, C., Delgado, G., Michailidis, G.: Multi-phase structural optimization via a level-set method. ESAIM Control Optim. Calc. Var. **20**, 576–611 (2014)
12. Allaire, G., Dapogny, C., Frey, P.: Shape optimization with a level set based mesh evolution method. Comput. Methods Appl. Mech. Eng. **282**, 22–53 (2014)
13. Allaire, G., Jouve, F., Michailidis, G.: Thickness control in structural optimization via a level set method. Struct. Multidiscip. Optim. **53**, 1349–1382 (2016). HALipriprint: hal-00985000, version 1 (April 2014)
14. Bendsøe, M.P.: Optimal shape design as a material distribution problem. Struct. Multidiscip. Optim. **1**(4), 193–202 (1989)
15. Bendsøe, M.P., Kikuchi, N.: Generating optimal topologies in structural design using a homogenization method. Comput. Methods Appl. Mech. Eng. **71**(2), 197–224 (1988)
16. Bendsoe, M.P., Sigmund, O.: Topology Optimization: Theory, Methods and Applications. Springer, Berlin (2004)
17. Blank, L., Farshbaf-Shaker, H., Garcke, H., Styles, V.: Relating phase field and sharp interface approaches to structural topology optimization. ESAIM Control Optim. Calc. Var. **20**(4), 1025–1058 (2014)
18. Bourdin, B., Chambolle, A.: Design-dependent loads in topology optimization. ESAIM Control Optim. Calc. Var. **9**, 19–48 (2003)
19. Campbell, J.: Castings, 1991. Butterworth Heinemann, Great Britain (1964)
20. de Gournay, F.: Velocity extension for the level-set method and multiple eigenvalues in shape optimization. SIAM J. Control. Optim. **45**(1), 343–367 (2006)
21. de Gournay, F., Allaire, G., Jouve, F.: Shape and topology optimization of the robust compliance via the level set method. ESAIM Control Optim. Calc. Var. **14**(01), 43–70 (2008)
22. Fu, M.W., Nee, A.Y.C., Fuh, J.Y.H.: The application of surface visibility and moldability to parting line generation. Comput. Aided Des. **34**(6), 469–480 (2002)
23. Harzheim, L., Graf, G.: Optimization of engineering components with the SKO method. SAE Technical Paper, 951104 (1995)
24. Harzheim, L., Graf, G.: Topshape: an attempt to create design proposals including manufacturing constraints. Int. J. Veh. Des. **28**(4), 389–409 (2002)

25. Harzheim, L., Graf, G.: A review of optimization of cast parts using topology optimization: Ii - topology optimization without manufacturing constraints. Struct. Multidiscip. Optim. **30**(5), 491–497 (2005)
26. Harzheim, L., Graf, G.: A review of optimization of cast parts using topology optimization: I - topology optimization with manufacturing constraints. Struct. Multidiscip. Optim. **31**(5), 388–399 (2006)
27. Henrot, A., Pierre, M.: Variation et optimisation de formes: une analyse géométrique, vol. 48. Springer, Berlin (2005)
28. Laadhari, A., Misbah, C., Saramito, P.: On the equilibrium equation for a generalized biological membrane energy by using a shape optimization approach. Physica D **239**(16), 1567–1572 (2010)
29. Leiva, J.P., Watson, B.C., Kosaka, I.: An analytical directional growth topology parameterization to enforce manufacturing requirements. In: Proceedings of 45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Material Conference, Palm Springs, CA (2004)
30. Lewis, R.W., Manzari, M.T., Gethin, D.T.: Thermal optimisation in the sand casting process. Eng. Comput. **18**(3/4), 392–417 (2001)
31. Mattheck, C.: Design and growth rules for biological structures and their application to engineering. Fatigue Fract. Eng. Mater. Struct. **13**(5), 535–550 (1990)
32. Michailidis, G.: Manufacturing constraints and multi-phase shape and topology optimization via a level-set method. Ph.D. thesis, Ecole Polytechnique X (2014). Available at: http://pastel.archives-ouvertes.fr/pastel-00937306
33. Mohammadi, B., Pironneau, O.: Applied Shape Optimization for Fluids, vol. 28. Oxford University Press, Oxford (2001)
34. Murat, F., Simon, J.: Etude de problèmes d'optimal design. In: Optimization Techniques Modeling and Optimization in the Service of Man Part 2, pp. 54–62 (1976)
35. Osher, S.J., Santosa, F.: Level set methods for optimization problems involving geometry and constraints: I. frequencies of a two-density inhomogeneous drum. J. Comput. Phys. **171**(1), 272–288 (2001)
36. Osher, S.J., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
37. Sethian, J.A.: Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press, Cambridge (1999)
38. Sethian, J.A., Wiegmann, A.: Structural boundary design via level set and immersed interface methods. J. Comput. Phys. **163**(2), 489–528 (2000)
39. Sokołowski, J., Zolésio, J.-P.: Introduction to Shape Optimization. Springer, Berlin (1992)
40. SYSTUS2014, SYSWELD2014, User's manual, SYSTUS International (2014)
41. Takezawa, A., Nishiwaki, S., Kitamura, M.: Shape and topology optimization based on the phase field method and sensitivity analysis. J. Comput. Phys. **229**(7), 2697–2718 (2010)
42. Tavakoli, R., Davami, P.: Optimal riser design in sand casting process by topology optimization with SIMP method I: Poisson approximation of nonlinear heat transfer equation. Struct. Multidiscip. Optim. **36**(2), 193–202 (2008)
43. Tortorelli, D.A., Tiller, M.M., Dantzig, J.A.: Optimal design of nonlinear parabolic systems. Part I: fixed spatial domain with applications to process optimization. Comput. Methods Appl. Mech. Eng. **113**(1), 141–155 (1994)
44. Wang, X., Wang, M.Y., Guo, D.: Structural shape and topology optimization in a level-set-based framework of region representation. Struct. Multidiscip. Optim. **27**(1), 1–19 (2004)
45. Xia, Q., Shi, T., Wang, M.Y., Liu, S.: A level set based method for the optimization of cast part. Struct. Multidiscip. Optim. **41**(5), 735–747 (2010)
46. Xia, Q., Shi, T., Wang, M.Y., Liu, S.: Simultaneous optimization of cast part and parting direction using level set method. Struct. Multidiscip. Optim. **44**(6), 751–759 (2011)
47. Xia, Q., Shi, T., Liu, S., Wang, M.Y.: A level set solution to the stress-based structural shape and topology optimization. Comput. Struct. **90–91**, 55–64 (2012)

48. Yamada, T., Izui, K., Nishiwaki, S., Takezawa, A.: A topology optimization method based on the level set method incorporating a fictitious interface energy. Comput. Methods Appl. Mech. Eng. **199**(45), 2876–2891 (2010)
49. Zhou, M., Rozvany, G.I.N.: The COC algorithm, Part II: topological, geometrical and generalized shape optimization. Comput. Methods Appl. Mech. Eng. **89**(1), 309–336 (1991)
50. Zhou, S., Wang, M.Y.: Multimaterial structural topology optimization with a generalized Cahn–Hilliard model of multiphase transition. Struct. Multidiscip. Optim. **33**(2), 89–111 (2007)
51. Zhou, M., Shyy, Y.K., Thomas, H.L.: Topology optimization with manufacturing constraints. In: Proceedings of the 4th World Congress of Structural and Multidisciplinary Optimization (2001)

Adaptive Control for Weakly Minimum Phase Linear Infinite-Dimensional Systems in Hilbert Space Using a Zero Filter

Mark J. Balas and Susan A. Frost

Abstract Given a linear continuous-time infinite-dimensional plant on a Hilbert space and disturbances of known waveform but unknown amplitude and phase, we show that there exists a stabilizing direct model reference adaptive control law with persistent disturbance rejection and robustness properties. The plant is described by a closed, densely defined linear operator that generates a continuous semigroup of bounded operators on the Hilbert space of states. For this paper, the plant will be weakly minimum phase, i.e., there will be a finite number of unstable zeros with real part equal to zero. All other zeros will be exponentially stable.

The central result will show that all errors will converge to a prescribed neighborhood of zero in an infinite-dimensional Hilbert space even though the plant is not truly minimum phase. The result will not require the use of the standard Barbalat's lemma which requires certain signals to be uniformly continuous. This result is used to determine conditions under which a linear infinite-dimensional system can be directly adaptively controlled to follow a reference model. In particular we examine conditions for a set of ideal trajectories to exist for the tracking problem. Our principal result will be that the direct adaptive controller can be compensated with a zero filter for the unstable zeros which will produce the desired robust adaptive control results even though the plant is only weakly minimum phase. Our results are applied to adaptive control of general linear infinite-dimensional systems described by self-adjoint operators with compact resolvent.

Keywords Adaptive control • Infinite dimensional systems • Distributed parameter systems

M.J. Balas (✉)
Aerospace Engineering Department, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
e-mail: balasm@erau.edu

S.A. Frost
Intelligent Systems Division, MS 269-3, NASA Ames Research Center, Moffett Field, CA 94035, USA
e-mail: susan.frost@nasa.gov

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_2

1 Introduction

Many new fields are beginning to seek the benefits of control of systems described by partial differential equations, especially for flexible aerospace structures, smart electric power grids, and the quantum control field [1–3] and [21]. New general results in the theory of control of partial differential equations can be found in [4–6]. In our previous finite-dimensional work [7–10] we have accomplished direct model reference adaptive control and disturbance rejection with very low order adaptive gain laws for MIMO systems. When systems are subjected to an unknown internal delay, these systems are also infinite-dimensional in nature. Direct adaptive control theory can be modified to handle this time delay situation for infinite-dimensional spaces [11]. However, this approach does not handle the situation when partial differential equations (PDEs) describe the open-loop system.

In [12–14] and [24], we considered how to make a linear infinite-dimensional system track the output of a finite-dimensional reference model in a robust fashion in the presence of persistent disturbances. This requires that the transmission zeros of a linear infinite-dimensional system will be exponentially stable. But many systems are not perfectly minimum phase; they may contain a finite number of isolated zeros with real part equal to zero. This paper uses the results on transmission zeros and their relationship to almost strict dissipativity in [14] and [22] to expand the capability of robust direct adaptive control to handle these weakly minimum phase systems.

We will apply this robust theory to general linear PDEs governed by self-adjoint operators with compact resolvent such as linear diffusion systems.

2 Adaptive Robust Tracking with Disturbance Rejection

Let X be an infinite-dimensional separable Hilbert space with *inner product* (x, y) and corresponding norm $\|x\| \equiv \sqrt{(x,x)}$. Also let A be a closed linear operator with domain $D(A)$ dense in X. Consider the *linear infinite-dimensional plant with persistent disturbances:*

$$\begin{cases} \dfrac{\partial x}{\partial t}(t) = Ax(t) + Bu(t) + \Gamma\, u_D(t) + v, \quad x(0) \equiv x_0 \in D(A) \\ Bu \equiv \displaystyle\sum_{i=1}^{m} b_i u_i \\ y(t) = Cx(t), \quad y_i \equiv (c_i, x(t))\,, \quad i = 1 \ldots m \end{cases} \tag{1}$$

where $x \in D(A)$ is the plant state, $b_i \in D(A)$ are actuator influence functions, $c_i \in D(A)$ are sensor influence functions, $u, y \in \Re^m$ are the control input and plant output m-dimensional vectors respectively, and u_D is a disturbance with known basis functions ϕ_D. We assume v is a bounded but unknown disturbance with $\|v\| \leq M_v < \infty$.

In order to accomplish some degree of disturbance rejection in a direct adaptive scheme, we will make use of a definition, given in [11], for persistent disturbances:

Definition 1 *A disturbance vector $u_D \in R^q$ is said to be* ***persistent*** *if it satisfies the* ***disturbance generator equations****:*

$$\begin{cases} u_D(t) = \theta z_D(t) \\ \dot{z}_D(t) = F z_D(t) \end{cases} \text{or} \quad \begin{cases} u_D(t) = \theta z_D(t) \\ z_D(t) = L\phi_D(t) \end{cases} \tag{2}$$

where F is a marginally stable matrix and $\phi_D(t)$ is a vector of known functions forming a basis for all such possible disturbances. This is known as "a disturbance with known waveform but unknown amplitudes."

The *objective of control* in this paper will be to cause the output $y(t)$ of the plant to robustly asymptotically track the output $y_m(t)$ of a linear finite-dimensional reference model given by:

$$\begin{cases} \dot{x}_m = A_m x_m + B_m u_m \\ y_m = C_m x_m; \quad x_m(0) = x_0^m \end{cases} \tag{3}$$

where the reference model state $x_m(t)$ is an $\mathrm{N_m}$-dimensional vector with reference model output $y_m(t)$ having the *same* dimension as the plant output $y(t)$. In general, the plant and reference models need **not** have the same dimensions. The excitation of the reference model is accomplished via the vector $u_m(t)$ which is generated by

$$\dot{u}_m = F_m u_m; \; u_m(0) = u_0^m \tag{4}$$

The reference model parameters (A_m, B_m, C_m, F_m) will be assumed completely known. What is meant by "robust asymptotic tracking" is the following:

We define the **output error vector** to be

$$e_y \equiv y - y_m \underset{t\to\infty}{\to} N(0) \tag{5}$$

where $N(0)$ is a predetermined neighborhood of the vector *0*.

The control objective will be accomplished by a direct a**daptive control law** of the form:

$$u = G_m x_m + G_u u_m + G_e e_y + G_D \phi_D \tag{6a}$$

The direct adaptive controller will have adaptive gains given by:

$$\begin{cases} \dot{G}_u = -aG_u - e_y u_m^* \gamma_u; \gamma_u > 0 \\ \dot{G}_m = -aG_m - e_y x_m^* \gamma_m; \gamma_m > 0 \\ \dot{G}_e = -aG_e - e_y e_y^* \gamma_e; \gamma_e > 0 \\ \dot{G}_D = -aG_D - e_y \phi_D^* \gamma_D; \gamma_D > 0 \end{cases} \tag{6b}$$

3 Ideal Trajectories

We define the i**deal trajectories** for (1) in the following way:

$$\begin{cases} x_* = S_{11}^* x_m + S_{12}^* u_m + S_{13}^* z_D = S_1 z \\ u_* = S_{21}^* x_m + S_{22}^* u_m + S_{23}^* z_D = S_2 z \end{cases} \quad \text{with } z \equiv \begin{bmatrix} x_m \\ u_m \\ z_d \end{bmatrix} \in \Re^L \tag{7}$$

where the **ideal trajectory** $x_*(t)$ is generated by the **ideal control** $u_*(t)$ from

$$\begin{cases} \frac{\partial x_*}{\partial t} = Ax_* + Bu_* + \Gamma u_D \\ \quad y_* = Cx_* = y_m \end{cases} \tag{8}$$

If such ideal trajectories exist, they will be linear combinations of the reference model state, disturbance state, and reference model input (7), and they will produce exact output tracking in a disturbance-free plant (8).

By substitution of (7) into (8) using (3) and (4), we obtain the linear **model matching conditions**:

$$AS_{11}^* + BS_{21}^* = S_{11}^* A_m \tag{9a}$$

$$AS_{12}^* + BS_{22}^* = S_{12}^* F_m + S_{11}^* B_m \tag{9b}$$

$$CS_{11}^* = C_m \tag{9c}$$

$$CS_{12}^* = 0 \tag{9d}$$

$$AS_{13}^* + BS_{23}^* + \Gamma \Theta = S_{13}^* F \tag{9e}$$

$$CS_{13}^* = 0 \tag{9f}$$

The model matching conditions (9a), (9b), (9c), (9d), (9e), and (9f) are *necessary and sufficient* conditions for the existence of the ideal trajectories in the form of (7). These model matching conditions (9a), (9b), (9c), (9d), (9e), and (9f) can be rewritten as:

$$\begin{cases} AS_1 + BS_2 = S_1 L_m + H_1 \\ CS_1 = H_2 \end{cases} \tag{10}$$

where $S_1 \equiv \left[S_{11}^* \; S_{12}^* \; S_{13}^* \right] : \Re^L \to D(A) \subset X$, $S_2 \equiv \left[S_{21}^* \; S_{22}^* \; S_{23}^* \right] : \Re^L \to \Re^m$, $L_m \equiv \begin{bmatrix} A_m & B_m & 0 \\ 0 & F_m & 0 \\ 0 & 0 & F \end{bmatrix}$, and $\begin{cases} H_1 \equiv \left[0 \; 0 \; -\Gamma\theta \right] \\ H_2 \equiv \left[C_m \; 0 \; 0 \right] \end{cases}$. Because (S_1, S_2) are both of finite rank, they are bounded linear operators on their respective domains.

4 Ideal Trajectory Existence and Uniqueness: Normal Form

To determine conditions for the existence and uniqueness of the ideal trajectories, we need two lemmae:

Lemma 1 If CB is nonsingular, then $P_1 \equiv B(CB)^{-1}C$ is a (non-orthogonal) bounded projection onto the *range of* B, $R(B)$, along the *null space of* C, $N(C)$ with $P_2 \equiv I - P_1$ the complementary bounded projection, and $X = R(B) \oplus N(C)$ as well as $D(A) = R(B) \oplus [N(C) \cap D(A)]$.

Now for the above pair of projections $(P_1.P_2)$ we will have

$$\begin{cases} \frac{\partial P_1 x}{\partial t} = P_1 \frac{\partial x}{\partial t} = \underbrace{(P_1 A P_1)}_{A_{11}} P_1 x + \underbrace{(P_1 A P_2)}_{A_{12}} P_2 x + \underbrace{(\; P_1 B \;)}_{B} u \\ \frac{\partial P_2 x}{\partial t} = P_2 \frac{\partial x}{\partial t} = \underbrace{(P_2 A P_1)}_{A_{21}} P_1 x + \underbrace{(P_2 A P_2)}_{A_{22}} P_2 x + \underbrace{(\; P_2 B \;)}_{=0} u \\ y = \underbrace{(\; C P_1 \;)}_{C} P_1 x + \underbrace{(\; C P_2 \;)}_{=0} P_2 x \end{cases}$$

which implies that

$$\begin{cases} \frac{\partial P_1 x}{\partial t} = A_{11} P_1 x + A_{12} P_2 x + B u \\ \frac{\partial P_2 x}{\partial t} = A_{21} P_1 x + A_{22} P_2 x \\ y = C P_1 x = C x \end{cases}$$

because $y = Cx = C\left(B(CB)^{-1}C\right)x = CP_1 x$, $P_1 x = B(CB)^{-1}Cx = B(CB)^{-1}y$, $CP_2 = C - CB(CB)^{-1}C = 0$, and $P_2 B = B - B(CB)^{-1}CB = 0$.

Lemma 2 If CB is nonsingular, then there exists an invertible, bounded linear operator $W \equiv \begin{bmatrix} C \\ W_2 P_2 \end{bmatrix} : X \to \tilde{X} \equiv R(B) x l_2$ such that $\overline{B} \equiv WB = \begin{bmatrix} CB \\ 0 \end{bmatrix}$ and $\overline{C} \equiv CW^{-1} = \left[I_m \; 0 \right]$ and $\overline{A} \equiv WAW^{-1}$.

This coordinate transformation can be used to put (1) into *normal form:*

$$\begin{cases} \dot{y} = \overline{A}_{11}y + \overline{A}_{12}z_2 + CBu \\ \frac{\partial z_2}{\partial t} = \overline{A}_{21}y + \overline{A}_{22}z_2 \end{cases} \tag{11}$$

where the subsystem: $(\bar{A}_{22}, \bar{A}_{12}, \bar{A}_{21})$ is called the *zero dynamics* of (1) and

$$\overline{A}_{11} \equiv CA_{11}B(CB)^{-1} = CAB(CB)^{-1}; \overline{A}_{12} \equiv CA_{12}W_2^* = CAW_2^*$$

$$\overline{A}_{21} \equiv W_2A_{21}B(CB)^{-1}; \overline{A}_{22} \equiv W_2A_{22}W_2^*$$

and $W_2 : X \to l_2$ by $W_2x \equiv \begin{bmatrix} (\theta_1, P_2x) \\ (\theta_2, P_2x) \\ (\theta_3, P_2x) \\ \dots \\ \end{bmatrix}$ is an isometry from $N(C)$ into l_2.

Now we can prove the following theorem about the *existence and uniqueness of ideal trajectories*:

Theorem 1 Assume CB is nonsingular. Then $\sigma(L_m) = \sigma(A_m) \cup \sigma(F_m) \cup \sigma(F) \subset \rho(\overline{A}_{22})$ where $\rho(\overline{A}_{22}) \equiv \left\{\lambda \in C \text{ such that } (\lambda I - \overline{A}_{22})^{-1} : l_2 \to l_2 \text{ is a bounded linear operator}\right\}$ if and only if there exist unique bounded linear operator solutions (S_1, S_2) satisfying the matching conditions (10).

Note that we can also write $\sigma(L_m) \cap \sigma(\overline{A}_{22}) = \phi$ where $\sigma(\overline{A}_{22}) \equiv \left[\rho(\overline{A}_{22})\right]^c$.

5 Transmission Zeros of a Linear Infinite-Dimensional System

It is possible to relate the point spectrum $\sigma_p(\overline{A}_{22}) \equiv \{\lambda \text{ such that } \lambda I - \overline{A}_{22} \text{ is not one to one}\}$ to the set Z of *transmission (or blocking) zeros* of (A, B, C). As in the finite-dimensional case [15] or [23], we can see that $Z \equiv \left\{\lambda \text{ such that } V(\lambda) \equiv \begin{bmatrix} \lambda I - A & B \\ C & 0 \end{bmatrix} : D(A)x\Re^m \to Xx\Re^m \text{ linear operator is not one to one}\right\}$

Lemma 3 $Z = \sigma_p(\overline{A}_{22}) \equiv \{\lambda \text{ such that } \lambda I - \overline{A}_{22} \text{ is not one to one}\}$ is the *point spectrum* of $\bar{A}_{22}$.

Proof of Lemma 3 From $\overline{V}(\lambda) = \begin{bmatrix} \lambda I - \overline{A} & \overline{B} \\ \overline{C} & 0 \end{bmatrix} = \begin{bmatrix} W^{-1} & 0 \\ 0 & I \end{bmatrix} \underbrace{\begin{bmatrix} \lambda I - A & B \\ C & 0 \end{bmatrix}}_{V(\lambda)} \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}$

we obtain $\begin{bmatrix} \lambda I - \overline{A} & \overline{B} \\ \overline{C} & 0 \end{bmatrix}$ not one to one if and only if $\begin{bmatrix} \lambda I - A & B \\ C & 0 \end{bmatrix}$ is not one to one.

But, using the normal form from Lemma 2,

$$\overline{V}(\lambda) \equiv \begin{bmatrix} \lambda I - \overline{A} & \overline{B} \\ \overline{C} & 0 \end{bmatrix} = \begin{bmatrix} \lambda I - \overline{A}_{11} & -\overline{A}_{12} & CB \\ -\overline{A}_{21} & \lambda I - \overline{A}_{22} & 0 \\ I_m & 0 & 0 \end{bmatrix}$$

And therefore $0 = \overline{V}(\lambda) h = \overline{V}(\lambda) \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}$ if and only if $h_1 = 0$, $h_3 = (CB)^{-1}\overline{A}_{12}h_2$, and $\left(\lambda I - \overline{A}_{22}\right) h_2 = 0$. So $h \neq 0$ if and only if $h_2 \neq 0$.

Therefore $\begin{bmatrix} sI - \overline{A} & \overline{B} \\ \overline{C} & 0 \end{bmatrix}$ is not one to one if and only if $\lambda \in \sigma_p \left(\overline{A}_{22}\right)$.

This completes the proof of Lemma 3.

Using Lemma 3 and Theorem 1, we have the following i*nternal model principle*:

Corollary 1 Assume CB is nonsingular and $\sigma\left(\overline{A}_{22}\right) = \sigma_p\left(\overline{A}_{22}\right) = \sigma_p\left(P_2AP_2\right)$ where $\overline{A}_{22} \equiv W_2^* P_2 A P_2 W_2$.

There exist unique bounded linear operator solutions (S_1, S_2) satisfying the matching conditions (10) if and only if $\sigma(L_m) \cap Z = [\sigma(A_m) \cup \sigma(F_m) \cup \sigma(F)] \cap Z = \varphi$, i.e., *no eigenvalues of* (A_m, F_m, F) *can be transmission zeros of* (A, B, C).

Note : $\lambda I - \overline{A}_{22}$ is not 1-1 $\iff \exists x \neq 0 P_2 x \neq 0 \& z_2$

$$= W_2 P_2 x \neq 0 \& \left(\lambda I - \overline{A}_{22}\right) z_2 = 0$$

$$\iff \exists x \neq 0 P_2 x \neq 0 \& 0 = \left(\lambda I - \overline{A}_{22}\right) W_2 P_2 x$$

$$= \left(\lambda \underbrace{W_2 W_2^*}_{I} - W_2 PAP_2 W_2^* \right) W_2 P_2 x$$

$$= \left[W_2 \left(\lambda I - P_2 A P_2\right) W_2^* \right] W_2 P_2 x$$

$$\iff W_2 \left(\lambda I - P_2 A P_2\right) W_2^* \text{ is not 1-1 on } N(C)$$

But W_2 is an isometry on $N(C)$

$$\therefore \sigma_p\left(\overline{A}_{22}\right) = \sigma_p\left(P_2 A P_2\right).$$

6 Stability of the Error System: Almost Strict Dissipativity

The error system can be found from (1) and (8) by first defining $e \equiv x - x_*$ and $\Delta u \equiv u - u_*$. Then we have

$$\begin{cases} \frac{\partial e}{\partial t} = Ae + B\Delta u + v \\ e_y \equiv y - y_m = y - y_* = Ce \end{cases} \tag{12}$$

Now consider the definition of strict dissipativity for infinite-dimensional systems and the general form of this **adaptive error system** to prove stability. The main theorem of this section will later be utilized to assess the convergence and stability of the adaptive controller with disturbance rejection for linear diffusion systems.

Noting that there can be some ambiguity in the literature with the definition of strictly dissipative systems, we modify the suggestion of Wen in [16] for finite-dimensional systems and expand it to include infinite-dimensional systems.

Definition 2 *The triple (A_c, B, C) is said to be* ***strictly dissipative (SD)*** *if A_c is a densely defined, closed operator on $D(A_c) \subseteq X$ a complex Hilbert space with inner product (x, y) and corresponding norm $\|x\| \equiv \sqrt{(x,x)}$ and generates a C_0 semigroup of bounded operators $U(t)$, and (B, C) are bounded finite rank input/output operators with rank M where $B : R^m \to X$ and $C : X \to R^m$. In addition there exist symmetric positive bounded operators P and Q* on *X such that $0 \le p_{\min}\|e\|^2 \le (Pe, e) \le p_{\max}\|e\|^2$; $0 \le q_{\min}\|e\|^2 \le (Qe, e) \le q_{\max}\|e\|^2$, i.e., P and Q are bounded and coercive, and*

$$\begin{cases} \operatorname{Re}(PA_c e, e) \equiv \frac{1}{2}\left[(PA_c e, e) + \overline{(PA_c e, e)}\right] = \frac{1}{2}\left[(PA_c e, e) + (e, PA_c e)\right] \\ \qquad = -(Qe, e) \le -q_{\min}\|e\|^2; \quad e \in D(A_c) \\ \qquad PB = C^* \end{cases} \tag{13}$$

where C^* is the adjoint of the operator C.

We also say that (A, B, C) is *almost strictly dissipative (ASD)* when there exists a $G_* \in \Re^{mxm}$ such that (A_c, B, C) is strictly dissipative with $A_c \equiv A + BG_*C$.

Note that if $P = I$ in (13), by the Lumer–Phillips theorem [17], p. 405, we would have

$$\|U_c(t)\| \le e^{-\sigma t}; \quad t \ge 0; \quad \sigma \equiv q_{\min} > 0$$

Henceforth, we will make the following set of assumptions:

Hypothesis 1 Assume the following:

(i) *There exists a gain, G_e^* such that the triple $\left(A_C \equiv A + BG_e^*C, B, C\right)$ is strictly dissipative, i.e., (A, B, C) is ASD,*
(ii) *A is a densely defined, closed operator on $D(A) \subseteq X$ and generates a C_0 semigroup of bounded operators $U(t)$, and*
(iii) *ϕ_D is bounded.*

From (7), we have $u_* = S^*_{21}x_m + S^*_{22}u_m + S^*_{23}z_D$ and using (6), we obtain

$$\Delta u \equiv u - u_* = \left(G_m x_m + G_u u_m + G_e e_y + G_D \phi_D\right) - \left(S^*_{21}x_m + S^*_{22}u_m + S^*_{23}\underbrace{z_D}_{L\phi_D}\right)$$

$$= G^*_e e_y + \Delta G_e e_y + \left[\Delta G_m \; \Delta G_u \; \Delta G_D\right]\begin{bmatrix} x_m \\ u_m \\ \phi_D \end{bmatrix} = G^*_e e_y + \Delta G \eta$$

where $\Delta G \equiv G - G_*; G \equiv \left[G_e \; G_m \; G_u \; G_D\right]; G_* \equiv \left[G^*_e \; S^*_{21} \; S^*_{22} \; S^*_{23}L\right]$; and

$$\eta \equiv \left[e_y \; x_m \; u_m \; \phi_D\right]^T$$

From (1), (6), (12), and (13), the *error system* becomes

$$\begin{cases} \frac{\partial e}{\partial t} = \left(\underbrace{A + BG^*_e C}_{A_c}\right) e + B\Delta G\eta + v = A_c e + B\rho + v; e \in D(A); \rho \equiv \Delta G\eta \\ e_y = Ce \\ \Delta\dot{G} = \dot{G} - \dot{G}_* = \dot{G} = -e_y \eta^* \gamma; \quad \gamma \equiv \begin{bmatrix} \gamma_e & 0 & 0 & 0 \\ 0 & \gamma_m & 0 & 0 \\ 0 & 0 & \gamma_u & 0 \\ 0 & 0 & 0 & \gamma_D \end{bmatrix} > 0 \end{cases} \tag{14}$$

Since B, C are finite rank operators, so is $BG^*_e C$. Therefore, $A_c \equiv A + BG^*_e C$ with $D(A_c) = D(A)$ generates a C_0 semigroup $U_c(t)$ because A does; see [18] Theorem 2.1 p. 497. Furthermore, by Theorem 8.10 p. 157 in [4], $x(t)$ remains in $D(A)$ and is differentiable there for all $t \geq 0$.This is because $F(t) \equiv B\rho = B\Delta G\eta$ is continuously differentiable in $D(A)$.

We see that (14) is the *feedback interconnection* of an infinite-dimensional linear subsystem with $e \in D(A) \subseteq X$ and a finite-dimensional subsystem with $\Delta G \in \Re^{mxm}$. This can be written in the following form using $w \equiv \begin{bmatrix} e \\ \Delta G \end{bmatrix} \in D \equiv D(A)x\Re^{mxm} \subseteq \overline{X} \equiv Xx\Re^{mxm}$:

$$\begin{cases} \frac{\partial w}{\partial t} = w_t = f(t, w) \equiv \begin{bmatrix} A_c e + B\rho(t) + v \\ -e_y \eta^* \gamma \end{bmatrix} \\ w(t_0) = w_0 \in D \text{ dense in } \overline{X} \equiv Xx\Re^{mxm} \end{cases} \tag{15}$$

The inner product on $\overline{X} \equiv Xx\Re^{mxm}$ can be defined as

$$(w_1, w_2) \equiv \left(\begin{bmatrix} x_1 \\ \Delta G_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ \Delta G_2 \end{bmatrix} \right)$$ which will make it a Hilbert space also.

$$\equiv (x_1, x_2) + \text{trace}\left(\Delta G_2 \Delta G_1{}^*\right)$$

The following r**obust stabilization theorem** shows that convergence to a neighborhood with radius determined by the supremum norm of ν is possible for a specific type of adaptive error system. In the following, we denote $\|M\|_2 \equiv \sqrt{\text{tr}\left(M\gamma^{-1}M^T\right)}$ as the trace norm of a matrix M where $\gamma > 0$.

Theorem 2 (Robust Stabilization) Consider the coupled system of differential equations

$$\begin{cases} \frac{\partial e}{\partial t} = \left(\underbrace{A + BG_e^* C}_{A_c} \right) e + B\Delta G\eta = A_c e + B\rho; e \in D(A); \rho \equiv \Delta G\eta \\ e_y = Ce \\ \Delta\dot{G} = \dot{G} - \dot{G}_* = \dot{G} = -e_y \eta^* \gamma; \quad \gamma > 0 \end{cases} \tag{16}$$

where $e, v \in D(A_C), z \in R^m$ and $\begin{bmatrix} e \\ G \end{bmatrix} \in \overline{X} \equiv XxR^{mxm}$ is a Hilbert space with inner product $\left(\begin{bmatrix} e_1 \\ G_1 \end{bmatrix}, \begin{bmatrix} e_2 \\ G_2 \end{bmatrix} \right) \equiv (e_1, e_2) + \text{tr}\left(G_1 \gamma^{-1} G_2\right)$, norm $\left\| \begin{bmatrix} e \\ G \end{bmatrix} \right\| \equiv \left(\|e\|^2 + \text{tr}\left(G\gamma^{-1}G\right) \right)^{\frac{1}{2}}$ and where $G(t)$ is the mxm adaptive gain matrix and γ is any positive definite constant matrix, each of appropriate dimension. *Assume the following:*

(i) (A, B, C) *is ASD with* $A_c \equiv A + BG_* C$

(ii) *there exists* $M_G > 0$ *such that* $\sqrt{\text{tr}\left(G^* G^{*T}\right)} \leq M_G$

(iii) *there exists* $M_\upsilon > 0$ *such that* $\sup\limits_{t \geq 0} \|\nu(t)\| \leq M_\nu < \infty$

(iv) *there exists* $\alpha > 0$ *such that* $a \leq \frac{q_{\min}}{p_{\max}}$, *where* $q_{\min}, p_{\max}$ *are defined in Definition 2*

(v) *the positive definite matrix* γ *satisfies* $\text{tr}\left(\gamma^{-1}\right) \leq \left(\frac{M_\nu}{aM_G} \right)^2$, then the gain matrix, G(t), is bounded, and the state, e(t), exponentially with rate e^{-at} approaches the ball of radius $R_* \equiv \frac{(1+\sqrt{p_{\max}})}{a\sqrt{p_{\min}}} M_\nu$

Proof of Theorem 2 See Appendix 1. Now we can prove the robust stability and convergence of the direct adaptive controller (4) in closed-loop with the linear infinite-dimensional plant (1)–(2).

Theorem 3 Under Hypothesis 1, we have robust state and output tracking of the reference model:

$\begin{bmatrix} e \\ \Delta G \end{bmatrix} \underset{t \to \infty}{\longrightarrow} N(0, R_*)$ and since C is a bounded linear operator, we have

$e_y = y - y_m = Ce \underset{t\to\infty}{\to} N(0,\ R_*)$ with bounded adaptive gains $G \equiv \left[G_e\ G_m\ G_u\ G_D \right] = G_* + \Delta G$

Proof of Theorem 3 Follows directly from application of Theorem 2 to the error system (12) or (16). Note that uniform continuity is not needed since Barbalat's lemma [19] is not invoked here.

7 Robust Stabilization of Weakly Nonminimum Phase Infinite-Dimensional Systems

From Lemma 3, the transmission zeros of the infinite-dimensional open-loop plant (A, B, C) are the eigenvalues of its zero dynamics $(\bar{A}_{22}, \bar{A}_{12}, \bar{A}_{21})$. We will say that a linear infinite-dimensional system (A, B, C) is *minimum phase* when $\bar{A}_{22}$ generates an exponentially stable C_0 − semigroup $\overline{U}_{22}(t)$. Also, we will say (A, B, C) is *weakly minimum phase* when $\bar{A}_{22}$ can be rewritten, via coordinate transformation, as $\begin{bmatrix} \overline{A}_{22}^u & 0 \\ 0 & \overline{A}_{22}^s \end{bmatrix}$ where $\bar{A}_{22}^s$ generates an exponentially stable C_0 − semigroup $\overline{U}_{22}^s(t)$ and $\bar{A}_{22}^u$ is finite-dimensional with spectrum $\bar{A}_{22}^u$ being $\sigma\left(\overline{A}_{22}^u\right) = \sigma_p\left(\overline{A}_{22}^u\right) = \{\lambda_1, \ldots, \lambda_l\}$ isolated unrepeated eigenvalues with $\mathrm{Re}\lambda_k = 0$. In this case: $Z = \sigma\left(\overline{A}_{22}\right) = \sigma_p\left(\overline{A}_{22}^u\right) \cup \sigma\left(\overline{A}_{22}^s\right)$ and there are only a finite number of unrepeated marginally stable zeros. Then the unstable zero dynamics $(\bar{A}_{22}^u, \bar{A}_{12}^u, \bar{A}_{21}^u)$ form a finite-dimensional subsystem.

We have the following relationship between minimum phase systems and almost strict dissipativity:

Theorem 4 Assume $CB = \left[\left(c_i, b_j\right)\right]$ is a (symmetric) positive definite *mxm* matrix and the *zero dynamics* $(\bar{A}_{22}, \bar{A}_{21}, \bar{A}_{12})$ are exponentially stable, i.e., there exists $\overline{P}_{22}, \overline{Q}_{22}$ bounded self-adjoint coercive operators such that $\mathrm{Re}\left(\overline{P}_{22}\overline{A}_{22}z_2, z_2\right) = -\left(\overline{Q}_{22}z_2, z_2\right) \leq -\mu_2\|z_2\|^2$, then (A, B, C) is ASD and conversely it is true.

Proof: see [13].

Now for weakly nonminimum phase systems, we can use the following *modified adaptive controller which includes a zero filter to compensate for the unstable zero dynamics:*

$$\begin{aligned} & u = G_m x_m + G_u u_m + G_e e_y + G_D \phi_D + \widehat{\nu}_u \\ & \textit{The Zero Filter} \begin{cases} \widehat{\nu}_u = -(CB)^{-1}\overline{A}_{12}^u \widehat{z}_u \\ \dot{\widehat{z}}_u = \overline{A}_{21}^u e_y + \overline{A}_{22}^u \widehat{z}_u \end{cases} \end{aligned} \tag{17}$$

$$Adaptive\ Gains\begin{cases}\dot{G}_u = -aG_u - e_y u_m^* \gamma_u; \gamma_u > 0\\ \dot{G}_m = -aG_m - e_y x_m^* \gamma_m; \gamma_m > 0\\ \dot{G}_e = -aG_e - e_y e_y^* \gamma_e; \gamma_e > 0\\ \dot{G}_D = -aG_D - e_y \phi_D^* \gamma_D; \gamma_D > 0\end{cases}$$

Theorem 5 Assume $CB = \left[\left(c_i, b_j\right)\right]$ is a (symmetric) positive definite *mxm* matrix and (A, B, C) is a weakly nonminimum phase system then $\left(\overline{\overline{A}}, \overline{\overline{B}}, \overline{\overline{C}}\right)$ is ASD with

$$\overline{\overline{A}} \equiv \begin{bmatrix} \overline{A}_{11} & \overline{A}^{s}_{12} \\ \overline{A}^{s}_{21} & \overline{A}^{s}_{22} \end{bmatrix}, \overline{\overline{B}} \equiv \begin{bmatrix} CB \\ 0 \end{bmatrix}, \overline{\overline{C}} \equiv \begin{bmatrix} I & 0 \end{bmatrix}$$

and, under the assumptions of the robust stabilization Theorem 4, all adaptive gain matrices are bounded, and the state tracking error $e(t) \equiv x(t) - x_*(t)$, as well as the output tracing error $e_y(t) \equiv y(t) - y_m(t)$, converges exponentially with rate e^{-at} to the ball of radius

$$R_* \equiv \frac{\left(1 + \sqrt{p_{\max}}\right)}{a\sqrt{p_{\min}}} M_u$$

where the size of M_u depends on the zero filter error.

Proof We define the zero filter error $e_u \equiv \widehat{z}_u - z_u$ and obtain:

$$\begin{cases} \widehat{v}_u = -(CB)^{-1}\overline{A}^{u}_{12}\left(z_u + e_u\right) = v_u - (CB)^{-1}\overline{A}^{u}_{12} e_u \\ \dot{e}_u = \overline{A}^{u}_{22} e_u \end{cases} \tag{18}$$

This yields the following *error system:*

$$\begin{cases} \frac{\partial e}{\partial t} = Ae + B\Delta u \\ e_y \equiv y - y_m = y - y_* \equiv \Delta y = Ce \end{cases}$$
$$\text{where} \begin{cases} \Delta u \equiv u - u_* = G_e^* e_y + \underbrace{\Delta G}_{G - G_*} \eta + v_u + \underbrace{(CB)^{-1}\overline{A}^{u}_{12} e_u}_{\Delta v} \\ \dot{G} = \begin{bmatrix} \dot{G}_m & \dot{G}_u & \dot{G}_e & \dot{G}_D \end{bmatrix} = -e_y \eta^* \gamma + aG \end{cases} \tag{19}$$

When CB is nonsingular, we can take the above error system in the normal form of (11) with $\overline{A}_{22} = \begin{bmatrix} \overline{A}^{s}_{22} & 0 \\ 0 & \overline{A}^{u}_{22} \end{bmatrix}$, to obtain:

$$\begin{cases} \dot{e}_y = \overline{A}_{11}e_y + \overline{A}^s_{12}z_s + \overline{A}^u_{12}z_u + CB\left(\Delta G\eta + \nu_u + \Delta\nu_u\right) \\ \quad = \overline{A}_{11}e_y + \overline{A}^s_{12}z_s + CB\left(\Delta G\eta\right) + \overline{A}^u_{12}e_u \\ \frac{\partial z_s}{\partial t} = \overline{A}^s_{21}e_y + \overline{A}^s_{22}z_s \\ \dot{e}_u = \overline{A}^u_{22}e_u \end{cases}$$

$$\Rightarrow \begin{cases} \frac{\partial}{\partial t}\begin{bmatrix} e_y \\ z_s \end{bmatrix} = \underbrace{\begin{bmatrix} \overline{A}_{11} & \overline{A}^s_{12} \\ \overline{A}^s_{21} & \overline{A}^s_{22} \end{bmatrix}}_{\overline{\overline{A}}}\begin{bmatrix} e_y \\ z_s \end{bmatrix} + \underbrace{\begin{bmatrix} CB \\ 0 \end{bmatrix}}_{\overline{\overline{B}}}\Delta G\eta + \underbrace{\begin{bmatrix} \overline{A}^u_{12} \\ 0 \end{bmatrix} e_u}_{\nu} \\ e_y = \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{\overline{\overline{C}}}\begin{bmatrix} e_y \\ z_s \end{bmatrix} \end{cases}$$

When the open-loop system (A, B, C) is weakly minimum phase, e_u is bounded and therefore

$$\|\nu\| \leq \|A^u_{12}\| \, \|e_u\| \leq M_u < \infty$$

8 Application: Adaptive Control of Unstable Diffusion Equations Described by Self-Adjoint Operators with Compact Resolvent

We will apply the above direct adaptive controller on the following single-input/single-output Cauchy problem:

$$\begin{cases} \frac{\partial x}{\partial t} = Ax + b\left(u + u_D\right), x(0) \equiv x_0 \in D(A) \\ y = (c, x)\,, \text{ with } b, c \in D(A) \end{cases}$$

And the *reference model* will be a simple output regulator:

$$y_m = x_m = 0$$

For this application we will *assume the disturbances are step functions*. Note that the disturbance functions can be any basis function as long as ϕ_D is bounded, in particular sinusoidal disturbances are often applicable. So we have $\phi_D \equiv 1$ and $\begin{cases} u_D = (1)z_D \\ \dot{z}_D = (0)z_D \end{cases}$ which implies $F = 0$ and $\theta_D = 1$.Let $u = G_e y + G_D + \widehat{\nu}_u$

$$Zero\ Filter \quad \begin{cases} \widehat{\nu}_u = -(CB)^{-1}\overline{A}^u_{12}\widehat{z}_u \\ \dot{\widehat{z}}_u = \overline{A}^u_{21}e_y + \overline{A}^u_{22}\widehat{z}_u \end{cases} \tag{20}$$

$$\textit{Adaptive Gains}\ \begin{cases} \dot{G}_e = -aG_e - yy^*\gamma_e; \gamma_e > 0 \\ \dot{G}_D = -aG_D - y\phi_D^*\gamma_D; \gamma_D > 0 \end{cases}$$

We will assume not only that A is closed and densely defined, but it is also a *self-adjoint operator with compact resolvent*. This means A has discrete real spectrum: $\lambda_1 \geq \lambda_2 \geq .... \to -\infty$ and $\{\phi_k\}_{k=1}^{\infty}$ an orthonormal sequence of eigenfunctions; see [18] Theorem 6.29 p. 187.

Assume for this example that $\begin{cases} \lambda_1 > 0; \lambda_2 = -\lambda_1 < 0 \\ \lambda_k < 0 \forall k = 3, 4, ... \end{cases}$ and

Define $b = c = \theta_1 \equiv \frac{\phi_1+\phi_2}{\sqrt{2}}, \theta_2 \equiv \frac{\phi_1-\phi_2}{\sqrt{2}}, \theta_k \equiv \phi_k \forall k - 3, 4, \ldots$

Then $\begin{cases} R(B) = sp\,\{b\} = sp\,\{\theta_1\} \\ N(C) = \overline{sp}\,\{\theta_2, \phi_3, \phi_4, ...\} \end{cases}$ and $\|\theta_1\| = \|\theta_2\| = 1 \& (\theta_1, \theta_2) = 0, (\theta_i, \phi_k) = 0 \forall i - 1, 2.k = 3, 4, ...$So $\{\theta_2, \phi_3, \phi_4, ...\}$ is an orthonormal basis for $N(C)$. We see that $CB = (c, b) = \|b\|^2 = 1$. Consequently $P_1 x \equiv B(CB)^{-1}Cx = BCx = b\,(c, x) = b\,(b, x) = \theta_1\,(\theta_1, x)$, which for this example is orthogonal projection onto $R(B)$. Also $P_2 x \equiv (I - P_1)\,x = \theta_2\,(\theta_2, x) + \sum_{k=3}^{\infty} \underbrace{(x, \phi_k)}_{x_k}\phi_k$ which is orthogonal projection onto $N(C)$. In general these projections will <u>not</u> be orthogonal ones.

Now

$$AP_1 x = \sum_{k=1}^{\infty} (\phi_k, P_1 x)\,\phi_k = \frac{\lambda_1 + \lambda_2}{2}\,(\theta_1, x)\,\theta_1 + \tfrac{\lambda_1 - \lambda_2}{2}\,(\theta_1, x)\,\theta_2$$
$$\Rightarrow A_{11} x \equiv P_1 A P_1 x = \tfrac{\lambda_1+\lambda_2}{2}\,(\theta_1, x)\,\theta_1$$

Also,

$$AP_2 x = \sum_{k=1}^{\infty} \lambda_k\,(\phi_k, P_2 x)\,\phi_k = \sum_{k=1}^{\infty} \left(\phi_k, (\theta_2, x) + \sum_{l=3}^{\infty} \underbrace{(x, \phi_l)}_{x_l}\phi_l \right)\phi_k$$
$$= \left(\tfrac{\lambda_1-\lambda_2}{2}\theta_1 + \tfrac{\lambda_1+\lambda_2}{2}\theta_2\right)(\theta_2, x) + \sum_{k=3}^{\infty} \lambda_k\,(x, \phi_k)\phi_k;$$
$$\Rightarrow A_{12} x \equiv P_1 A P_2 x = \tfrac{\lambda_1-\lambda_2}{2}\,(\theta_2, x)\,\theta_1 + P_1\left(\sum_{k=3}^{\infty} \lambda_k\,(x, \phi_k)\phi_k\right) = \tfrac{\lambda_1-\lambda_2}{2}\,(\theta_2, x)\,\theta_1;$$
$$A_{22} x \equiv P_2 A P_2 x = P_2\left[\left(\tfrac{\lambda_1-\lambda_2}{2}\theta_1 + \tfrac{\lambda_1+\lambda_2}{2}\theta_2\right)(\theta_2, x) + \sum_{k=3}^{\infty} \lambda_k\,(x, \phi_k)\phi_k\right]$$
$$= \tfrac{\lambda_1+\lambda_2}{2}\,(\theta_2, x)\,\theta_2 + \sum_{k=3}^{\infty} \lambda_k\,(x, \phi_k)\phi_k;$$
$$A_{21} x \equiv P_2 A P_1 x = P_2\left(\tfrac{\lambda_1+\lambda_2}{2}\,(\theta_1, x)\,\theta_1 + \tfrac{\lambda_1-\lambda_2}{2}\,(\theta_1, x)\,\theta_2\right) = \tfrac{\lambda_1-\lambda_2}{2}\,(\theta_1, x)\,\theta_2.$$

We have

$$\begin{cases} W_2x \equiv \begin{bmatrix} (\theta_2, P_2x) \\ (\theta_3, P_2x) \\ (\theta_4, P_2x) \\ .... \end{bmatrix} = \begin{bmatrix} (\theta_2, x) \\ (\theta_3, x) \\ (\theta_4, x) \\ .... \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ x_4 \\ .... \end{bmatrix} \equiv \underline{z} \in l^2 \\ W_2^* z = \sum\limits_{k=2}^{\infty} z_k \theta_k \end{cases}$$

$$\Rightarrow \begin{cases} \overline{A}_{11} \equiv CA_{11}B(CB)^{-1} = (\theta_1, A_{11}\theta_1) = \frac{\lambda_1+\lambda_2}{2} = 0 \text{ since } \lambda_2 = -\lambda_1 \\ \overline{A}_{12} \equiv CA_{12}W_2^* = \left[\frac{\lambda_1-\lambda_2}{2}, 0, 0, ...\right] \\ \overline{A}_{21} \equiv W_2A_{21}B(CB)^{-1} = W_2A_{21}\theta_1 = \left[\frac{\lambda_1-\lambda_2}{2}, 0, 0, ...\right] \\ \overline{A}_{22} \equiv W_2A_{22}W_2^* = diag\left[\frac{\lambda_1+\lambda_2}{2} = 0, \lambda_3, \lambda_4, ...\right] \end{cases}$$

$$\Rightarrow Z\,(A, B, C) = \sigma_p\left(\overline{A}_{22}\right) = \left\{\frac{\lambda_1 + \lambda_2}{2} = 0, \lambda_3, \lambda_4, ...\right\}$$

From this we see that there is one transmission zero at 0 and the rest are exponentially stable since $\lambda_k < 0 \forall k = 2, 3, 4, \ldots$; consequently this open-loop system is weakly minimum phase.

Furthermore we have the information for the *zero filter* to compensate for the single unstable transmission zero at 0:

$$\begin{cases} \overline{A}^u_{12} = \frac{\lambda_1-\lambda_2}{2} = \lambda_1 > 0 \\ \overline{A}^u_{22} = 0 \\ \overline{A}^u_{21} = \frac{\lambda_1-\lambda_2}{2} = \lambda_1 > 0 \end{cases}$$

$\Rightarrow$ *Zero Filter* $\begin{cases} \widehat{v}_u = -(CB)^{-1}\overline{A}^u_{12}\widehat{z}_u = -\lambda_1\widehat{z}_u \\ \dot{\widehat{z}}_u = \overline{A}^u_{21}e_y + \overline{A}^u_{22}\widehat{z}_u = \lambda_1 y + (0)\widehat{z}_u \end{cases}$ which has the transfer function $P(s) = -\frac{\lambda_1^2}{s}$.

And this is exactly the right transfer function to compensate the transmission zero at 0.

9 Perturbation Results

The previous results depend upon $b = P_N b \in S_N$.

However, it is possible to allow $b \equiv P_N b + \varepsilon P_R b \in D(A); \varepsilon \geq 0$.

Define $x_N \equiv P_N x$ and $x_R \equiv P_R x$

this implies that

$$
\begin{aligned}
\text{Re}\,(A(\varepsilon)_c x, x) &= \text{Re}\left(\begin{bmatrix} A_N^c & \varepsilon A_{12} \\ \varepsilon A_{21} & A_R + \varepsilon A_{22} \end{bmatrix}\begin{bmatrix} x_N \\ x_R \end{bmatrix}, \begin{bmatrix} x_N \\ x_R \end{bmatrix}\right) \\
&= \text{Re}\left(\begin{bmatrix} A_N^c & 0 \\ 0 & A_R \end{bmatrix}\begin{bmatrix} x_N \\ x_R \end{bmatrix}, \begin{bmatrix} x_N \\ x_R \end{bmatrix}\right) + \varepsilon \underbrace{\text{Re}\,(\Delta A x, x)}_{\leq |(\Delta A x, x)|} \\
&\leq -\sigma \underbrace{\left(\|x_N\|^2 + \|x_R\|^2\right)}_{\|x\|^2} + \varepsilon\, \|\Delta A\|\, \|x\|^2 = \text{-}\,(\sigma - \varepsilon\, \|\Delta A\|)\, \|x\|^2.
\end{aligned}
$$

And this proves: $\text{Re}\,(A(\varepsilon)_c x, x) \leq -\left(\underbrace{\sigma - \varepsilon\, \|\Delta A\|}_{\gamma > 0}\right) \|x\|^2$ for all $0 \leq \varepsilon < \frac{\sigma}{\|\Delta A\|}$.

And we have $(A(\varepsilon)_c, B, C)$ strictly dissipative and we can apply Theorem 2 again.

Therefore, for small $\varepsilon > 0$, all previous results are still true and we do not need b entirely confined to S_N.

10 Conclusions

In Theorems 2 and 3 we have robust stabilization results for linear dynamic systems on infinite-dimensional Hilbert spaces under the hypothesis of almost strict dissipativity. This idea is an extension of the concept of m-accretivity for infinite-dimensional systems; see [18] pp. 278–280. In Theorem 3, we showed that adaptive model tracking is possible with a very simple direct adaptive controller that knows very little specific information about the system it is controlling. This controller can also mitigate persistent disturbances. There was no use of Barbalat's lemma which requires certain signals to be uniformly continuous. However, we do not get something for nothing; we must relax the idea that all signals will converge to 0 and replace it with the idea that they will be attracted exponentially to a prescribed neighborhood whose size depends on the norm of the completely unknown disturbance. In order to cause such an infinite-dimensional system to track a finite-dimensional reference model, we used the idea of ideal trajectories, and in Theorem 1 we showed conditions for the existence and uniqueness of these ideal trajectories without requiring any deep knowledge of the infinite-dimensional plant.

In Theorem 4, we connect the idea of almost strict dissipativity, which is essential for robust direct adaptive control, to the exponential stability of the open-loop system transmission zeros. But we recognize that many applications will not be truly minimum phase; so we consider the case of weakly minimum phase systems, i.e., ones where there are a finite number of isolated zeros with zero real part. In Theorem 5, we develop the idea of extracting the unstable zero dynamics from these weakly minimum phase systems and using these dynamics to compensate the direct adaptive controller with a zero filter to perform robustly in their presence. Certainly an argument can be made that the knowledge of the unstable zero dynamics is hard to obtain, but we think that, because these unstable zero dynamics are finite-dimensional, it is entirely possible to obtain them via some offline system identification method and use them as we have described. Also note that the zero filter is added onto the existing adaptive controller and does not require any further redesign for its use. A related but different approach for decoupling nonminimum phase zeros in a finite-dimensional linear system appears in [20]; it does not use adaptive control or normal form, and assumes all parametric information for the plant is available.

We applied these results to a general linear infinite-dimensional weakly minimum phase (with a single zero at zero) linear system described by a self-adjoint operator with compact resolvent. An example of such a general system was shown to be able to robustly track the outputs of a finite-dimensional reference model in the presence of persistent disturbances using a zero filter augmentation to the adaptive controller without further modification.

Appendix 1: Proofs of Lemmae 1, 2, and Theorem 1

Proof of Lemma 1 Consider $\begin{aligned} P_1^2 &= \left(B(CB)^{-1}C\right)\left(B(CB)^{-1}C\right) \\ &= B(CB)^{-1}C \equiv P_1 \end{aligned}$.

Hence P_1 is a projection.

Clearly, $R\left(P_1\right) \subseteq R(B)$ and $z = Bu \in R(B)$ which implies $\begin{aligned} P_1 z &= \left(B(CB)^{-1}C\right)Bu \\ &= Bu = z \in R\left(P_1\right) \end{aligned}$.

Therefore $R\left(P_1\right) = R(B)$.

Also $N\left(P_1\right) = N(C)$ because $N(C) \subseteq N\left(P_1\right)$ and $z \in N\left(P_1\right)$ implies that $P_1 z = 0$ which implies that $CP_1 z = CB(CB)^{-1}Cz = 0$ or $N\left(P_1\right) \subseteq N(C)$.

So P_2 is a projection onto $R(B)$ along $N(C)$ but $P_2^* \neq P_2$ so it is not an orthogonal projection in general. We have $X = R\left(P_1\right) \oplus N\left(P_1\right)$; hence $X = R(B) \oplus N(C)$.

Since $b_i \in D(A)$, we have $R(B) \subset D(A)$.

Consequently $D(A) = (R(B) \cap D(A)) \oplus (N(C) \cap D(A)) = R(B) \oplus (N(C) \cap D(A))$.

The projection P_1 is bounded since its range is finite-dimensional, and the projection P_2 is bounded because $\|P_2\| \leq 1 + \|P_1\| < \infty$.

This completes the proof of Lemma 1.

Proof of Lemma 2 Since X is separable, we can let $N(C) = \overline{sp}\{\theta_k\}_{k=1}^{\infty}$ be an orthonormal basis.

Define $W_2 : X \to l_2$ by $W_2 x \equiv \begin{bmatrix} (\theta_1, P_2 x) \\ (\theta_2, P_2 x) \\ (\theta_3, P_2 x) \\ \cdots \\ \end{bmatrix}$.

Note that $\|W_2 x\|^2 = \sum_{k=1}^{\infty} |(\theta_k, P_2 x)|^2 = \|P_2 x\|^2 < \infty$ which implies $W_2 x \in l_2$.

So W_2 is a bounded linear operator, and an isometry of $W_2 N(C)$ into l_2.

Consequently $W_2 W_2^* = I$ on $N(C)$.

Then we have $W_2^* W_2 = P_2$ and the retraction: $z_2 = W_2 P_2 x \in l_2$.

Also $W_2^* z_2 = W_2^* \left(W_2 P_2 x\right) = P_2 x$.

Now, using $x = P_1 x + P_2 x$ from Lemma 1, we have

$$\begin{aligned} \dot{y} &= CP_1 \dot{x} \\ &= CP_1 A\left(P_1 x + P_2 x\right) + CP_1 Bu \\ &= C\left(B(CB)^{-1}C\right)AB(CB)^{-1}y + C\left(B(CB)^{-1}C\right)A\left(W_2^* z_2\right) + C\left(B(CB)^{-1}C\right)Bu \\ &= \overline{A}_{11} y + \overline{A}_{12} z_2 + CBu \end{aligned}$$

and

$$\begin{aligned} \dot{z}_2 &= W_2 P_2 \dot{x} \\ &= WP_2\left[A\left(P_1 x + P_2 x\right) + Bu\right] \\ &= W_2 P_2 A\left(B(CB)^{-1}y + W_2^8 z_2\right) + W_2 P_2 Bu \\ &= W_2\left(I - B(CB)^{-1}B\right)AB(CB)^{-1}y + W_2\left(I - B(CB)^{-1}B\right)AW_2^* z_2 \\ &= \overline{A}_{21} y + \overline{A}_{22} z_2. \end{aligned}$$

This yields the *normal form* (11).

Choose $W \equiv \begin{bmatrix} C \\ W_2 P_2 \end{bmatrix}$ which is a bounded linear operator. Then W has a bounded inverse explicitly stated as $W^{-1} \equiv \begin{bmatrix} B(CB)^{-1} & W_2^* \end{bmatrix}$.

This gives

$$\begin{aligned} WW^{-1} &= \begin{bmatrix} CB(CB)^{-1} & CW_2^* \\ W_2 P_2 B(CB)^{-1} & W_2 P_2 W_2^* \end{bmatrix} \\ &= \begin{bmatrix} I_m & 0 \\ 0 & W_2 W_2^* \end{bmatrix} = \begin{bmatrix} I_m & 0 \\ 0 & I \end{bmatrix} = I \end{aligned}$$

because $R\left(W_2^*\right) \subseteq N(C)$.

Furthermore, $W^{-1}W = P_1 + W_2^* W_2 P_2 = P_1 + P_2 = I$ because $W_2 W_2^* = I$ on $N(C)$.

Also direct calculation yields

$$\begin{cases} \overline{B} \equiv WB = \begin{bmatrix} CB \\ W_2 P_2 B \end{bmatrix} = \begin{bmatrix} CB \\ 0 \end{bmatrix} \\ \overline{C} \equiv CW^{-1} = \begin{bmatrix} CB(CB)^{-1} & CW_2^* \end{bmatrix} = \begin{bmatrix} I_m & 0 \end{bmatrix} \\ \overline{A} \equiv WAW^{-1} = \begin{bmatrix} CAB(CB)^{-1} & CAW_2^* \\ W_2 P_2 AB(CB)^{-1} & W_2 P_2 A P_2 W_2^* \end{bmatrix} \end{cases}$$

This completes the proof of Lemma 2.

Proof of Theorem 1 Define $\overline{S}_1 \equiv W^{-1} S_1 = \begin{bmatrix} \overline{S}_a \\ \overline{S}_b \end{bmatrix}$ and $\overline{H}_1 \equiv WH_1 = \begin{bmatrix} \overline{H}_a \\ \overline{H}_b \end{bmatrix}$.

From (10), we obtain

$$\begin{cases} \overline{AS}_1 + \overline{B} S_2 = \overline{S}_1 L_m + \overline{H}_1 \\ \overline{CS}_1 = H_2 \end{cases}$$

where $\left(\overline{A}, \overline{B}, \overline{C}\right)$ is the normal form (11).

From this we obtain

$$\begin{cases} \overline{S}_a = H_2 \\ S_2 = (CB)^{-1} \left[H_2 L_m + \overline{H}_a - \left(\overline{A}_{11} H_2 + \overline{A}_{12} \overline{S}_b\right)\right] . \\ \overline{A}_{22} \overline{S}_b - \overline{S}_b L_m = \overline{H}_b - \overline{A}_{21} H_2 \end{cases}$$

We can rewrite the last of these equations as

$\left(\lambda I - \overline{A}_{22}\right) \overline{S}_b - \overline{S}_b \left(\lambda I - L_m\right) = \overline{A}_{21} H_2 - \overline{H}_b \equiv \overline{H}$ for all complex λ.

Now assume that L_m is simple and therefore provides a basis of eigenvectors $\{\emptyset_k\}_{k=1}^L$ for $\Re^L$. This is not essential but will make this part of the proof easier to understand. The proof can be done with generalized eigenvectors and the Jordan form. So we have

$$\left(\lambda_k I - \overline{A}_{22}\right) \overline{S}_b \phi_k - \overline{S}_b \underbrace{\left(\lambda_k I - L_m\right) \phi_k}_{=0} = \overline{A}_{21} H_2 - \overline{H}_b \equiv \overline{H}$$

which implies that

$$\overline{S}_b \phi_k = \left(\lambda_k I - \overline{A}_{22}\right)^{-1} \overline{H} \phi_k$$

because $\lambda_{\mathrm{k}} \in \sigma\left(L_m\right) \subset \rho\left(\overline{A}_{22}\right)$.

Thus we have $\overline{S}_b z = \sum\limits_{k=1}^{L} \alpha_k \left(\lambda_k I - \overline{A}_{22}\right)^{-1} \overline{H} \phi_k \forall z = \sum\limits_{k=1}^{L} \alpha_k \phi_k \in \Re^L$.

Since $\lambda_{\mathrm{k}} \in \sigma\left(L_m\right) \subset \rho\left(\overline{A}_{22}\right)$, all $\left(\lambda_k I - \overline{A}_{22}\right)^{-1}$ are bounded operators.

Also $\overline{H} \equiv \overline{A}_{21} H_2 - \overline{H}_b$ is a bounded operator on $\Re^L$.

Therefore $\overline{S}_b$ is a bounded linear operator, and this leads to S_1 also bounded linear.

If we look at the converse statement and let $\lambda_* \in \sigma\left(L_m\right) \cap \sigma\left(\overline{A}_{22}\right) = \emptyset$.

Then there exists $\phi_* \neq 0$ such that $\left(\lambda_* I - \overline{A}_{22}\right)\overline{S}_b\phi_* - \overline{S}_b\underbrace{\left(\lambda_* I - L_m\right)\phi_*}_{=0} =$ $\left(\lambda_* I - \overline{A}_{22}\right)\overline{S}_b\phi_* = \overline{H}$.

In this case *three* things can happen when $\lambda_* \in \sigma\left(\overline{A}_{22}\right)$:

(1) $\left(\lambda_* I - \overline{A}_{22}\right)$ can fail to be one to one so multiple solutions of $\overline{S}_b$ will exist

(2) $R\left(\lambda_* I - \overline{A}_{22}\right)$ can fail to be all of X so no solutions $\overline{S}_b$ may occur, or

(3) $\left(\lambda_* I - \overline{A}_{22}\right)^{-1}$ can fail to be a bounded operator so solutions $\overline{S}_b$ may be unbounded.

In all cases these three alternatives lead to a lack of unique bounded operator solutions for S_1.

The proof of Theorem 1 is complete.

Appendix 2: Proof of Theorem 2

From (15) and Pazy Cor 2.5 p. 107 [1], we have a well-posed system in (16) where A_c is a closed operator, densely defined on $D\left(A_C\right) \subseteq X$ and generates a C_0 semigroup on X, and all trajectories starting in $D(A_C)$ will remain there. Hence we can differentiate signals in $D(A_C)$.

Consider the positive definite function,

$$V \equiv \frac{1}{2}\left(Pe, e\right) + \frac{1}{2}\mathrm{tr}\left[\Delta G\gamma^{-1}\Delta G^{\mathrm{T}}\right] \tag{21}$$

where $\Delta G(t) \equiv G(t) - G^*$ and P satisfies (13).

Taking the time derivative of (21) (this can be done $\forall e \in D\left(A_C\right)$) and substituting (2a) into the result yields $\dot{V} = \frac{1}{2}\left[\left(PA_c\, e, e\right) + \left(e, PA_c e\right)\right] + \left(PBw, e\right) + \mathrm{tr}\left[\Delta\dot{G}\gamma^{-1}\Delta G^{\mathrm{T}}\right] + \left(Pe, v\right); w \equiv \Delta Gz$.

Invoking the equalities in Definition 2 of strict dissipativity, using $x^{\mathrm{T}}y = \mathrm{tr}[yx^{\mathrm{T}}]$, and substituting (16) into the last expression, we get (with $\left\langle e_y, w\right\rangle \equiv e_y^* w$),

$$\left\{\begin{aligned}
\dot{V} &= \mathrm{Re}\left(PA_c e, e\right) + \left\langle e_y, w\right\rangle - a\cdot\mathrm{tr}\left[G\gamma^{-1}\Delta G^{\mathrm{T}}\right] - \underbrace{\mathrm{tr}\left(e_y z^T \Delta G^T\right)}_{\left\langle e_y, w\right\rangle} + \left(Pe, v\right)\\
&\leq -q_{\min}\|e\|^2 - a\cdot\mathrm{tr}\left[\left(\Delta G + G^*\right)\gamma^{-1}\Delta G^{\mathrm{T}}\right] + \left(Pe, v\right)\\
&\leq -\left(q_{\min}\|e\|^2 + a\cdot\mathrm{tr}\left[\Delta G\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right) + a\cdot\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| + \left|\left(Pe, v\right)\right|\\
&\leq -\left(\tfrac{2q_{\min}}{p_{\max}}\cdot\tfrac{1}{2}\left(Pe, e\right) + 2a\cdot\tfrac{1}{2}\mathrm{tr}\left[\Delta G\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right) + a\cdot\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| + \left|\left(Pe, v\right)\right|\\
&\leq -2aV + a\cdot\left|\mathrm{tr}\ \left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| + \left|\left(Pe, v\right)\right|
\end{aligned}\right.$$

Now, using the Cauchy-Schwarz inequality

$$\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| \le \|G^*\|_2\|\Delta G\|_2$$

and

$$|(Pe,v)| \le \left\|P^{\frac{1}{2}}\nu\right\|\ \left\|P^{\frac{1}{2}}e\right\| = \sqrt{(P\nu,v)}\cdot\sqrt{(Pe,e)}$$

We have

$$\begin{aligned}\dot{V}+2aV &\le a\cdot\|G^*\|_2\|\Delta G\|_2+\sqrt{p_{\max}}\,\|\nu\|\,\sqrt{(Pe,e)}\\ &\le a\cdot\|G^*\|_2\|\Delta G\|_2+\left(\sqrt{p_{\max}}M_\nu\right)\sqrt{(Pe,e)}\\ &\le \left(a\|G^*\|_2+\sqrt{p_{\max}}M_\nu\right)\underbrace{\sqrt{2}\left[\frac{1}{2}\,(Pe,e)+\frac{1}{2}\|\Delta G\|_2^2\right]^{\frac{1}{2}}}_{V^{\frac{1}{2}}}\end{aligned}$$

Therefore,

$$\frac{\dot{V}+2aV}{V^{\frac{1}{2}}} \le \left(a\|G^*\|_2+\sqrt{p_{\max}}M_\nu\right)\sqrt{2}$$

Now, using the identity $\mathrm{tr}\,[ABC] = \mathrm{tr}\,[CAB]$,

$$\begin{aligned}\|G^*\|_2 &\equiv \left[\mathrm{tr}\left(G^*\gamma^{-1}(G^*)^T\right)\right]^{\frac{1}{2}} = \left[\mathrm{tr}\left((G^*)^TG^*\gamma^{-1}\right)\right]^{\frac{1}{2}}\\ &\le \left[\left(\mathrm{tr}\left((G^*)^TG^*(G^*)^TG^*\right)\right)^{\frac{1}{2}}\left(\mathrm{tr}\left(\gamma^{-1}\gamma^{-1}\right)\right)^{\frac{1}{2}}\right]^{\frac{1}{2}}\\ &= \left[\mathrm{tr}\left(G^*(G^*)^T\right)\right]^{\frac{1}{2}}\left[\mathrm{tr}\left(\gamma^{-1}\right)\right]^{\frac{1}{2}}\\ &\le \tfrac{M_\nu}{aM_G}\cdot M_G = \tfrac{M_\nu}{a}\end{aligned}$$

which implies that

$$\frac{\dot{V}+2aV}{V^{\frac{1}{2}}} \le \left(1+\sqrt{p_{\max}}\right)M_\nu\sqrt{2} \tag{22}$$

From

$$\begin{aligned}\frac{d}{dt}\left(2e^{at}V^{\frac{1}{2}}\right) &= e^{at}\frac{\dot{V}+2aV}{V^{\frac{1}{2}}}\\ &\le e^{at}\left(1+\sqrt{p_{\max}}\right)M_\nu\sqrt{2}\end{aligned}$$

Integrating this expression we have $e^{at}V(t)^{1/2} - V(0)^{1/2} \leq \frac{\left(1+\sqrt{p_{\max}}\right) M_\nu}{a}\left(e^{at}-1\right)$. Therefore,

$$V(t)^{1/2} \leq V(0)^{1/2}e^{-at} + \frac{\left(1+\sqrt{p_{\max}}\right)\ M_\nu}{a}\left(1-e^{-at}\right) \tag{23}$$

The function $V(t)$ is a norm function of the state $e(t)$ and matrix $G(t)$. So, since $V(t)^{1/2}$ is bounded for all t, then $e(t)$ and $G(t)$ are bounded. We also obtain the following inequality:

$$\sqrt{p_{\min}}\,\|e(t)\| \leq V(t)^{1/2}$$

Substitution of this into (23) gives us an exponential bound on state $e(\tau)$:

$$\|e(t)\| \leq \frac{e^{-at}}{\sqrt{p_{\min}}}V(0)^{1/2} + \frac{\left(1+\sqrt{p_{\max}}\right)\ M_\nu}{a\sqrt{p_{\min}}}\left(1-e^{-at}\right) \tag{24}$$

Taking the limit superior of (24), we have

$$\overline{\lim_{\tau\to\infty}}\,\|e(t)\| \leq \frac{\left(1+\sqrt{p_{\max}}\right)}{a\sqrt{p_{\min}}}M_\nu \equiv R_* \tag{25}$$

And the proof is complete.

References

1. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Springer, Berlin (1983)
2. D'Alessandro, D.: Introduction to Quantum Control and Dynamics. Chapman & Hall, London (2008)
3. Kothari, D., Nagrath, I.: Modern Power System Analysis. McGraw-Hill, New York (2003)
4. Curtain, R., Pritchard, A.: Functional Analysis in Modern Applied Mathematics. Academic, London (1977)
5. Cannarsa, P., Coron, J.-M. (eds.): Control of Partial Differential Equations (Lecture Notes in Mathematics 2048/C.I.M.E. Foundation Subseries). Springer, Heidelberg (2012)
6. Troltzsch, F.: Optimal Control of Partial Differential Equations. American Mathematical Society, Providence, RI (2010)
7. Balas, M., Erwin, R.S., Fuentes, R.: Adaptive control of persistent disturbances for aerospace structures. Proceedings of the AIAA Guidance, Navigation and Control Conference, Denver (2000)
8. Fuentes, R., Balas, M.: Direct adaptive rejection of persistent disturbances. J. Math. Anal. Appl. **251**, 28–39 (2000)
9. Fuentes, R., Balas, M.: Disturbance accommodation for a class of tracking control systems. Proceedings of the AIAA Guidance, Navigation and Control Conference, Denver, Colorado (2000)

10. Fuentes, R., Balas, M.: Robust model reference adaptive control with disturbance rejection. Proceedings of the American Control Conference, Anchorage (2002)
11. Balas, M., Gajendar, S., Robertson, L.: Adaptive tracking control of linear systems with unknown delays and persistent disturbances (or Who You Callin' Retarded?). Proceedings of the AIAA Guidance, Navigation and Control Conference, Chicago, IL, August 2009
12. Balas, M., Frost, S.: Adaptive model tracking for distributed parameter control of linear infinite-dimensional systems in Hilbert space. Proceedings of AIAA SCITECH, Boston, July 2013
13. Balas, M., Frost, S.: Adaptive regulation in the presence of persistent disturbances for linear infinite-dimensional systems in Hilbert space: conditions for almost strict dissipativity. Proceedings of the European Control Conference, Linz, July 2015
14. Balas, M., Frost, S.: Robust adaptive model tracking for distributed parameter control of linear infinite-dimensional systems in Hilbert space. IEEE/CAA J Automat Sin **1**(3), 294–301 (2014)
15. Kailath, T.: Linear Systems, pp. 448–449. Prentice-Hall (1980)
16. Wen, J.: Time domain and frequency domain conditions for strict positive realness". IEEE Trans. Autom. Control **33**(10), 988–992 (1988)
17. Renardy, M., Rogers, R.: An Introduction to Partial Differential Equations. Springer, Berlin (1993)
18. Kato, T.: Perturbation Theory for Linear Operators. Springer, Berlin (1980)
19. Popov, V.M.: Hyperstability of Control Systems. Springer, Berlin (1973)
20. Snell, A.: Decoupling of nonminimum phase plants and application to flight control. AIAA Guidance Navigation and Control Conference, Monterey, CA, August 2002
21. Balas, M.: Trends in large space structure control theory: fondest hopes, wildest dreams. IEEE Trans Automatic Control, AC-27, No. 3 (1982)
22. Balas, M., Fuentes, R.: A non-orthogonal projection approach to characterization of almost positive real systems with an application to adaptive control. Proceedings of the American Control Conference, Boston (2004)
23. Antsaklis, P., Michel, A.: A Linear Systems Primer. Birkhauser, Basel (2007)
24. Balas, M., Frost, S.: Distributed parameter direct adaptive control using a new version of the Barbalat-Lyapunov stability result in Hilbert Space. Proceedings of AIAA Guidance, Navigation and Control Conference, Boston, MA, 2013

Aeroelasticity of the PrandtlPlane: Body Freedom Flutter, Freeplay, and Limit Cycle Oscillation

Rauno Cavallaro, Rocco Bombardieri, Simone Silvani, Luciano Demasi, and Giovanni Bernardini

Abstract Aeroelasticity of PrandtlPlane configurations is a yet unexplored field. The overconstrained structural system and the mutual aerodynamic interference of the wings enhance the complexity of the aeroelastic response. In this work the aeroelastic behavior of several models based on wing system of 250-seat PrandtlPlane design is studied. When an aluminum version of the structure is considered, flutter is associated with a coalescence of the first two elastic modes, the first being characterized by a classic upward bending of both wings, and the second one being associated with an out-of-phase bending of the two wings and tilting of the lateral joint. Analyses show that energy is injected into the structure mainly at the tip of the front wing, close to the aileron. Effects of freeplay of mobile surfaces are evaluated, showing how, in some cases, an increase in the flutter speed is observed. When flutter analyses are repeated considering the configuration free to pitch and plunge, flutter speed does increase due to a particular interaction between rigid-body pitching and elastic modes. Several of the above findings are demonstrated on more detailed structural models considering the local stiffness distribution, and taking into consideration compressibility effects. When composite materials are employed, flutter issues are completely overcome.

R. Cavallaro (✉)
Technion - Israel Institute of Technology, Haifa, Israel
e-mail: rauno.cavallaro@gmail.com

R. Bombardieri
Università di Pisa, Pisa, Italy

S. Silvani • G. Bernardini
Università degli Studi Roma Tre, Rome, Italy

L. Demasi
San Diego State University, San Diego, CA, USA

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_3

1 PrandtlPlane: An Introduction

With the growth of passenger volume and an increasing attention towards a more sustainable aeronautic traffic, a consistent improvement of the efficiency of aircraft is required. Following the trend of the last decades it looks difficult to even think of reaching the ambitious goals set in documents such as *Horizon* 2020 and 2050. A technological revolution seems to be necessary.

Unconventional aircraft configurations were proposed as breakthrough in pursuing a higher efficiency. Among them, Joined Wings and Blended Wing Body were the most studied ones. Within the category of Joined Wings, the Box Wing is a layout in which the wing system is characterized by two wings (front and rear) connected at their tips by a vertical joint, and resembling, so, a box when observed frontally. The reader is warned that the name Box Wing is herein used when referring to the general case of a configuration featuring the typical box shape of the lifting system; PrandtlPlane is a different nomenclature for the same lifting system arrangement as given by several authors [1] at the University of Pisa. It has to be said that since their early work these authors focused not only on the lifting system, but also on its integration in a more comprehensive aircraft design perspective. The name PrandtlPlane will then be used when specifically referring to the Box Wing studied by the University of Pisa and partner groups.

A very brief introduction to the features of the PrandtlPlane is given next. For a more complete and detailed coverage of the topic please refer to the review paper [2] and the therein cited references.

1.1 Aerodynamic Properties of Box Wings

1.1.1 Prandtl's Work and Successive Studies

The first scientific study featuring a box-wing layout is the paper [3]. The German scientist Ludwig Prandtl introduced the concept of *Best Wing System* (BWS): among all possible layouts the *optimal* box shape was demonstrated to have the lowest level of induced drag for a fixed wingspan and lift (and also for a fixed maximum vertical dimension). An approximated formula relating the induced drag and a geometric parameter (the vertical aspect ratio, ratio between the vertical dimension and the wingspan, see Fig. 1) was given in the reference, although Prandtl did not explain its origin.

Several recent works focused on the lift distribution on the two wings performing in optimal conditions [4–6] (see Fig. 1), on the asymptotic value of the induced drag against the vertical aspect ratio, and on other aspects. A good review of this topic is also given in a chapter of this book [7].

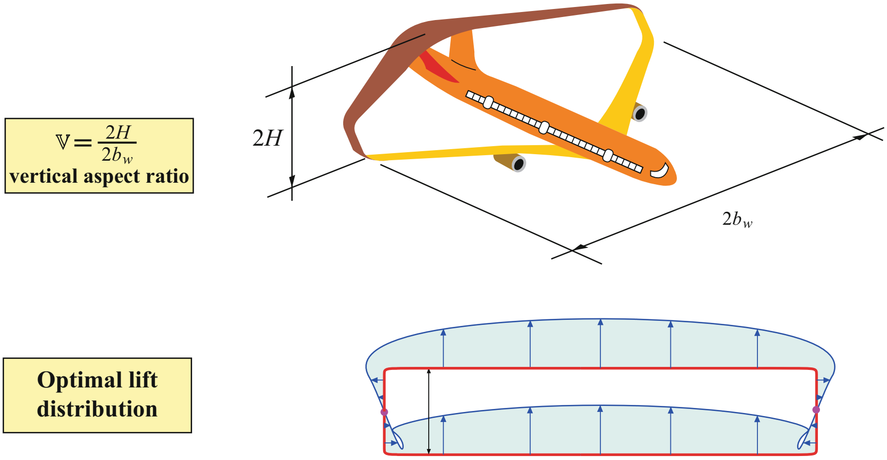


Fig. 1 Definition of vertical aspect ratio for a Box Wing (*top*). Distribution of the lift in optimal conditions (*bottom*). See also [4]

1.1.2 Extension of Prandtl's Results

Thanks to Munk's stagger theorem [8, 9], stating that the induced drag does not change when the wings are moved (and swept) along the undisturbed flow direction, Prandtl's results can be extended and gain a practical relevance: box-wing system with swept wings is then possible without induced drag penalties.

The above result, however, is valid only for potential aerodynamics and assuming a wake aligned with the freestream.

1.1.3 More Refined Model: Wake Shape

Several efforts studied effects of a more accurate description of the wake [10–12]. Significantly lower levels of induced drag were observed. It is important to notice that redistribution of aerodynamic forces has also effects on flight mechanics; for example, in the case shown in [12], the wake roll up contributed in distorting the aerodynamic field decreasing the load on the tip region of the rear-upper wing. In the paper [13] the way the wake was modeled had nonnegligible effects on the prediction of the aeroelastic behavior.

1.1.4 Viscous and Compressible Aerodynamics

Several numerical investigations considered the effects of compressibility. With an appropriate design of critical areas like junctions (wing-joint and wing-fin) issues related to shock waves were overcome [14, 15].

Besides numerical experiments, also several wind tunnel tests were performed. Results regarding the aerodynamic efficiency of the Box Wing [16–18] agreed with the predictions, confirming the advantages.

1.2 Aircraft Synthesis

The first effort tackling the design of an aircraft based on the Prandtl's lifting system was carried out by Lockheed [19] in the 1970s. The research team studied an application of the Box Wing for a transonic mid-range aircraft concept. Several variants of the wing-system layout were considered in order to satisfy different requirements. Not considering the aeroelastic constraints, the Box Wing turned out to be slightly lighter than a conventional (but employing unconventional technologies) layout reference aircraft.

Within a multiyear and multi-institution effort [1, 20–23] a 250-seat PrandtlPlane version (an artistic representation is given in Fig. 2) was designed considering aerodynamics, flight mechanics and dynamics, structures, aeroelasticity, engine integration, and infrastructures. Aerodynamic advantages were predicted, and the structural weight of the lifting system to the MTOW was found to be comparable to the one relative to vehicles of the same class.

1.3 Aeroelasticity

The most important aeroelastic analyses carried out on Box Wing/PrandtlPlane are briefly discussed in the following. The reader is referred to the following bibliography [2, 24] and the therein cited references for a more comprehensive discussion of the topic.

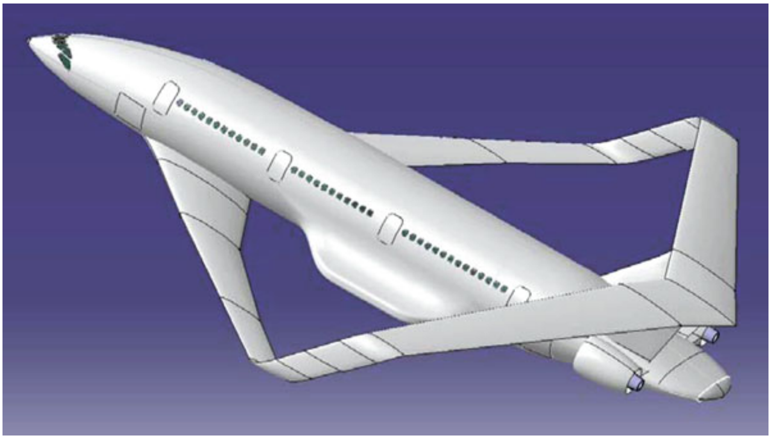


Fig. 2 250-Seat PrandtlPlane [1]

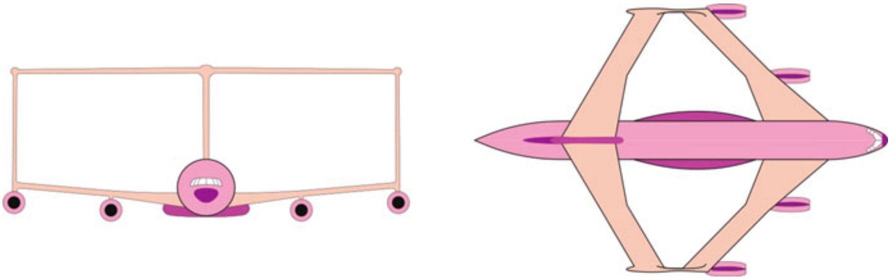


Fig. 3 Transonic Box Wing studied at Lockheed [19] in the 1970s

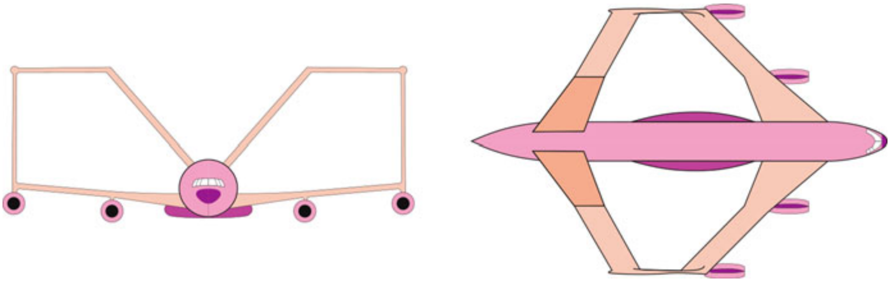


Fig. 4 Transonic Box Wing studied at Lockheed [19] in the 1970s; an alternative configuration designed to (unsuccessfully) increase flutter speed

1.3.1 Box Wing

Aeroelasticity of the Box Wing has been first studied within the Lockheed's investigation [19]. On the interim configuration (shown in Fig. 3), flutter occurred at very low speeds. Several modifications, from marginal to radical, were pursued with the aim of reducing aeroelastic issues. Figure 4 shows an alternative configuration built with that purpose. Despite the big effort, flutter issues were not overcome.

1.3.2 PrandtlPlane

In the case of the PrandtlPlane, the first aeroelastic studies can be found in [21, 25]. The configuration was designed for a range of 6000 nm, maximum take-off weight (MTOW) of 230 tons, and featured a wingspan of 55 m. In the structural optimization design, aeroelasticity was considered as a constraint. It was noted that a weight penalty was introduced by the flutter constraint, especially on the rear wing.

A more detailed study on flutter, taking into account the FAR/JAR regulation, was then carried out [26, 27]. The flight envelope of the aircraft was enlarged by 15 %, and "matched" flutter analyses were carried out (i.e., compressibility effects were taken into account). The same configuration that was previously assessed as flutter free [21, 25] was found not to be as such. Since in the previous optimization process [21, 25] the increase of stiffness to comply with the flutter constraint didn't follow too sophisticated design approaches (e.g., compressibility effects were not taken into account), later a pre-optimized configuration (obtained without the flutter constraint being considered) was chosen as the starting (baseline) one. Several modifications were then implemented on the baseline layout. Skin thickness on the front and rear wings was selectively or simultaneously increased, without touching the other structural components of the wing box. It was found that the most efficient way to have a flutter-free design was to increase front-wing skin thickness by 30 %, with a 3.58 % penalty in the total weight of the aircraft.

A different option was explored: the addition of a tip tank. With a 1000 kg tip tank on the front-wing flutter speed requirements were met. Interestingly, the addition of the tip tank on the aft wing was detrimental, lowering the speed of instability.

2 Contribution of the Present Study

This paper aims to summarize some aeroelastic researches carried out on the PrandtlPlane configuration. One of the first contributions is towards a more in-depth analysis of flutter of such configurations. This is accomplished [28] considering an aeroelastically "similar" model built from [26], and tracking flutter and postflutter (LCO) responses; moreover, mapping the energy transfer between fluid and structures gives a significant physical insight on the instability mechanism.

Control surfaces are necessary to guarantee maneuverability of the aircraft. However, they can be characterized by a nonnegligible level of freeplay which can induce aeroelastic instabilities. In this paper, flutter analyses are conducted considering freeplay of selected control surfaces [28]. An attempt to explain the results is also given considering the energy transfer diagram.

As a first (and often reliable) approximation inertial and elastic contribution of the fuselage are neglected during aeroelastic analyses. However, in some cases this approach is nonconservative. In this paper the fuselage inertial effects are retained, and flutter analyses repeated. For an in-depth insight, studies are performed in which the fuselage moment of inertia is varied. All flutter calculations consider only symmetric flight conditions (in respect to the longitudinal plane); as a consequence, the included rigid-body modes are the pitching and plunging ones.

The above investigations, however, do not consider effects of compressibility and rely on a structural description of the aircraft [28] obtained starting from a beam-like model [26]. With the aid of a detailed wing-box model featuring its typical elements (e.g., stiffeners, spars) and based on the one shown in [29], flutter analyses are repeated considering the effects of Mach number and, eventually, also the inertial properties of the fuselage.

The above employed models are based on a metallic (aluminum) structure. A realization of the same concept with composite materials, adapted from [29], is then selected, and aeroelastic properties of the detailed model are carried out considering also effects of compressibility and inertia of the fuselage.

This contribution is based on results shown in [28, 30, 31], which are here rearranged and augmented with new investigations and verifications.

3 Theoretical Highlights Regarding the Computational Tools

The herein performed investigations have been carried out both with in-house and commercial tools.

Flutter has been mostly studied with frequency-domain solvers, namely the commercial code NASTRAN and an in-house tool. The in-house capability consists of a finite element method for computational structural dynamics (CSD) and a doublet lattice method (DLM) based on [32].

Time-domain analyses have been carried out with an in-house tool to study the limit cycle oscillation (LCO) and also the time evolution of the system. An unsteady vortex lattice method (UVLM) was used for the aerodynamics. The coupling (time integration) and interfacing (load and displacement transferring) are described in detail in [13, 33]. Briefly, an implicit integration scheme is used to advance in time, being the (structural) nonlinear problem at each time-step solved by means of a Newton's iterative approach. Moreover, the interface information is passed either through an *infinite plate spline* (IPS) or a *moving least squares* (MLS) approach.

For both in-house capabilities, the geometrically nonlinear finite element (refer to [13]) is based on the constant strain triangle (CST) membrane element and the discrete Kirchhoff (DKT) plate element.

In all cases, the aerodynamics is based on the hypothesis of potential flow (non-viscid and irrotational). Considering attached flow, the potential theory underlying the computational method is adequate to simulate the aerodynamic field with relatively low computational costs.

4 Aeroelastic Models

The configurations studied in this paper refer to the 250-seat PrandtlPlane first introduced in [20, 34]. An artistic representation was already given in Fig. 2, and some technical details were shared in Sect. 1.3.2.

The acronyms used for the different models are the following: "PrP250" refers to the 250-seat PrandtlPlane, "v" identifies the version of the models, and "Al/Comp" to the material (aluminum or composite, respectively).

For all cases similar aerodynamic meshes have been used, see Fig. 5. The main differences concern the structural description.

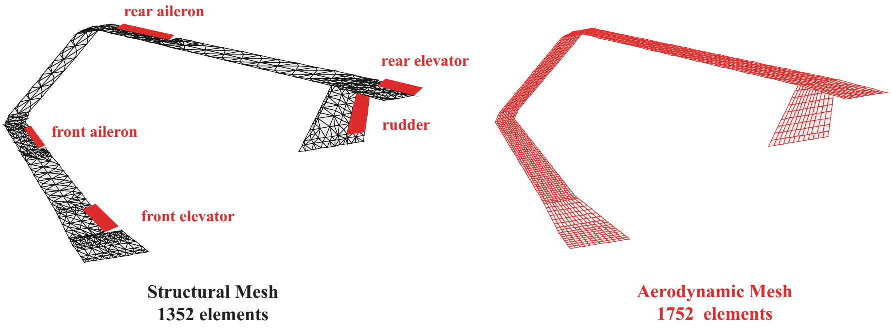


Fig. 5 PrP250v1 structural and aerodynamic models

4.1 PrP250v1

In [26] an aeroelastic model of the wing system was provided taking into account also the inertia of nonstructural elements (e.g., fuel). The structural model was described by beams.

For a compatibility issue with the in-house tool and for the purpose of this research, the above structure has been selected as a starting point, and has been more conveniently described by shell elements. Given the two different topologies (beam and shell), a modal aeroelastic equivalence is pursued through an optimization problem, as better described in [28]. This process, however, has not the aim of having identical aeroelastic behaviors; on the contrary, only meaningful stiffness and inertial distributions which are representative of the aircraft are sought. The final model is shown in Fig. 5. The structural description is characterized by shell elements having several layers (composite materials) to properly model the stiffness characteristics.

The location of control surfaces are extrapolated from [35] (based on master thesis work [36]). The control architecture is considered as a "realistic" one, as it was originally obtained through an optimization process taking into account several constraints. The wing-control surface connection is modeled through hinges and springs acting on the hinge line. These springs (described in detail in [28]) are also capable of modeling typical nonlinear freeplay motions.

In cases in which aeroelastic properties are studied considering the "constrained" (or fixed) case, the nodes on the root section of the fin are clamped; to simulate the interface with the fuselage this set of constraints is employed: the sections of the wing system on the symmetry plane are subjected to constraints complying with symmetry, whereas the nodes on the front-wing-fuselage intersection are free to rotate in the streamwise direction (simulating a simple support).

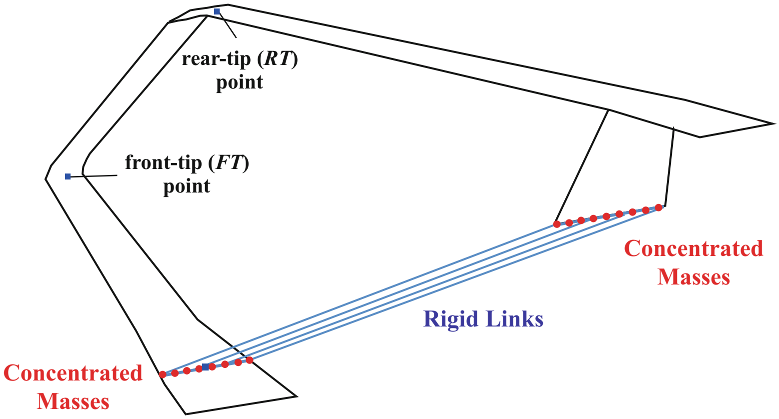


Fig. 6 "Free" version of the PrP250v1, free to plunge and pitch

In the early design stages, effects of fuselage are often considered negligible and only wings are modeled. However, this can eventually not be true. Fuselage has inertial and stiffness properties that can have an impact to the overall aeroelastic response of the system. If the aircraft is considered free to rigidly move in the space, the rigid-body modes can interact with the elastic ones changing the aeroelastic response; the so-called body-freedom flutter (BFF) can be observed, see [37, 38]. If such cases are sought to be studied, then fuselage inertia needs to be included and wing system–fuselage interfaces have to be rethought.

In this model, fuselage properties have been extrapolated from [20], obtaining a Mass $M_{\text{fus}}^{\text{ref}} = 9.1 \cdot 10^4$ kg, and pitching moment of inertia $I_{\text{fus}}^{\text{ref}} = 1.22 \cdot 10^7$ kg m^2. The target values have been obtained placing concentrated masses on the front-wing section intersecting the fuselage, and the root section of the fin. These areas were also stiffened to prevent local deformations. Moreover, in order to model a rigid fuselage, these sections were linked through rigid elements, as shown in Fig. 6. In this first attempt to study BFF, only symmetric motions are considered.

During the investigations, inertial properties have been modified simply reallocating and changing the mass values. In this process, the Center of Gravity (CoG) of the whole system was maintained fixed.

4.2 PrP250v2Al

4.2.1 The Original Model

In effort [29] (based on Master Thesis [39]) results obtained in [21] were investigated with a higher-fidelity approach. In particular, the configuration obtained with the structural optimization carried out with the constraints on maximum allowable stress, and featuring the nonsymmetric wing box was chosen as starting point.

A detailed FE model was built describing the structure of the wing box, whereas in the previous work [21] these structures were modeled to include their inertial effect at sectional level. Structures were all made of 2024-T3 aluminum alloy. The limit load analysis (relative to a load factor of 2.5 in symmetric flight) showed very circumscribed areas for which the equivalent tension was higher than the allowable one, 233 MPa, obtained reducing the yield stress (290 MPa) to take into consideration local stress concentrations, fatigue strength, etc. These peaks often coincided with areas of action of the external loads. Considering that wing sections showing these high stresses were globally under loaded, it was reasonable to expect that with an appropriate more refined local design, the peaks would have been cut without weight penalties.

It was noticed that deformations for the limit load case were a bit large, enough to cast doubt on the validity of the structural geometric linear approach.

4.2.2 The Modified Model: PrP250v2Al

The model of [29] featured a large amount of finite elements in order to appropriately predict the stress levels. Since the purpose of the present effort is the investigation of the aeroelastic properties, and also for compatibility with the in-house nonlinear aeroelastic code capable of performing both time- and frequency-domain analyses, the original model was modified using a smaller number of finite elements. In this process, it was accurately checked that natural frequencies and deformations were not changing. A sketch of the final model is given in Fig. 7.

The aerodynamic surfaces were adapted from the one of [26]. The infinite plate spline (IPS) method was used to transfer loads and displacements between the

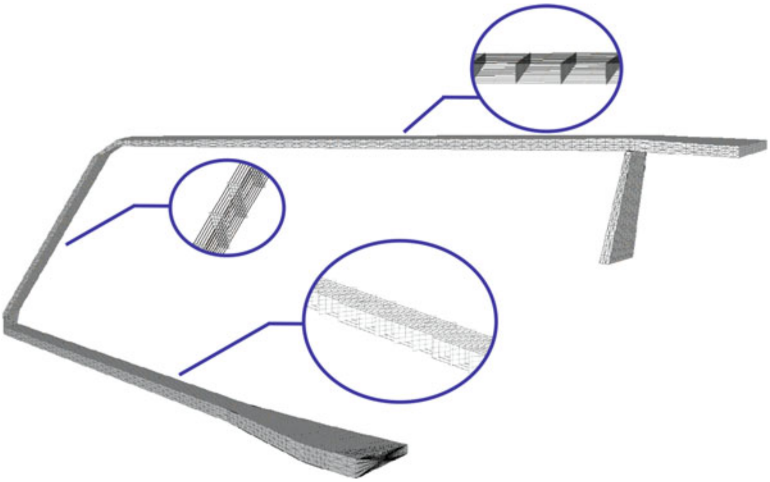


Fig. 7 PrP250v2Al structural model. *Circles* highlight the structure inside the wing box: 1D beam elements are used to model stringers and 2D plane elements to model spars and ribs

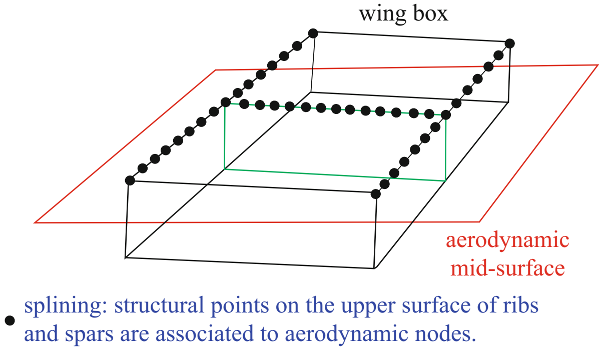

wing box
aerodynamic mid-surface
● splining: structural points on the upper surface of ribs and spars are associated to aerodynamic nodes.

Fig. 8 Splining for the PrP250v2Al model

structural and aerodynamic models. Figure 8 shows a conceptual sketch of this coupling: the structural nodes on the upper surface of the spars and the ribs have been associated with the nodes on the mean aerodynamic surface. Constraints to fix the structure follow the same logics as the one described in the previous section for the PrP250v1 case.

When fuselage inertial properties have to be considered, the following strategy was adopted. Linkages were used to connect the root of the fin and the wing section supported by the fuselage (the fuselage was idealized as rigid). Then, the configuration was able to have a rigid motion in the vertical axis (plunging) and a rigid rotation in the plane of symmetry (pitching). Of course, the constraints employed to fix the structure in the space were discarded.

The inertial data of the fuselage were extrapolated from [36]. To retain the inertial effects, two concentrated masses were connected with linkages to the fin root and wing-fuselage intersection. The exact position and the value of the masses were determined in such way to match the inertial properties of the fuselage (weight and pitching moment of inertia). Moreover, when different target pitching moment of inertia of the whole configuration were required to perform a sensitivity analysis, the masses were repositioned in such a way to avoid shifting of the CoG (and thus, the same value of the static margin of flight stability was maintained).

4.3 PrP250v2Comp

4.3.1 The Original Model

In effort [29] a realization of the wing-system structure of the PrandtlPlane was pursued using composite materials. This was not only driven by the aim of having a lighter structure, but also because with the aluminum version large displacements were observed in the limit load case. Due to the required high skin thicknesses the wing box was designed without stringer, increasing the number of ribs to avoid instability under compressive loads. More details can be found in [29, 39].

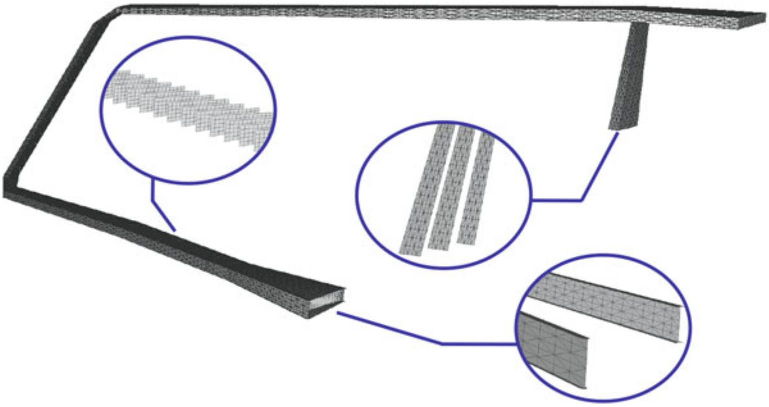


Fig. 9 PrP250v2Comp structural model. *Circles* highlight elements inside the wing box, such as ribs and spars (both webs and caps)

4.3.2 The PrP250v2Comp Model

The structural model used in cited references was provided. Small modifications in the number of finite elements were carried out to lighten the computational burden as the original model was built with the purpose of measuring stress levels. A sketch of the final model is given in Fig. 9.

Load and displacement transferring between the structural and aerodynamic meshes, constraints of the "fixed" case, and fuselage inertial effects of the "free" case, followed the same logics adopted for the PrP250v2Al case, as described in Sect. 4.2.2.

5 Flutter Analysis of the PrP250v1

This section is based on the findings shown in paper [28]. Part of these results are relevant as starting point for the analysis carried out in Sect. 6.

5.1 "Constrained" Model

Hereinafter "constrained" (or fixed) is used to indicate a configuration which is not free to move in the space. The opposite condition, in which the model is free to rigidly move (as the analyses are symmetric, the considered rigid-body motions are plunging and pitching), is called "free" case.

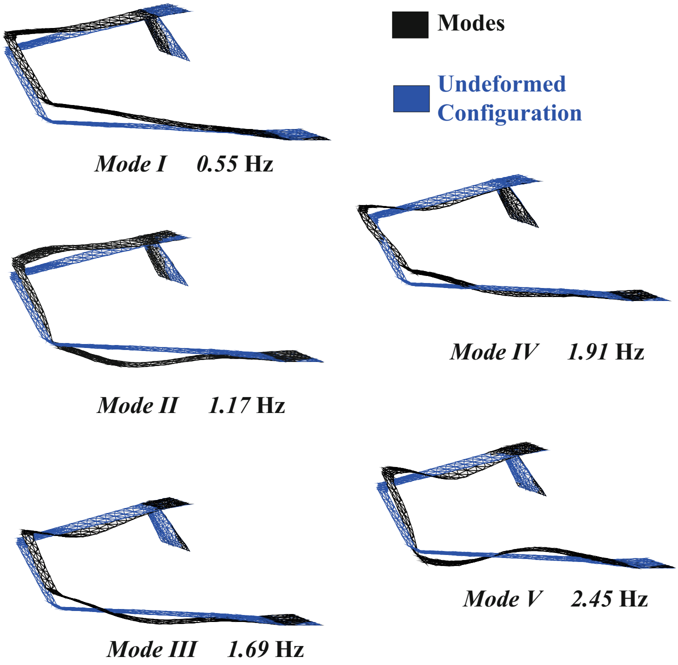


Fig. 10 Modal analysis of the PrP250v1

5.1.1 Modal Analysis

Outcome of the modal analysis of the PrP250v1 is given in Fig. 10. Low frequency of the first mode is observed. The mode has an in-phase bending of the two wings; conversely, the second mode has an out-of-phase deflection accompanied by a tilting of the vertical joint in the longitudinal plane.

5.1.2 Flutter

Flutter analysis was performed neglecting compressibility effects ($M = 0$). As it can be inferred from Fig. 11, mode II loses stability at a speed of approximately 257 m/s. A coalescence of modes I and II is observed. These results correlate qualitatively well with those presented in paper [26].

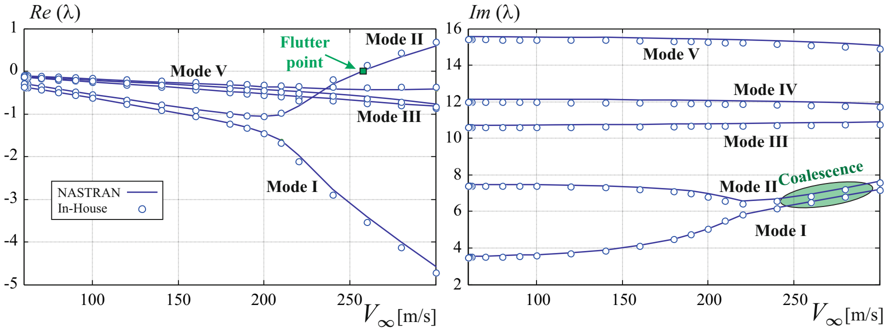


Fig. 11 Flutter analysis: real and imaginary parts of the eigenvalues

5.1.3 Energy Transfer and Limit Cycle Oscillation

For analyses in the postcritical regime a speed slightly higher than the flutter one was chosen (260 m/s). A disturbance in the angle of attack was given to trigger the instability, and the time response was tracked as shown in Fig. 12. A limit cycle oscillation develops after the transient, having a frequency of 1.1 Hz.

To gain more insight in the instability process, a wave in the time response (between $t = 16$ and 17.1 s in Fig. 13) was chosen and the power transferred by the fluid to the structure was calculated. Figure 14 shows the deformations (magnified 5×) and the power of aerodynamic forces at different snapshots taken in the interval. The upper wing alternated between energy extraction and transfer from and to the fluid. The lower wing, on the contrary, mostly extracts energy, especially in the tip region. Being energy transferring a possible way to interpret instability, this finding is relevant in characterizing flutter and, potentially, to design ad-hoc flutter suppression devices.

5.1.4 Effects of Freeplay

Freeplay of mobile surfaces may have a relevant impact on aeroelastic response of the aircraft. In this investigation, freeplay was considered only for the front aileron (the reader is referred to paper [28] for a more extensive treatise on the topic). Result of flutter analysis is shown in Fig. 15. Low speed instabilities relative to higher modes were observed. These instabilities, however, are easy to be dominated: modeling some source of structural damping would eventually fix the issue [28]. The most relevant aspect of this investigation is the increased speed (almost 10 %) of the main instability in respect to the case in which freeplay was not modeled. Figure 16 shows how coalescence of modes I and II were postponed.

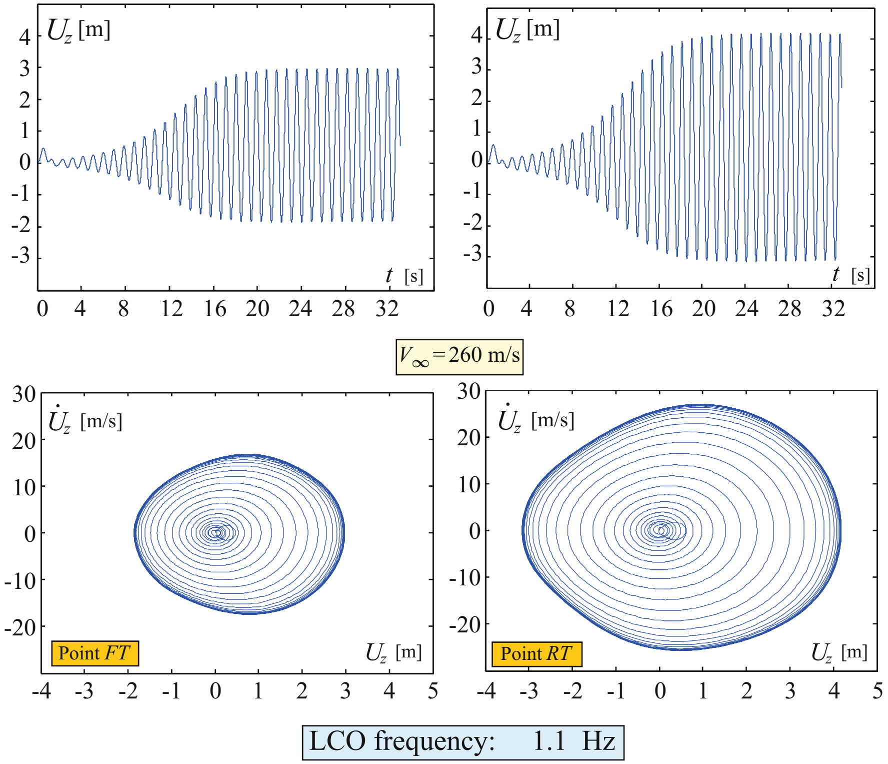


Fig. 12 Time response at 260 m/s after a perturbation in angle of attack is given: vertical displacement (U_z) of the tip of the front (*FT*) and rear (*RT*) wings. LCO is observed

It is not trivial to give an explanation of the reasons driving this increase in flutter speed. However, with reference to the energy diagram shown in Fig. 14, it is possible to speculate that freeplay of the front-wing aileron acts like a source of disturbance to the energy extraction mechanism, which was originally located in that area.

Cavallaro et al. [28] report the outcome of flutter studies when freeplay of all the control surfaces was considered. Instabilities always develop at lower speeds for small windows, and for particular combinations some extra source of damping was necessary to avoid their presence. In some other cases, instability occurred at lower speeds but, differently than the above described situation, persisted also in the higher-speed region. The *fundamental* instability with the interaction of modes I and II was still present at almost the same speed as for the cases without freeplay. Only freeplay on the front aileron seemed to have a significant effect postponing the flutter speed.

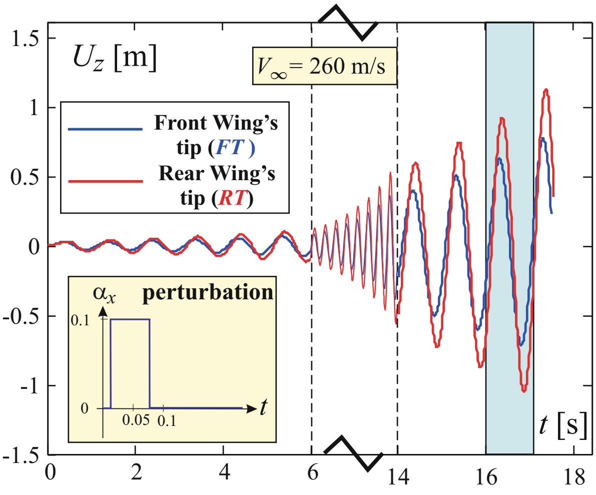


Fig. 13 Time response after a small perturbation in angle of attack is applied, speed is $V = 260\,\text{m/s}$. Interval between 16 and 17.1 s is used to track power transfer in a cycle

5.2 "Free" Model

When the configuration is considered free to move in space, rigid-body modes can interact with elastic ones changing the aeroelastic response compared to the case of constrained aircraft. Historically this was observed on flying-wing configurations [38] in the pre-World War II. This phenomenon is called BFF.

As observed in Sect. 5.1.1, the first elastic mode has a relatively low frequency. For this reason, an interaction with rigid-body modes can occur. The study conducted in [28] considered the PrP250v1 configuration free to pitch and plunge. The fuselage was considered rigid. Its inertia was extrapolated by previous works (see Sect. 4.1). Results of flutter analysis are shown in Fig. 17. No flutter instability was observed for the considered speed range. To better understand this behavior a sensitivity analysis was carried out in [28] considering fuselage mass and pitching moment of inertia.

Focusing on variation of pitching inertia (see Fig. 18) the following facts could be observed. For small values of pitching inertia, flutter occurred with a coalescence of the pitching and first elastic mode frequencies. On the other side of the spectrum, for large values the aeroelastic instability was associated with a coalescence of frequencies modes I and II, as already observed for the constrained case (in fact, in the limit process of infinite pitching moment of inertia, the configuration can be considered as constrained in pitching). In the nominal case, the moment of inertia falls between the two above limit cases. Mode I frequency did not get closer (coalesce) either to mode II or pitching mode frequencies; no aeroelastic instability was observed.

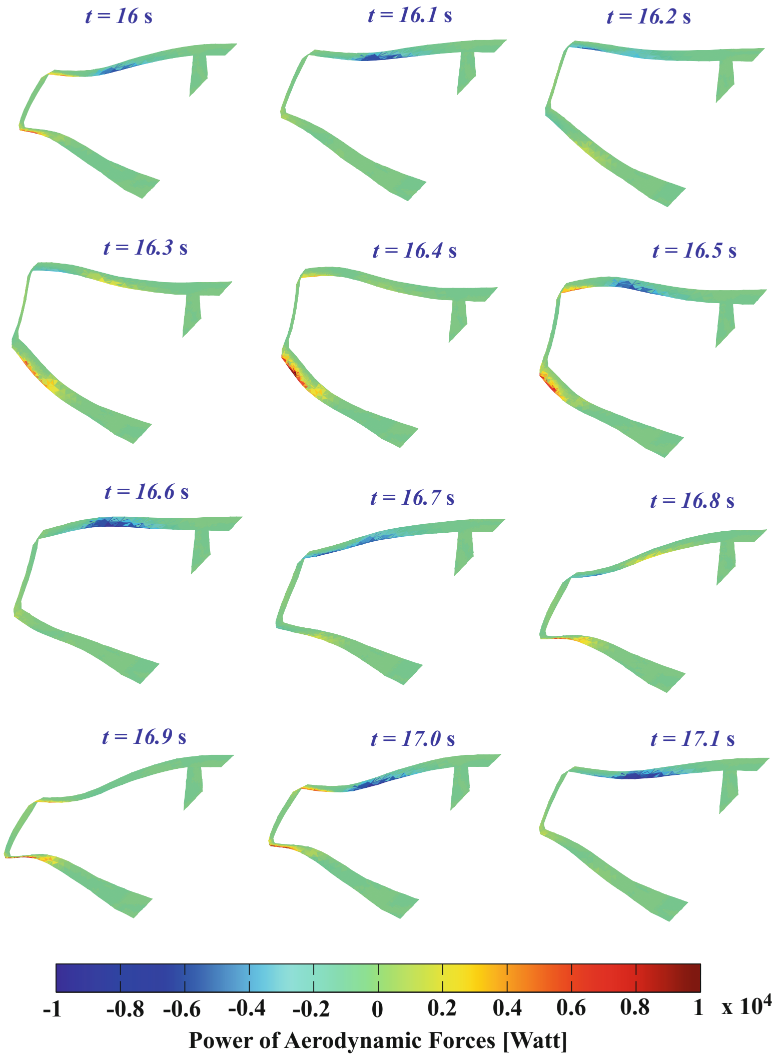


Fig. 14 Time response of the system for $V = 260\,\mathrm{m/s}$: magnified (5×) deformations and power of aerodynamic forces at different snapshots taken in the interval 16–17.1 s, see Fig. 13

What observed is interesting as it can turn a design with aeroelastic issues in a flutter-free one. For this specific case, thus, the typical flutter analysis considering the configuration constrained (sometimes called "cantilever" analysis) gives conservative results (at least a 20 % lower critical speed) when compared to the case of free aircraft (sometimes called "free-flying" analysis).

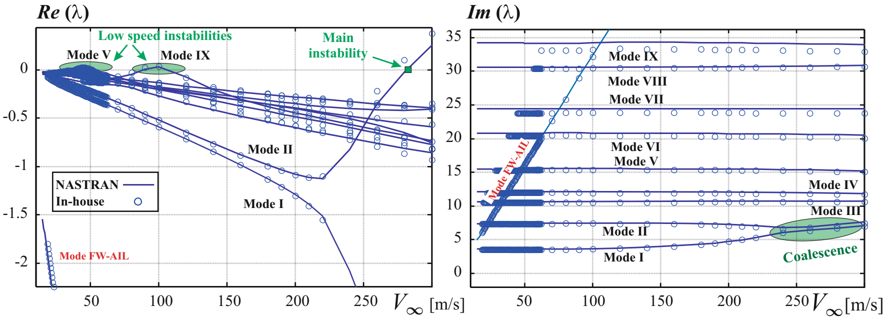


Fig. 15 Flutter analysis of the PrP250v1 when freeplay of the front aileron is considered. Real and imaginary parts of the eigenvalues of the system (the acronym FW-AIL refers to the free rotation of the forward aileron)

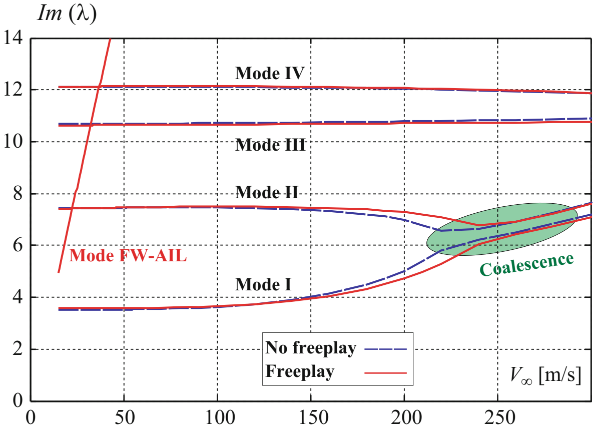


Fig. 16 Imaginary parts of the eigenvalues of the system at different speeds for cases in which front aileron has or has not freeplay (the acronym FW-AIL refers to the free rotation of the forward aileron)

6 Flutter Analysis of the PrP250v2Al

The relevant findings presented in the previous section have potentially a great impact on the aeroelastic design of the PrandtlPlane. Those results have been obtained on a configuration “similar” (from an aeroelastic point of view) to the one outcome of a structural optimization [21] and described by beams modeling the inertial and stiffness properties of the wing system. It is relevant to verify the above conceptual findings on a more refined model. Thus, the model presented in Sect. 4.2.2, describing the wing box with a good level of detail, represents an ideal choice.

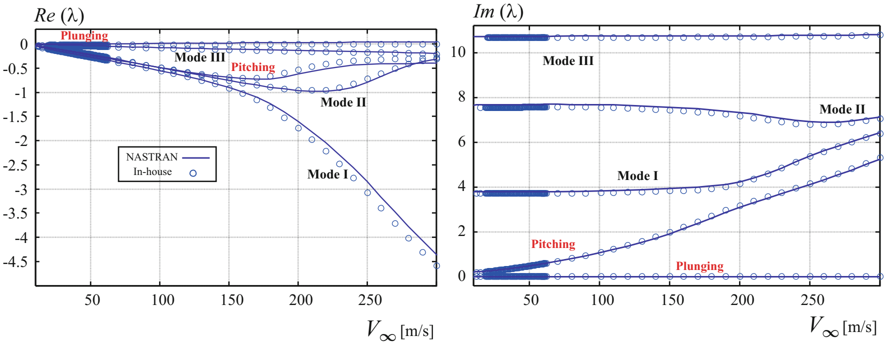


Fig. 17 "Free" PrP250v1 flutter analysis: real and imaginary parts of the eigenvalues, at different speeds

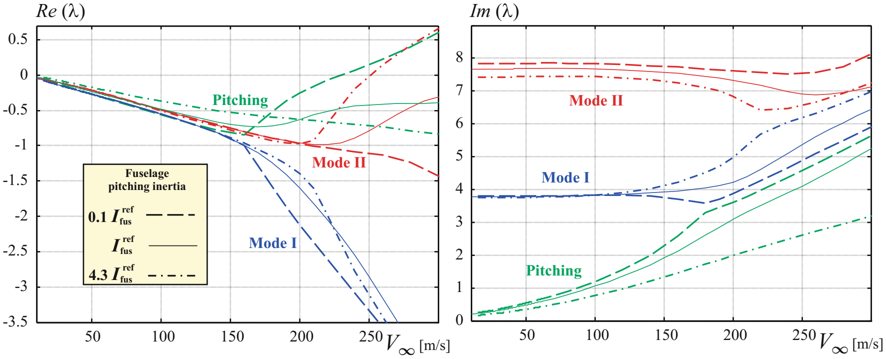


Fig. 18 "Free" PrP250v1 flutter analysis: sensitivity to pitching moment of inertia

It has to be underlined that the detailed model was built starting from the configuration obtained by the optimizer when constraints on maximum stress and stability were considered [29]. Thus, it is not expected this configuration to be flutter free.

6.1 "Constrained" Model

6.1.1 Modal Analysis

A preliminary modal analysis shows low frequencies relative to the first elastic modes. The shape of the modes (shown in Fig. 19) resembles the one already observed in previous investigations, as described in Sect. 5.1.1.

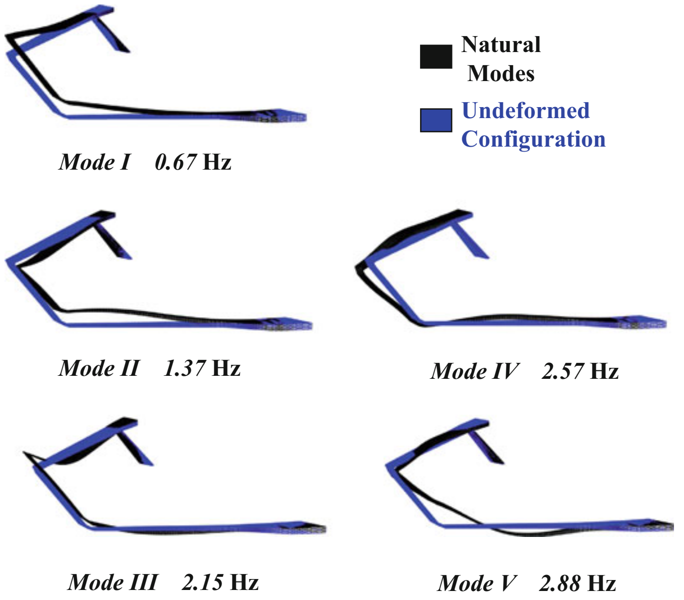


Fig. 19 PrP250v2Al: first five natural modes and relative frequencies

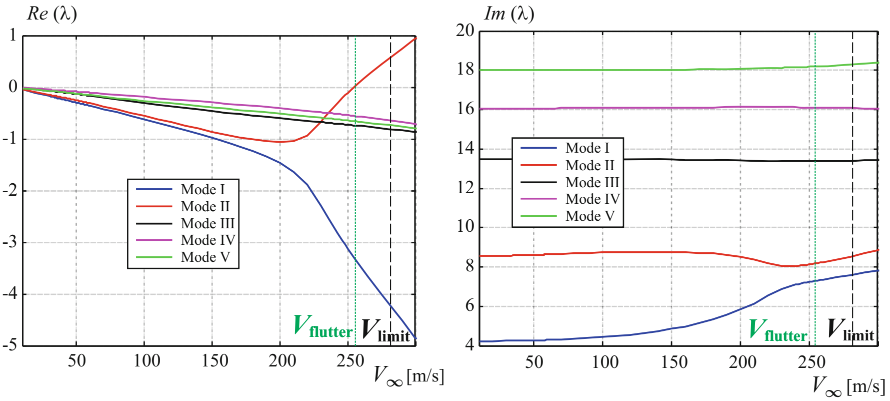


Fig. 20 Incompressible ($M = 0$) flutter analysis for the "constrained" PrP250v2Al. Real and imaginary parts of the eigenvalues. V_{limit} refers to the limit speed at sea level as prescribed by the regulations

6.1.2 Flutter (No Compressibility Effects)

The first flutter analysis was carried out considering $M = 0$, neglecting, thus, effects of compressibility. Figure 20 shows the real and imaginary parts of the eigenvalues versus the speed (relative to the sea level).

It can be inferred that flutter occurs approximately at $V = 255\,\mathrm{m/s}$, and the second elastic mode became unstable. Frequencies of the first and second elastic modes tend to coalesce. This scenario is extremely similar to the one of the PrP250v1 model.

6.1.3 Matched Flutter Analysis

Effects of Mach number have to be properly taken into account for a meaningful analysis. This is here pursued considering only linearized aerodynamics.

Regulations

Following the JAR-25, the aeroplane must be designed to be free from flutter with an appropriate margin (15 %, calculated on the equivalent airspeed—EAS) considering the flight envelope. At cruise level the limiting factor is the dive Mach number M_D, equal to the cruise Mach number increase by 0.05, and thus, $M_D = 0.9$. Being the speed of sound at cruise level equal to $297.4\,\mathrm{m/s}$, this means a diving speed of V_D of $267.5\,\mathrm{m/s}$. The relative EAS enlarged by 15 % is $173.1\,\mathrm{m/s}$.

At sea level, the calculation is based on the diving speed, which is conveniently calculated as suggested in [40]. The V_D is $245.1\,\mathrm{m/s}$, and, enlarging the limit by 15 %, the limit speed of $281.9\,\mathrm{m/s}$ is obtained.

Flutter Analysis

Analyses are performed considering different Mach numbers (for a fixed altitude), and the relative flutter speeds are found. For each of them, the Mach number is obtained. The *matched* flutter occurs when the Mach number relative to the flutter speed is equal to the one used for the flutter analysis. Table 1 shows the results of the analysis at the sea level. The critical flutter condition is $V_\infty = 244\,\mathrm{m/s}$, relative to $M = 0.71$. The relative diagrams are shown in Fig. 21. Flutter mechanism is the same as described with the initial incompressible case.

Table 1 Flutter speeds at several Mach numbers for the "constrained" PrP250v2Al. Sea level. The bold line is the matched flutter condition

Mach (input)	Flutter speed (m/s)	Relative Mach
0.55	248	0.73
0.60	248	0.73
0.65	246	0.72
0.70	244	0.71
0.71	**244**	**0.71**
0.75	242	0.71
0.80	242	0.71
0.85	238	0.69

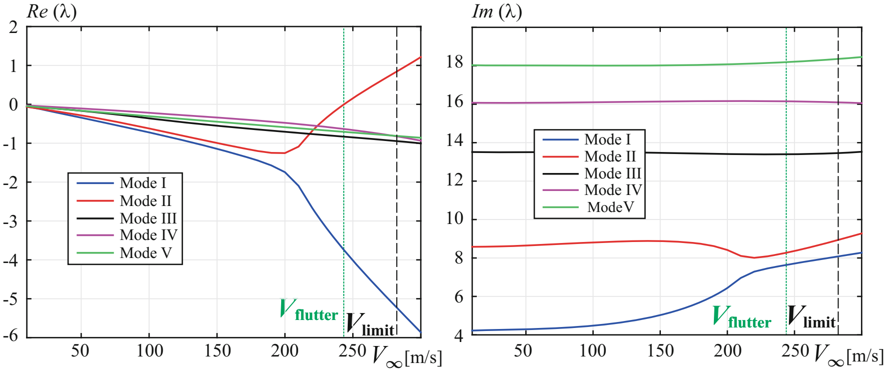


Fig. 21 Flutter analysis for the "constrained" PrP250v2Al at sea level for M=0.71. Real and imaginary part of the eigenvalues. V_{limit} refers to the speed limit at sea level as prescribed by regulations

Considering the cruise condition, flutter is not observed. Details are here omitted for brevity.

6.1.4 Considerations

Analyses showed that this configuration does *not* comply with regulations. At sea level, the speed limit under which flutter must not occur is $V_{\text{limit}} = 281.9\,\text{m/s}$, whereas flutter occurs at $V_\infty = 244\,\text{m/s}$. As already stated, it was not expected this configuration to be flutter free, as in the original structural design flutter constraint was not considered.

6.2 "Free" Model

Considering the results shown in Sect. 5.2, the question to be asked is if the fuselage inertial properties influence aeroelastic behavior of a PrandtlPlane to such large extent. Given these premises, confirmation of the previous trends are sought. First, the nominal pitching moment of inertia is considered, i.e., $1.22 \cdot 10^7\,\text{kg}\,\text{m}^2$.

6.2.1 Modal Analysis

Results of modal analysis are shown in Fig. 22. Compared to the "constrained" case, elastic modes do not change in a sensitive way, and only very minor differences in the frequencies are observed.

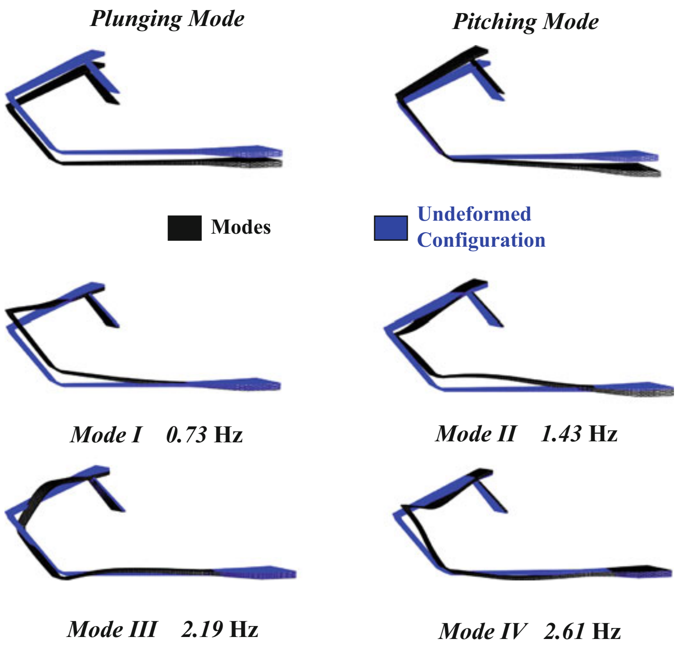


Fig. 22 "Free" PrP250v2Al: plunging and pitching modes, first four elastic modes and relative frequencies

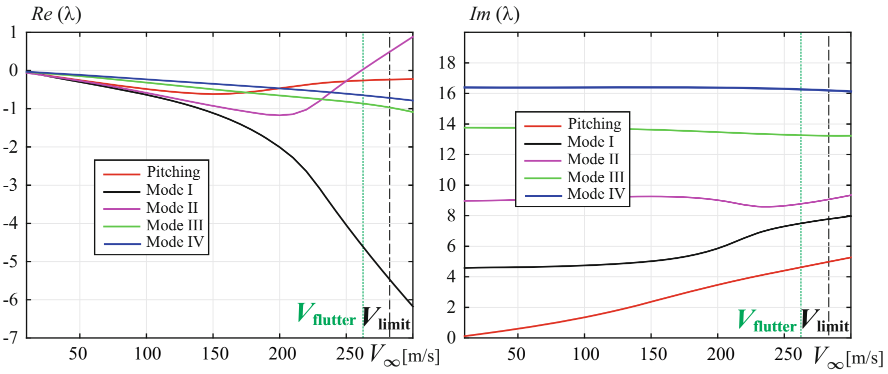


Fig. 23 Flutter analysis for the "free" PrP250v2Al at sea level for M=0.77. Real and imaginary part of the eigenvalues. V_{limit} refers to the speed limit at sea level as prescribed by the regulations

6.2.2 Matched Flutter Analysis

Figure 23 summarizes the critical condition (outcome of the matched flutter analysis) at sea level. Now the critical speed is $V = 262\,\text{m/s}$ (Mach number of 0.77), an increase of approximately 8 % in respect to the "constrained" case.

The mechanism is similar to the "constrained" case, mode II becomes unstable after showing a strong interaction with mode I. This mechanism reminds the one already seen in Sect. 5.2.

Results for the cruise condition are not shown, as in such case the configuration complies with flutter requirements.

6.2.3 Sensitivity to Fuselage Pitching (Moment of) Inertia

In order to gain further insight, a sensitivity analysis is performed changing the value of the pitching moment of inertia of the fuselage, as discussed also in Sect. 4.2.2. For each condition, a matched flutter analysis at sea level is performed. Results are summarized in Fig. 24. It is evident a change in flutter mechanism when varying the fuselage moment of inertia. For smaller than the nominal values, pitching mode becomes unstable (BFF) *following* a strong interaction (the frequencies of the modes getting very close) with the first elastic mode. On the opposite side, with larger moments of inertia, the behavior approaches the one of the "constrained" case, with the second elastic mode becoming unstable after a strong interaction with the first one. When the value of the moment of inertia is between these limiting cases, flutter speed increases. Looking at the imaginary part of the eigenvalues at different speeds, it may be noticed that the strong frequency interaction between mode I and pitching mode (small pitching inertia case) or mode I and II (large pitching inertia) seems to be milder.

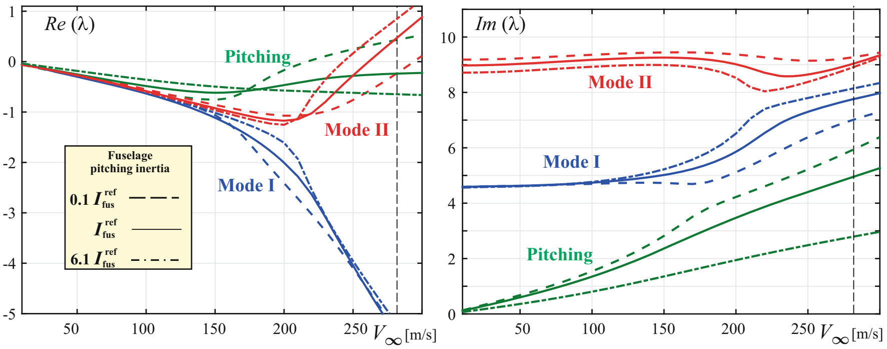


Fig. 24 "Free" PrP250v2Al flutter analysis: sensitivity to pitching inertia

7 Flutter Analysis on the PrP250v2Comp

In this section the model designed with composite material is aeroelastically tested.

7.1 "Constrained" Model

7.1.1 Modal Analysis

Results of modal analysis are shown in Fig. 25. Frequencies of the elastic modes are significantly higher than the ones of the aluminum version. Also, a larger separation between the first two frequencies is observed.

7.1.2 Flutter

A preliminary flutter analysis is carried out considering $M = 0$. Results are shown in Fig. 26. The configuration is flutter free. Even when compressibility effects are taken into account, flutter-free condition is found (for the sake of conciseness these results are not shown). Thus, the configuration built with composite materials, although not specifically designed against flutter, is flutter free.

7.2 "Free" Model

Analyses are now repeated considering the configuration free to plunge and pitch, and taking into account the inertial properties of the fuselage. In Sect. 4.3 the model is described in detail.

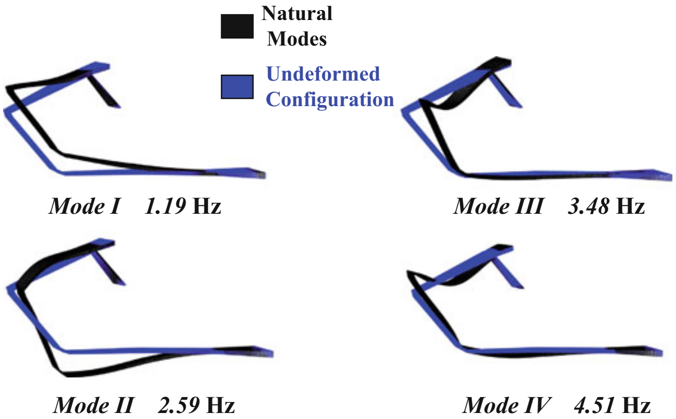


Fig. 25 PrP250v2Comp: first four natural modes and relative frequencies

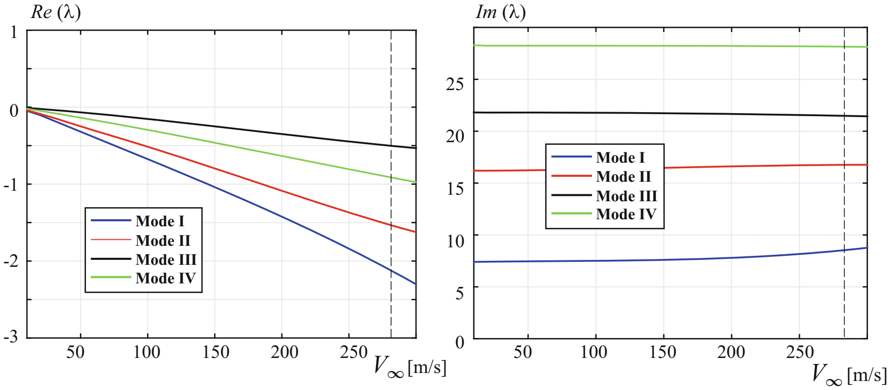


Fig. 26 Incompressible ($M = 0$) flutter analysis for the "constrained" PrP250v2Comp. Real and imaginary part of the eigenvalues

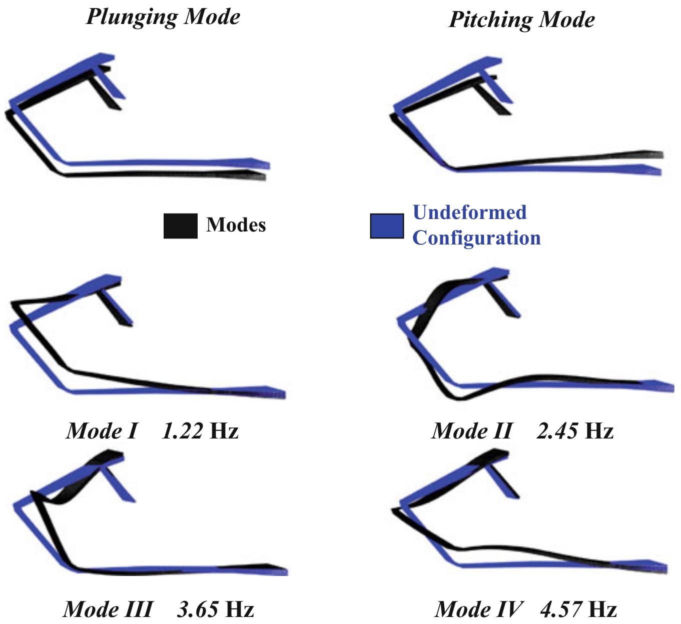


Fig. 27 "Free" PrP250v2Comp: plunging and pitching modes, first four elastic modes and relative frequencies

7.2.1 Modal Analysis

Results of modal analysis are shown in Fig. 27. A comparison with the "constrained" case shows relatively unvaried elastic modes with only very minor differences in frequencies.

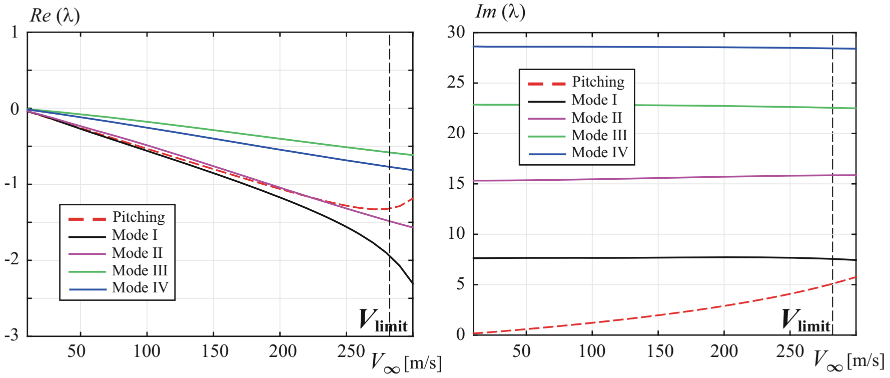


Fig. 28 Flutter analysis for the "free" PrP250v2Comp at sea level (M=0, compressibility effects are neglected). Real and imaginary part of the eigenvalues. V_{limit} refers to the speed limit at sea level as prescribed by the regulations

7.2.2 Flutter Analysis

Flutter analysis does not show any instability for the configuration. In Fig. 28 results are reported of the analysis carried out at sea level not considering compressibility effects, i.e., $M = 0$. Effects of compressibility do not change the main outcome: the configuration is flutter free.

8 Conclusions

In this work several flutter analyses of a PrandtlPlane configuration have been carried out. Considering a realization of the wing system in aluminum, aeroelastic properties have been studied on a constrained model (no rigid-body motion possible). The typical flutter occurred with a strong interaction between the first and the second elastic modes. The first natural mode presented a deformation characterized by an in-phase bending of the two wings, whereas the second one showed a strong tilting of the lateral joint in the longitudinal plane. The energy transferred by the fluid was tracked at a speed immediately higher than the flutter one, and it was noticed that the tip region of the front wing was active in introducing energy into the structure.

Freeplay of control surfaces was then considered, and its effect on flutter was studied. In the case of the front-wing aileron, flutter occurred with the same mechanism (instability of the second mode after a strong interaction with the first one) as for the case without freeplay, but the critical speed increased.

Aeroelastic analyses were then repeated considering the configuration free to plunge and pitch, and taking into account fuselage inertial properties. Flutter speed was found to be significantly higher. The interaction between pitching, first and second elastic modes resulted in a postponed loss of stability of the second mode. A sensitivity analysis varying the fuselage pitching moment of inertia better showed this behavior.

A more detailed model of a similar wing system was used to investigate the aeroelastic properties of the PrandtlPlane and, moreover, compressibility effects were taken into account. Results and trends seen on the lower fidelity model were confirmed.

In the last part of this study, a structural model of the same PrantlPlane made of composite materials was analyzed. The configuration was found to be flutter free.

8.1 Future Research

Future research to integrate the here presented results should focus on the lateral-directional dynamics. Work [24] did some preliminary analyses in this sense, revealing interesting aspects. Moreover, it is felt the need for a more integrated approach for the study of the aeroelasticity and flight dynamics properties of a free-flexible aircraft.

A topic to be studied concerns also the fuselage elasticity. Its inertial effects have been taken into account, however, what would happen when also its flexibility is considered? How would the aeroelastic response change?

These are only a few aspects among the many that need to be investigated to shed some light and gain a better understanding of the aeroelastic behavior of such unconventional configurations.

References

1. Frediani, A., Cipolla, V., Rizzo, E.: The PrandtlPlane configuration: overview on possible applications to civil aviation. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications, vol. 66, pp. 179–210. Springer US, New York (2012). doi:10.1007/978-1-4614-2435-2_8
2. Cavallaro, R., Demasi, L.: Challenges, ideas, and innovations of joined-wing configurations: a concept from the past, an opportunity for the future. Prog. Aerosp. Sci. (2016, accepted for publication) http://dx.doi.org/10.1016/j.paerosci.2016.07.002
3. Prandtl, L.: Induced drag of multiplanes. Technical Report TN 182, NACA, March 1924, reproduction of Der induzierte Widerstand von Mehrdeckern. Technische Ber. 3, pp. 309–315 (1918)
4. Demasi, L., Monegato, G., Dipace, A., Cavallaro, R.: Minimum induced drag theorems for joined wings, closed systems, and generic biwings: theory. J. Optim. Theory Appl. **169**(1), 200–235 (2016)
5. Frediani, A., Montanari, G.: Best wing system: an exact solution of the Prandtl's problem. In: Variational Analysis and Aerospace Engineering. Springer Optimization and Its Applications, vol. 33, pp. 183–211. Springer, New York (2009)

6. Demasi, L., Monegato, G., Cavallaro, R.: Minimum induced drag theorems for multi-wing systems. In: 57th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, No. AIAA 2016-0236. AIAA SciTech, San Diego, CA (2016)
7. Demasi, L., Monegato, G., Cavallaro, R.: Minimum induced drag theorems for nonplanar systems and closed wings. In: Frediani, A. (ed.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications. Springer US, New York (2016, to appear)
8. Munk, M.: Isoperimetrische Aufgaben aus der Theorie des Fluges. Dieterichsche Universitäts-Buchdruckerei, Göttingen (1919)
9. Munk, M.: The minimum induced drag of aerofoils. Report 121, NASA (1923)
10. Bernardini, G.: Problematiche aerodinamiche relative alla progettazione di configurazioni innovative. Ph.D. thesis, Politecnico di Milano (1999)
11. Bernardini, G., Frediani, A., Morino, L.: Aerodynamics for MDO of an innovative configuration. In: Research and Technology Organization, No. RTO Meeting Proceedings 35. RTO AVT Symposium on Aerodynamic Design and Optimisation of Flight Vehicles in a Concurrent Multi-Disciplinary Environment. Symposium of the Applied Vehicle Technology Panel, Ottawa (1999)
12. Cavallaro, R., Nardini, M., Demasi, L.: Amphibious PrandtlPlane: preliminary design aspects including propellers integration and ground effect. In: 2th SciTech2015, No. AIAA-2015-1185. Kissimmee, FL (2015)
13. Cavallaro, R., Iannelli, A., Demasi, L., Razón, A.M.: Phenomenology of nonlinear aeroelastic responses of highly deformable joined wings. Adv. Aircr. Spacecr. Sci. **2**(2), 125–168 (2015)
14. Frediani, A., Gasperini, M., Saporito, G., Rimondi, A.: Development of a PrandtlPlane aircraft configuration. In: XVII Congresso Nazionale AIDAA (17th National Congress AIDAA), Roma, pp. 2089–2104 (2003)
15. Gagnon, H., Zingg, D.W.: Aerodynamic optimization trade study of a box-wing aircraft configuration. In: AIAA SciTech, 56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, No. AIAA2015-0695 (2015)
16. Gall, P.D., Smith, H.C.: Aerodynamic characteristics of biplanes with winglets. J. Aircr. **24**(8), 518–522 (1987)
17. Chiocchia., G., Iuso, G., Carrera, E., Frediani, A.: A wind tunnel model of a ULM configuration of Prandt plane: design, manufacturing and aerodynamic testing. In: XVII Congresso Nazionale AIDAA (17th National Congress AIDAA), Rome, pp. 2089–2104 (2003)
18. Cipolla, V., Frediani, A., Oliviero, F., Gibertini, G.: Ultralight amphibious PrandPrandtl: wind tunnel tests. In: XXII Conference, Italian Association of Aeronautics and Astronautics, Napoli (2013)
19. Lange, R.H., Cahill, J.F., Bradley, E.S., Eudaily, R.R., Jenness, C.M., Macwilkinson, D.G.: Feasibility study of the transonic biplane concept for transport aircraft applications. NASA CR–132462. Lockheed-Georgia Company, Marietta, GA (1974)
20. Frediani, A., Rizzo, E., Bottoni, C., Scanu, J., Iezzi, G.: A 250 passenger PrandtlPlane transport aircraft preliminary design. Aerotecnica Missili e Spazio (AIDAA) **84**(4) 152–163 (2005)
21. Dal Canto, D., Frediani, A., Ghiringhelli, G.L., Terraneo, M.: The lifting system of a PrandtlPlane, Part 1: design and analysis of a light alloy structural solution. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications, vol. 66, pp. 211–234. Springer US, New York (2012). doi:10.1007/978-1-4614-2435-2_9
22. Voskuijl, M., Klerk, J., Ginneken, D.: Flight mechanics modeling of the PrandtlPlane for conceptual and preliminary design. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications, pp. 435–462. Springer US, New York (2012)
23. Dimartino, C., Baldini, M.: Analisi agli elementi finiti di un tronco di fusoliera di un velivolo PrandtlPlane sottoposto a carichi limite di pressurizzazione e di massa. Master's thesis, Universitä di Pisa (2009)

24. Bombardieri, R., Cavallaro, R., Demasi, L.: A historical perspective on the aeroelasticity of box wings and PrandtlPlane with new findings. In: 57th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, No. AIAA 2016-0238. AIAA SciTech, San Diego, CA (2016)
25. Dal Canto, D.: Progetto preliminare del cassone alare di un velivolo di tipo Prandtl-Plane mediante l'applicazione di un metodo di ottimizzazione strutturale. Master's thesis, Dipartimento di Ingegneria Aerospaziale, Università di Pisa (2009). Advisor: A. Frediani
26. Divoux, N., Frediani, A.: The lifting system of a PrandtlPlane, Part 2: Preliminary study on flutter characteristics. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications, vol. 66, pp. 235–267. Springer US, New York (2012). doi:10.1007/978-1-4614-2435-2_10
27. Divoux, N.: Preliminary study on flutter characteristics of a PrandtlPlane aircraft. Master's thesis, TU Delft (2008)
28. Cavallaro, R., Bombardieri, R., Demasi, L., Iannelli, A.: PrandtlPlane joined wing: body freedom flutter, limit cycle oscillation and freeplay studies. J. Fluids Struct. **59**, 57–84 (2015)
29. Frediani, A., Quattrone, F., Contini, F.: The lifting system of a PrandtlPlane, Part 3: structures made in composites. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications, vol. 66, pp. 269–288. Springer US, New York (2012). doi:10.1007/978-1-4614-2435-2_11
30. Bombardieri, R.: PrandtlPlane joined wing: body freedom flutter, limit cycle oscillation and freeplay studies. Master's thesis, Dipartimento di Ingegneria Aerospaziale, Univerisitä di Pisa (2015). Advisors: Aldo Frediani, Luciano Demasi and Rauno Cavallaro
31. Silvani, S.: Aeroelastic analysis of PrandtlPlane joined wings configuration. Master's thesis, Università degli Studi di Roma 3 (2015)
32. Rodden, W.P., Johnson, E.H.: User Guide V 68 MSC/NASTRAN Aeroelastic Analysis. MacNeal-Schwendler Corporation, Newport Beach, CA (1994)
33. Cavallaro, R., Iannelli, A., Demasi, L., Razón, A.M.: Phenomenology of nonlinear aeroelastic responses of highly deformable joined-wings configurations. In: 55th AIAA/ASMe/ASCE/AHS/SC Structures, Structural Dynamics, and Materials Conference, No. AIAA 2014-1199. AIAA Science and Technology Forum and Exposition (SciTech2014) National Harbor, MD (2014)
34. Bottoni, C., Scanu, J.: Preliminary design of a 250 passenger PrandtlPlane aircraft. Master's thesis, University of Pisa (2004)
35. Ginneken, D.A.J., Voskuijl, M., Van Tooren, M.J.L., Frediani, A.: Automated control surface design and sizing for the PrandtlPlane. In: 51st AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference, No. AIAA 2010-3060, Orlando, FL (2010)
36. Ginneken, D.V.: Automated control surface design and sizing for the PrandtlPlane. Master's thesis, TU Delft (2009)
37. Chipman, R., Rauch, F., Rimer, M., Muñiz, B.: Body-freedom flutter of a 1/2 scale forward-swept-wing model, an experimental and analytical Study. Contract Report NASA CR-172324, NASA, Grumman Aerospace Corporation, April 1984, http://naca.larc.nasa.gov/search.jsp?R=19840013485&qs=Ns%3DAcquired-Date%7C1%26N%3D4294950357%2B4294929324
38. Love, M.H., Zink, P.S., Wieselmann, P.A., Youngren, H.: Body freedom flutter of high aspect ratio flying wings. In: 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Material Conference, No. AIAA 2005-1947, Austin, TX (2005)
39. Quattrone, F., Contini, F.: Preliminary design and FEM analysis of a new conception non-standard wing structure: the PrandtlPlane 250 wing structure. Master's thesis, Dipartimento di Ingegneria Aerospaziale, Università di Pisa (2010)
40. Torenbeek, E.: Synthesis of Subsonic Airplane Design: An Introduction to the Preliminary Design of Subsonic General Aviation and Transport Aircraft, with Emphasis on Layout, Aerodynamic Design, Propulsion and Performance. Springer Netherlands, Dordrecht (1982)

HyPSim: A Simulation Tool for Hybrid Aircraft Performance Analysis

Vittorio Cipolla and Fabrizio Oliviero

Abstract This work presents the performance prediction activities carried out by the research team of Pisa University within the Euopean project "HYPSTAIR", concerning the development and validation of hybrid propulsion system components and sub-systems for electrical aircraft. The first part of the paper discusses the performance analysis of a serial hybrid general aviation airplane for a reference mission profile. In particular, the best flight performance is evaluated varying the relevant mission parameters (e.g. range, cruise altitude, and cruise speed) and the amount of available energy, in terms batteries and fuel. In the second part, a hybrid plane simulator, conceived to implement different mission profiles and to include pilot effects on power management by adopting a human-in-the-loop approach, is presented. Such simulator consists of three main software modules linked to each other in real time: a flight simulator, used to compute the aerodynamic forces and to visualize the airplane in flight, a flight planner, in which the mission profile can be defined, and a performance module, which calculates the instantaneous consumption of energy and provides the endurance prediction.

1 Introduction

In the *HYPSTAIR* project, the object of study is a general aviation aircraft in which the propulsion system is made of a propeller driven by an electric motor, that can be fed by both batteries and an internal combustion engine (ICE) used as generator. The architecture of the hybrid system is serial, which means that the ICE is not directly connected to the propeller but it is used as a source of electric power (Fig. 1).

V. Cipolla (✉)
Department of Civil and Industrial Engineering, Aerospace Section, University of Pisa,
Via G. Caruso 8, 56122 Pisa, Italy
e-mail: vittorio.cipolla@for.unipi.it

F. Oliviero
Faculty of Aerospace Engineering, Flight Performance and Propulsion (FPP), TU Delft,
Kluyverweg 1, 2629HS Delft, The Netherlands

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_4

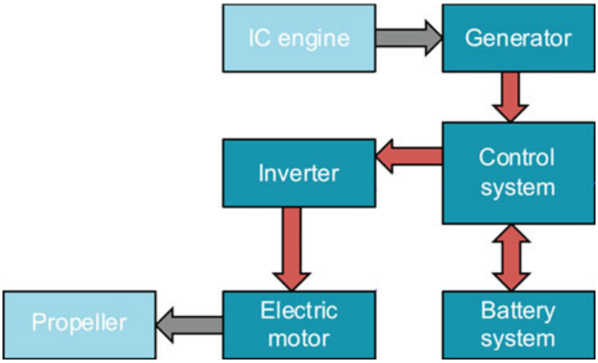


Fig. 1 Serial hybrid architecture

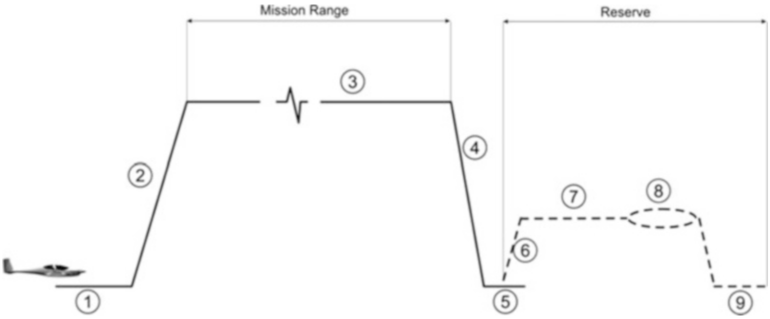


Fig. 2 Reference mission for the hybrid aircraft

This architecture gives more flexibility in terms of power management and it can provide a significant reduction of environmental impact, increasing safety at the same time. In fact, the presence of two independent energy sources introduces a redundancy and, in addition, an electric motor is more reliable than a piston engine.

1.1 Preliminary Analysis

The design of the hybrid system depends on mission requirements such as flight range, cruise altitude, and speed.

Concerning the novelty of the system here considered, a preliminary performance analysis has been performed in order to define the operating requirements limitations and to optimize the use of different energy sources. As detailed in [1], such analysis has been carried out implementing a simple but reliable performance model and considering the aerodynamic characteristic of an existing aircraft and the reference mission as shown in Fig. 2, where the adopted assumptions are indicated in Table 1.

Table 1 Hypotheses on mission segments

Segment	Parameters	Hypotheses
Climb	H_{in}, H_{fin}, P_{batt}	Fast climb
Cruise	V_{cruise} or P_{cruise}, H_{cruise}, range	Constant speed (or power) and altitude
Descent	–	Negligible for energy calculation
Landing	–	Negligible for energy calculation
Diversion	Range, H_{div}	Minimum power
Loiter	Time, H_{loi}	Maximum endurance

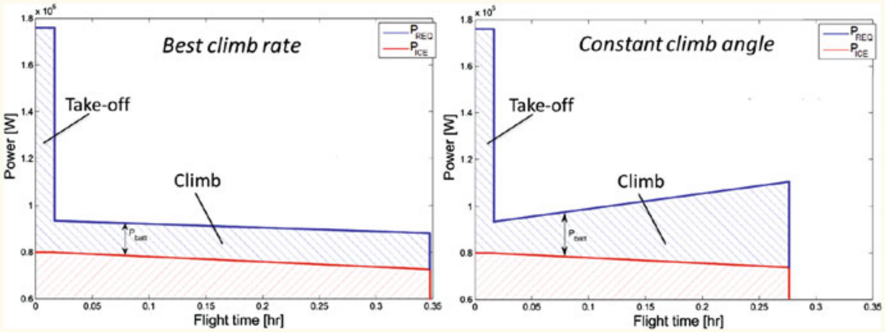


Fig. 3 Effect of different climb programs on power sources

The operating modes of the entire powertrain have been varied in accordance with the power request along the mission: battery packs provide additional energy during the most power demanding flight segments (take-off and climb), while during cruise the ICE generator gives sufficient power for both flight and battery charge.

The energy consumption has been evaluated through an energy balance. In particular, the efficiencies of all the components of the power architecture have been taken into account to define the mission parameters providing the best performance in terms of range or take-off weight. Therefore, two problems have been studied:

- evaluating the maximum flight range achievable with given amount of available energy (fuel + batteries) at take-off;
- evaluating the minimum energy amount (fuel + batteries) required at take-off in order to fly for a given range.

It has been observed that batteries affect performance mostly in flight segments such as climb or first part of the cruise, when batteries recharge occurs (see Fig. 3).

Therefore, as Fig. 4 suggests, differences between a hybrid and an internal combustion propulsion are more evident when the mission range is smaller. For long range mission, indeed, the energy contribution of batteries is less significant and flight performance is largely dominated by the ICE used as generator.

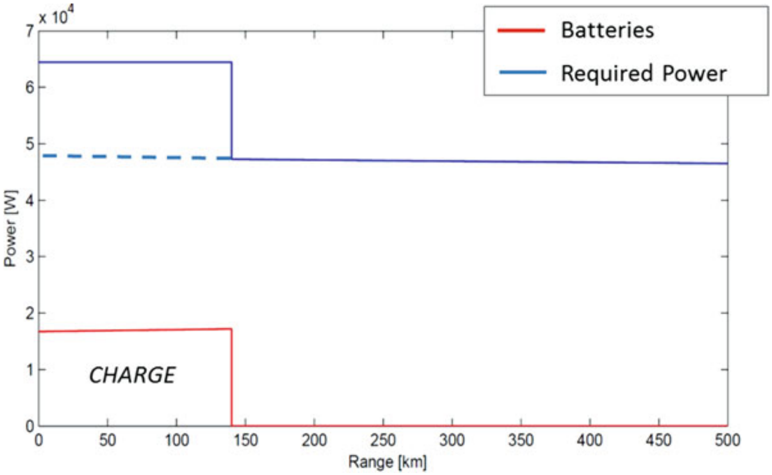


Fig. 4 Required power vs Range

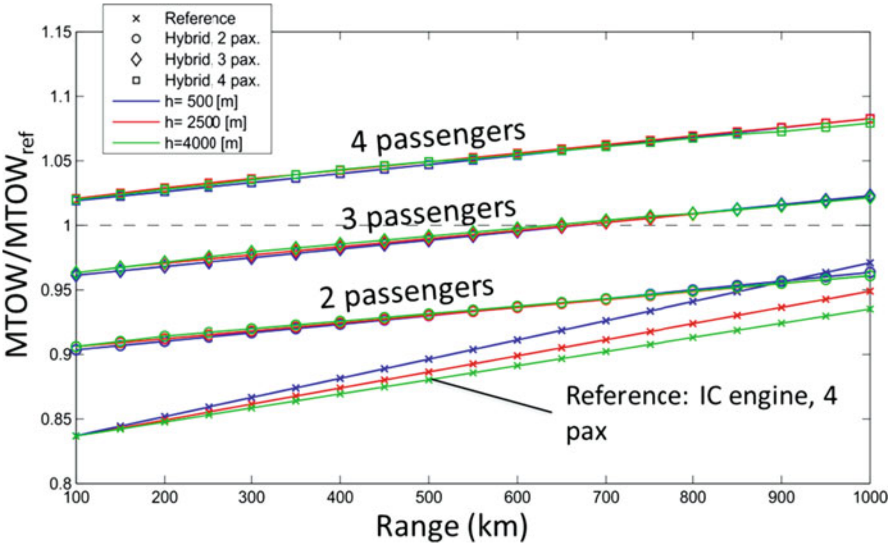


Fig. 5 Flexibility analysis of the hybrid aircraft and comparison with traditional propulsion

Finally, Fig. 5 shows the flexibility analysis of the hybrid aircraft compared to a traditional propulsion version, whose maximum Trke-off weight has been indicated as MTOW$_{ref}$.

As a preliminary result, the following conclusions can be given:

- if compared to internal combustion propulsion, the hybrid solution has significant influence on climb performance, whereas effects on cruise segment are smaller;

- since hybrid propulsion is less sensitive to altitude and range requirement has a weak influence on MTOW, the hybrid solution is more flexible than the traditional one;
- batteries energy and power densities play a key role.

1.2 Overview of the Simulator `HyPSim`

The simulator `HyPSim` (Hybrid Plane Simulator) has been set up in order to:

- validate the previously achieved results in order to meet the given requirements;
- evaluate the performance for different mission profiles (defined by the user and performed manually or using an autopilot);
- simulate the human-in-the-loop effects on flight performance and, in particular, on power management;
- simulate the instantaneous performance of the aircraft depending on the instantaneous battery state of charge (SoC);
- be used as a dissemination tool with user-friendly interfaces.

1.3 `HyPSim` Architecture

As described in [2], `HyPSim` is composed of three main software: a *Flight Simulator*, in which the airplane is displayed and flight data are calculated (position, angles, speed, forces, etc.); a *Flight Planner* which allows to define the mission profile and the flight mode (manual or automatic); a *Performance Module* in which the hybrid propulsion system is modelled by means of analytical relations and flight data are processed for performance estimations and endurance prediction.

Finally, the main flight parameters, such as flight speed and fuel/battery consumption, are shown on a human–machine interface (HMI) panel. The conceptual arrangements of the simulator are reported in Fig. 6.

The simulation is performed through the following process:

- the pilot indicates a reference mission at the beginning by indicating a set of waypoints (longitude, latitude, and altitude);
- a first estimation of endurance is provided and mission feasibility is evaluated;
- flight can be performed manually (joystick) and/or by means of an autopilot which helps in performing the waypoint flight;
- the pilot can both use the joystick and modify the waypoints by using the flight planner;
- at each time step, the energy consumption (fuel and batteries) is estimated and the endurance prediction updated;
- instantaneous data (flight parameters, power flows, etc.) and endurance prediction can be displayed via the HMI.

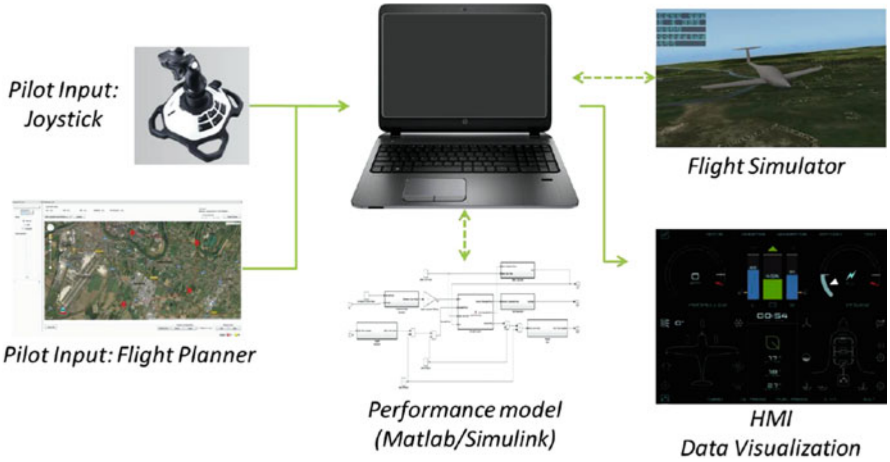

Pilot Input: Joystick
Pilot Input: Flight Planner
Flight Simulator
Performance model (Matlab/Simulink)
HMI Data Visualization

Fig. 6 Conceptual layout of the hybrid plane simulator

Data are exchanged in real time between the different modules through a set of plugins programmed in C++. Such data can be divided into the following datasets:

- *Aircraft dataset*: it contains information about the initial conditions of the aircraft, such as the amount of embarked fuel, the initial SoC of batteries, and the ICE generator characteristics;
- *Mission dataset*: it contains a description of the mission profile by means of waypoints, which can be modified during the mission;
- *Energy dataset*: it contains the instantaneous values of required flight power and energy consumption of both fuel and battery, thus it is updated continuously.

During the simulation, the main flight data are also recorded directly in a log file, in such a way that part of the calculations performed can be verified in a post-processing phase, and the user has a complete overview on mission parameters.

2 The *Flight Simulator*

The commercial flight simulator *X-Plane* [3] has been implemented in `HyPSim` and used as aerodynamic solver. It provides reliable data on aircraft aerodynamics by means of a panel method that computes the aerodynamic forces at each instant.

X-Plane has been chosen since it is easy to interface with other codes and contains a parametric tool for the creation of new aircraft, called *Plane-Maker*, by means of which the user can create all the aircraft components, such as wing, fuselage, blades, control surfaces, and landing gears. The resulting model, shown in Fig. 7, includes the airfoil characteristics which have been added in order to increase the accuracy of the panel method.

Fig. 7 X-Plane model of the reference aircraft

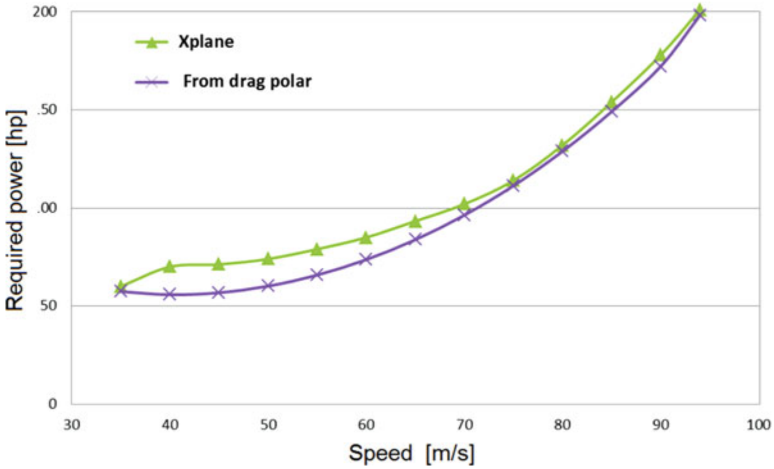


Fig. 8 Required power vs speed

Results provided by the aerodynamic model implemented in *X-Plane* have been validated through experimental data provided by the manufacturer. Figure 8 shows such comparison, which indicate a good accuracy for the model, although some differences are observed at low speed conditions.

During the simulation, aerodynamic forces, speed, etc., can be extracted from the *Flight Simulator* and mission parameters can be updated at the same time (Fig. 9).

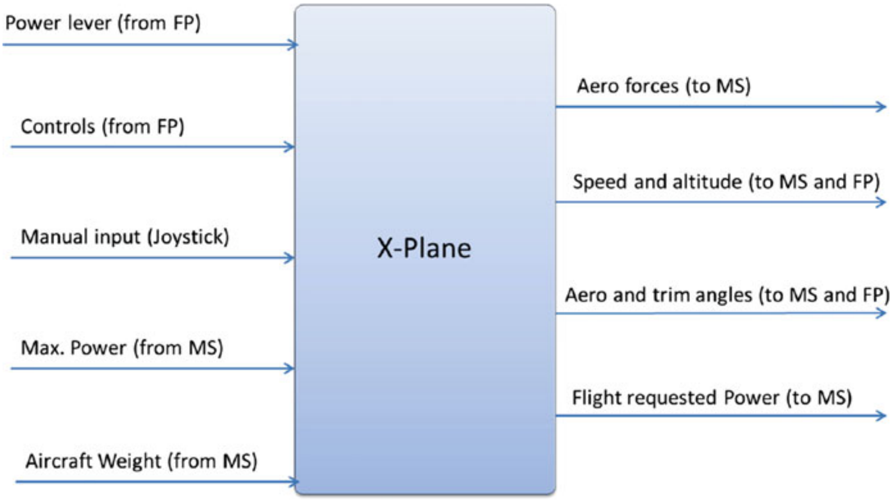

Power lever (from FP)
Controls (from FP)
Manual input (Joystick)
Max. Power (from MS)
Aircraft Weight (from MS)
X-Plane
Aero forces (to MS)
Speed and altitude (to MS and FP)
Aero and trim angles (to MS and FP)
Flight requested Power (to MS)

Fig. 9 Data exchange in the *Flight Simulator*. *XP* Flight simulator, *MS* Performance module, *FP* Flight planner

3 The *Performance Module*

Aerodynamic forces and other flight parameters are provided to the *Performance Module* by the *Flight Simulator* and used to compute the available energy in both fuel and batteries, in order to predict the remaining flight endurance.

The *Performance Module* has been developed by means of the Simulink software, implementing two independent blocks, the first one dedicated to the hybrid powertrain modelling and the second one for the endurance prediction.

The *Performance Module* is connected to the other modules as shown in Fig. 10.

3.1 Hybrid Powertrain Model

The hybrid powertrain, which includes also the ICE and the propeller, has been modelled using the scheme shown in Fig. 11. Since the maximum power provided by the brushless motor decreases with the batteries SoC, it is assumed that the rotating speed of the propeller is constant during the flight whereas the maximum torque can change.

In the first block, called ‘IN’ in Fig. 11, the input data coming from both the *Flight Planner* and the *Flight Simulator* are initialized and used as variables for the calculation. The main blocks are briefly described here after:

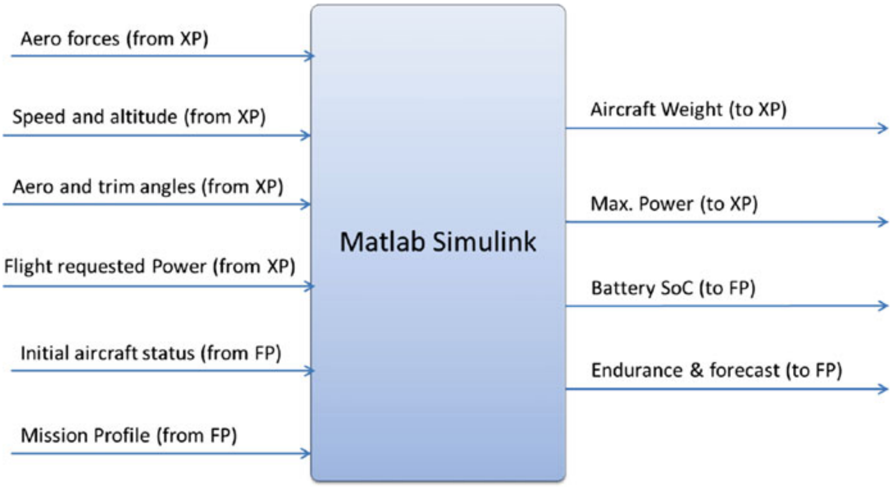

Aero forces (from XP)
Speed and altitude (from XP)
Aero and trim angles (from XP)
Flight requested Power (from XP)
Initial aircraft status (from FP)
Mission Profile (from FP)
Matlab Simulink
Aircraft Weight (to XP)
Max. Power (to XP)
Battery SoC (to FP)
Endurance & forecast (to FP)

Fig. 10 Data exchange in the *Performance Module*. *XP* Flight simulator, *MS* Performance module, *FP* Flight planner

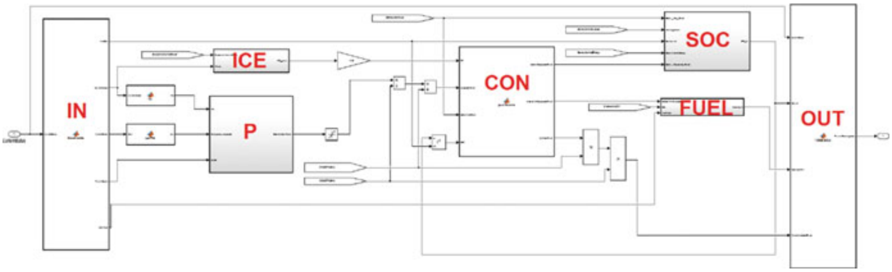


Fig. 11 Simulink scheme of the hybrid powertrain

- *ICE*: in this block the efficiency of the ICE is computed depending on the flight altitude. The relation is based on the interpolation of experimental data provided by the ICE manufacturer [4].
- *P*: here the propeller efficiency is defined and, by applying the actuator disk theory and taking the flight conditions into account, the power demand to the electric motor is calculated.
- *CON*: this block simulates the control system which manages the available power through proper control laws.
- *SOC*: once the instantaneous power request to batteries is known, the SoC of the batteries is calculated in this block thorough an energy balance, in which internal losses are taken into account.
- *FUEL*: given the specific fuel consumption and the power required to the ICE, the fuel consumption is calculated.

At the present stage of development, all the electric components (generator, motor, and inverters) are modelled by means of constant gains, representing the efficiencies, which can be modified during the initialization. In a similar way, the reference values for the batteries, such as maximum energy and initial SoC, are also set at the beginning of the mission.

The control block needs the instantaneous batteries SoC as input, that is computed in the battery block: thus, a closed loop is needed and an anticipator block is applied to the value of the SoC in order to synchronize the calculation.

The Simulink scheme is triggered with a value of 0.1 s in order to properly update the aircraft status.

3.2 *The Predictor*

The predictor has been conceived in order to estimate at each time step the remaining flight endurance. The prediction is performed taking the amount of available energy (fuel and batteries) and the reference mission defined during the initialization into account.

Mission parameters and aircraft status are used as input of the predictor, which is made of two blocks, as Fig. 12 shows

- *PROFILE*: here the mission profile is divided into flight segments: climb, cruise, descent, landing, diversion, and loiter. Each segment is defined entirely by the parameters listed in Table 1 and extrapolated from the *Flight Planner*.

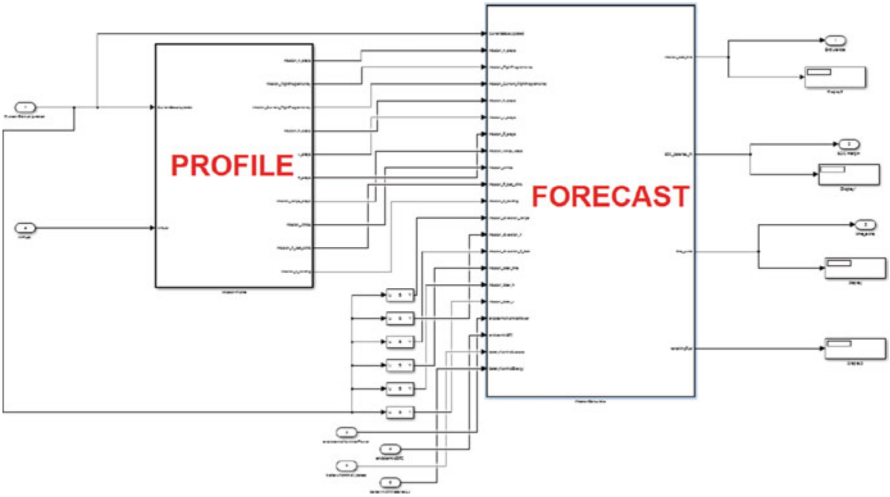


Fig. 12 Simulink scheme of the performance predictor

- *FORECAST*: in this block, the fuel consumption and the battery discharge are calculated for the mission defined in the previous steps, assuming the flight programs reported in Table 1.

Although the energy required to perform the emergency segments (diversion and loiter) is considered in the mission energy balance, the predicted endurance does not include the time needed to fly over such segments, hence the endurance prediction is always conservative.

In the *FORECAST* block, the amount of fuel required to complete the mission ($W_{f_{req}}$) is calculated and compared with the fuel available on the aircraft at the given time step ($W_f(t)$). Then, two conditions are possible:

(a) $W_f(t) > W_{f_{req}}$: the reference mission can be accomplished with some safety margin, which is indicated to the pilot as an additional flight endurance (t_{extra});
(b) $W_f(t) < W_{f_{req}}$: the reference mission cannot be accomplished and a negative extra flight time is provided as output to the pilot, together with a warning message.

In addition to these results, the calculation block returns also the final SoC that the batteries are expected to have at the end of the mission.

The prediction block is updated continuously during the mission in order to take possible external interferences during flight (e.g. wind or manual input) or possible modifications to the mission during the simulation into account. In this case the trigger frequency is lower because of the low computational speed of this block; since this calculation does not interfere with the other ones, this different frequency is considered acceptable.

4 The *Flight Planner*

The flight planner is an in-house developed software which is used for several purposes in `HyPSim`:

- to act as an autopilot, allowing to perform the given mission profile accurately (the pilot can always change the aircraft trajectory manually through the joystick);
- to allow the data exchange between the *Flight Simulator* and the *Performance Module*;
- to initialize the simulation, defining the initial status of the aircraft and the reference mission;
- to extract and visualize the results.

The main input/output data managed in the *Flight Planner* are reported in Fig. 13.

The *Flight Planner* interface consists of a plugin manager which allows to launch different software modules in a customizable layout. The main modules are

- a link module, which allows the communication between the different software of `HyPSim`;

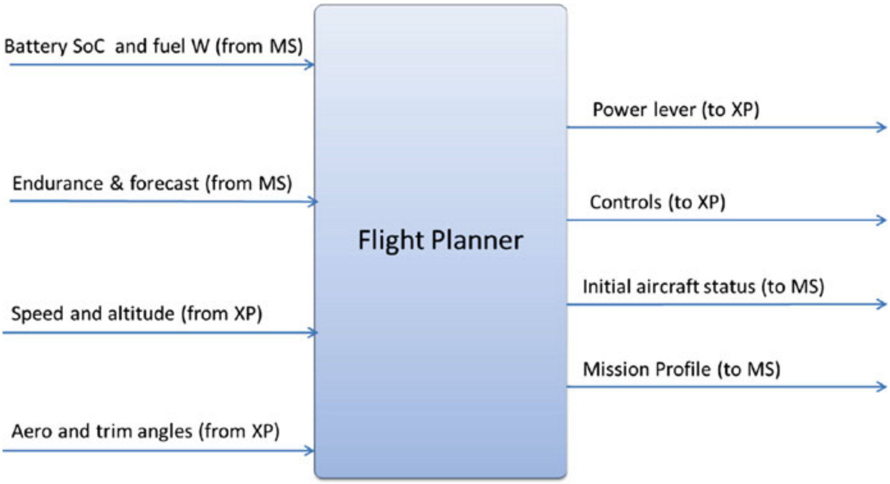

Battery SoC and fuel W (from MS)
Endurance & forecast (from MS)
Speed and altitude (from XP)
Aero and trim angles (from XP)
Flight Planner
Power lever (to XP)
Controls (to XP)
Initial aircraft status (to MS)
Mission Profile (to MS)

Fig. 13 Data exchange in the *Flight Planner*

- the map plugin, which is used to display the position, the direction, and the trajectory of the aircraft on a map as well as to define the mission profile (Fig. 14);
- the *Aircraft Management Module*, shown in Fig. 15, which allows to manage the aircraft during the flight simulation.

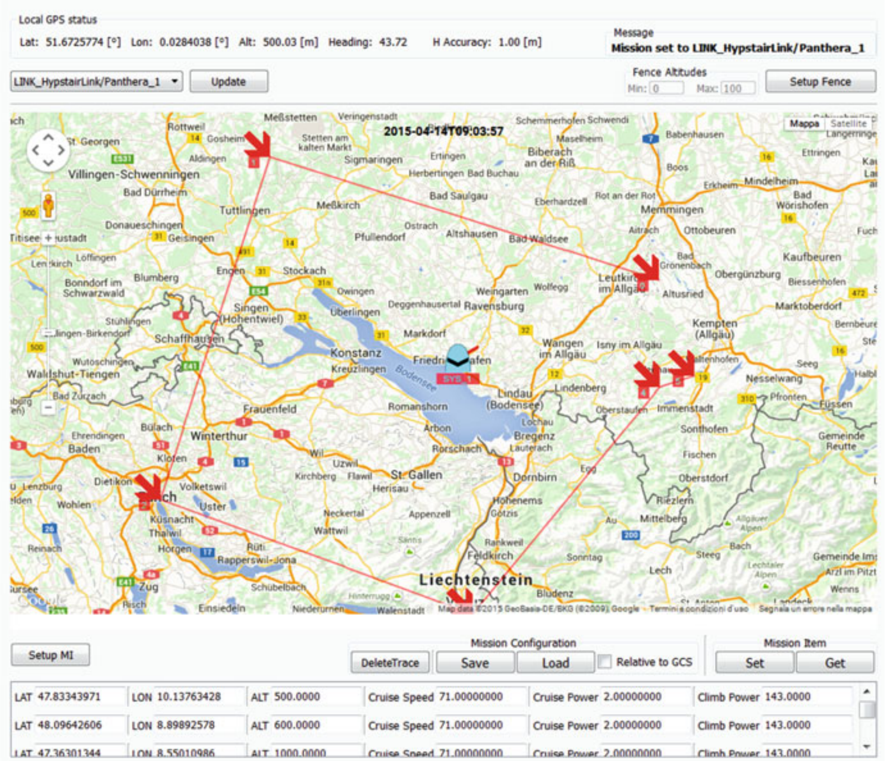
The mission profile can be defined by providing the waypoint list shown in the bottom part of Fig. 14. The waypoints are defined through the following values:

- latitude (*LAT*) and longitude (*LON*), whose values can be written in the related fields or provided by clicking on the map;
- altitude [m] (*ALT*);
- cruise speed [m/s] or cruise power [kW], which is neglected if cruise speed is assigned;
- climb power [kW], which is used when the altitude of the following waypoint is higher than the previous one.

The mission can be modified during the flight simulation by moving the waypoints on the map or modifying the parameters in the list; after any modification, the ‘Set’ button has to be clicked to make them active.

Some comments on both the initial and the last parts of the mission are remarked in the following points:

- the take-off point is not included in the waypoint list; the simulator recognizes whether the aircraft is on the ground and an automatic take-off procedure is performed in order to reach the first waypoint.
- two waypoints are required in the end of the list in order to perform an automatic landing: the first one is used to define the landing point and the second one provides the runway direction;



Local GPS status

Lat: 51.6725774 [°] Lon: 0.0284038 [°] Alt: 500.03 [m] Heading: 43.72 H Accuracy: 1.00 [m]

Message

Mission set to LINK_HypstairLink/Panthera_1

LINK_HypstairLink/Panthera_1

Fence Altitudes

Min: 0 Max: 100

2015-04-14T09:03:57

Mission Configuration

☐ Relative to GCS

LAT	47.83343971	LON	10.13763428	ALT	500.0000	Cruise Speed	71.00000000	Cruise Power	2.00000000	Climb Power	143.0000
LAT	48.09642606	LON	8.89892578	ALT	600.0000	Cruise Speed	71.00000000	Cruise Power	2.00000000	Climb Power	143.0000
LAT	47.36301344	LON	8.55010986	ALT	1000.0000	Cruise Speed	71.00000000	Cruise Power	2.00000000	Climb Power	143.0000

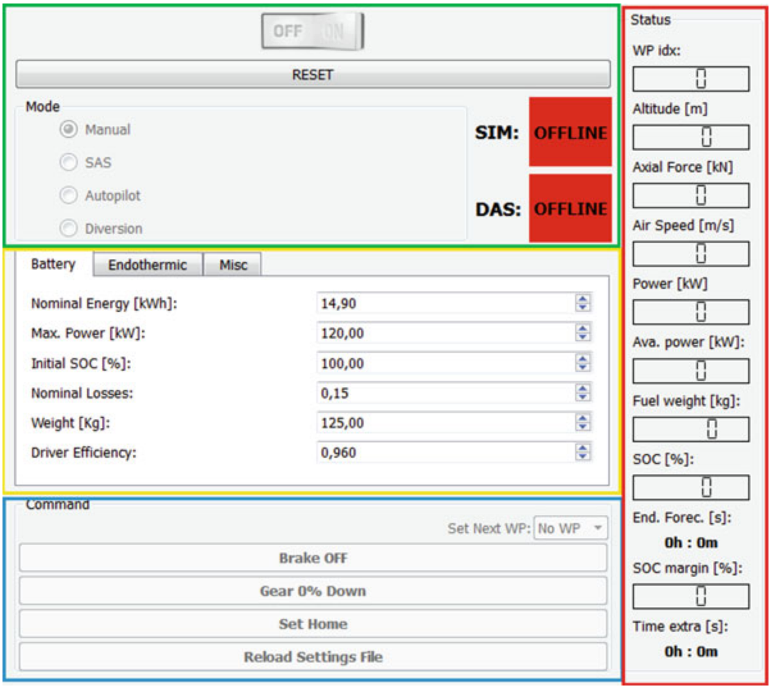
Fig. 14 Mission profile definition in the map plugin

- diversion and loiter segments are defined through the *Aircraft Management Module*, hence waypoints are not required.

4.1 The Aircraft Management Module

The *Aircraft Management Module*, shown in Fig. 15 is composed of the following parts:

- the *Simulation Control* section (green box);
- the *Power Control* section (yellow box);
- the *Commands* section (blue box);
- the *Status* section (red box);



Mode

● Manual
○ SAS
○ Autopilot
○ Diversion

SIM: OFFLINE

DAS: OFFLINE

Nominal Energy [kWh]: 14,90
Max. Power [kW]: 120,00
Initial SOC [%]: 100,00
Nominal Losses: 0,15
Weight [Kg]: 125,00
Driver Efficiency: 0,960

Set Next WP: No WP

WP idx: 0
Altitude [m] 0
Axial Force [kN] 0
Air Speed [m/s] 0
Power [kW] 0
Ava. power [kW]: 0
Fuel weight [kg]: 0
SOC [%]: 0
End. Forec. [s]: 0h : 0m
SOC margin [%]: 0
Time extra [s]: 0h : 0m

Fig. 15 The *Aircraft Management Module* window

4.1.1 Simulation Control Section

Once the connections of the *Flight Planner* with the *Flight Simulator* and the *Performance Module* are active (green color in *SIM* and *DAS* boxes, respectively), the *On/Off* button can be turned on in order to control the simulation. The following options can be activated:

- *AUTO*: the flight is controlled directly by the *Flight Planner* according to the mission profile defined in the map plugin;
- *MANUAL*: stick and throttle are manually controlled by means of a joystick;
- *DIVERSION*: the flight is automatically controlled and in addition the aircraft follows the path defined for the diversion and loiter.

When the autopilot mode is on, the *Flight Planner* manages the flight simulation directly. Most of the climb and cruise parameters can be inserted directly in the map plugin or in the aircraft management module, whereas the other flight segments (take-off, descent, diversion, and loiter) are managed by means of a *setting file* that must be loaded before starting the simulation control.

4.1.2 Power Control Section

The power sources can be defined through the interactive yellow box in the centre of the *Aircraft Management Module*. The data that must be provided are divided into three panels:

- *Battery panel*: required inputs concern the batteries characteristics (weight, maximum storable energy, initial SoC, etc.) and the internal losses of the electric propulsion components (propeller is not included);
- *Endothermic panel*: required inputs concern the nominal power, the efficiency and the Specific Fuel Consumption of the ICE, as well as data on the initial embarked fuel;
- *Misc panel*: required inputs are options about the power management during take-off (e.g. power provided by batteries, etc.).

4.1.3 Commands Section

When the autopilot is active, some commands can be managed using the *Commands* section in the bottom part of the interface.

Defined the reference mission as a waypoint list, the pilot can change the flight plan by selecting which waypoint has to be reached first (*Set next WP* button) or flying manually. In this latter case, by enabling again the autopilot mode, the *Flight Planner* automatically recognizes the nearest waypoint as the first one to be reached and the mission is then performed from that point ahead following the list.

Finally, the *Reload Setting File* button allows to modify the aircraft flight parameters and the control laws of moveable surfaces. Such file provides the following settings:

- take-off is performed at the maximum nominal power, with a given gain for rudder control in order to compensate the propeller torque effect;
- descent is performed with a given power throttle level;
- diversion and loiter parameters refer to the minimum power and maximum endurance conditions, respectively;
- landing is performed with given speed, flap deflection, and providing the runway altitude.
- manoeuvring limitations (e.g. maximum bank angle)

4.1.4 Status Section

The *Status* section is dedicated to the real time visualization of the main flight parameters; the first field (*WP idx*) refers the identification number of the waypoint which the aircraft is heading to, whereas the other fields provide instantaneous data on the hybrid aircraft performance.

Finally, data resulting from the prediction model are visualized: the predicted endurance (*End. Forecast*), the estimated SoC at the end of the mission (*SoC margin*), and the difference, positive or negative, between the expected flight time and the time required to complete the mission (*Time extra*).

5 The Human–Machine Interface

Since part of the HYPSTAIR project has been dedicated to the development of a dedicated HMI [5], the simplified HMI panel shown in Fig. 16 has been implemented in `HyPSim` in order to display the following information:

1. battery SoC;
2. discharging/charging state: the triangle is green and rotated upward during charge or yellow and rotated downward in discharge;
3. fuel amount in left and right tanks;
4. remaining flight time in hours and minutes, as the sum of mission time and extra time (if this latter is negative, the time is visualized in orange colour in order to create a warning for the pilot);

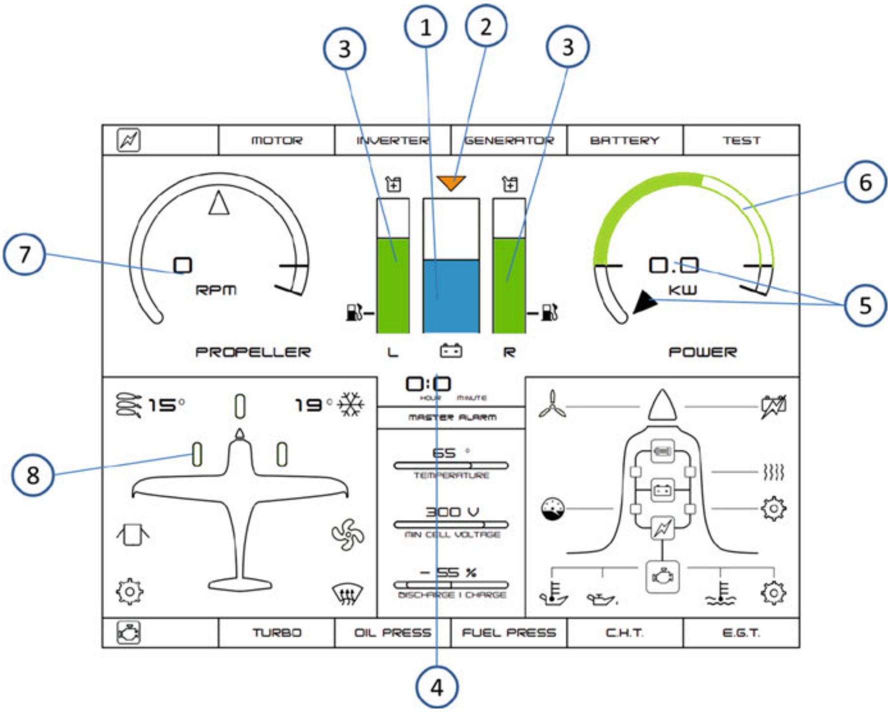


Fig. 16 The HMI panel

5. power consumption: the instantaneous required power is displayed through both numbers and a pointer which moves along the green arch;
6. available power, represented through the empty green arch, whose length changes if batteries are fully discharged or the ICE is switched off;
7. propeller revolutions per minute (RPM);
8. landing gear position: green if extracted, empty otherwise.

6 Simulator Testing

Several simulations have been conducted in order to assess the accuracy of the performance models implemented in `HyPSim`: the first test campaign has been focused on required power evaluation, whereas the second one has been carried out in order to study a critical condition in which batteries are fully discharged.

6.1 Required Power Evaluation

The simulator has been first tested by assigning the mission profile shown in Fig. 17, in which two level flight phases, at 200 and 1000 m, have been performed varying the speed from 55 to 85 m/s with a step input given to the throttle.

Figure 17 shows the required power calculated by the simulator, whose positive and negative peaks are due to the accelerations and decelerations of the aircraft. In fact, according to Eq. (1), the required power can be decomposed in three contributions: the first one associated with aerodynamic drag D (speed and altitude are considered constant), the second due to altitude variation on a constant slope (γ) trajectory, and the third one due to speed variations (dV/dt).

$$P_{req} = V \cdot D + V \cdot W \cdot \sin\gamma + V \cdot \frac{W}{g}\frac{dV}{dt} \tag{1}$$

In this case, positive peaks are due to the accelerations introduced at each step of the $V(t)$ input function, whereas negative peaks indicate that the aircraft is decelerated, hence P_{req} is set to 0.

The required power is multiplied by the efficiencies of all the powertrain components in order to calculate the power demanded to both ICE generator and batteries. Figure 18 shows the comparison between required power for flight and the power demand to energy sources as provided by the *Flight Simulator*.

The dashed grey line in Fig. 18 is the maximum power provided by ICE generator, hence this chart allows to define the batteries charge and discharge phases. Therefore the SoC chart has been obtained, observing that a big discharge (about 30 %) is needed to perform the climb from 200 to 1000 m, whereas during level flight charge and discharge phases alternate depending on speed variations (after take-off SoC has been limited to 90 %).

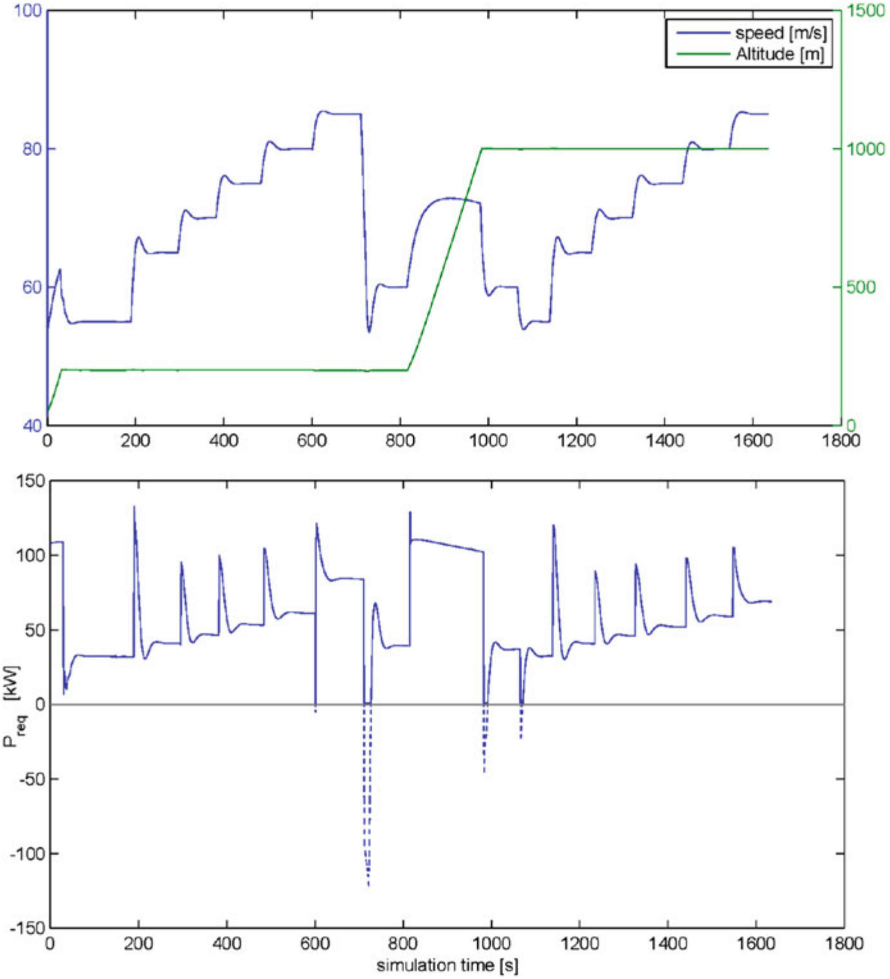


Fig. 17 Flight speed and altitude of the input mission (*top*) and required power output (*bottom*)

6.2 Fully Discharged Batteries

During this simulation campaign, the plane has been set in level flight conditions with a constant speed of 80 m/s, in such a way the batteries are continuously discharged until the minimum SoC threshold, set to 4 %, is reached. In such condition the only available energy source is the ICE generator, which is assumed to provide a constant power of 80 kW.

With the aim of evaluating the flight performance when the available power is limited, a simulation has been performed using the *Flight Planner* in automatic mode in order to force the aircraft to fly at 80 m/s although the available power is not sufficient.

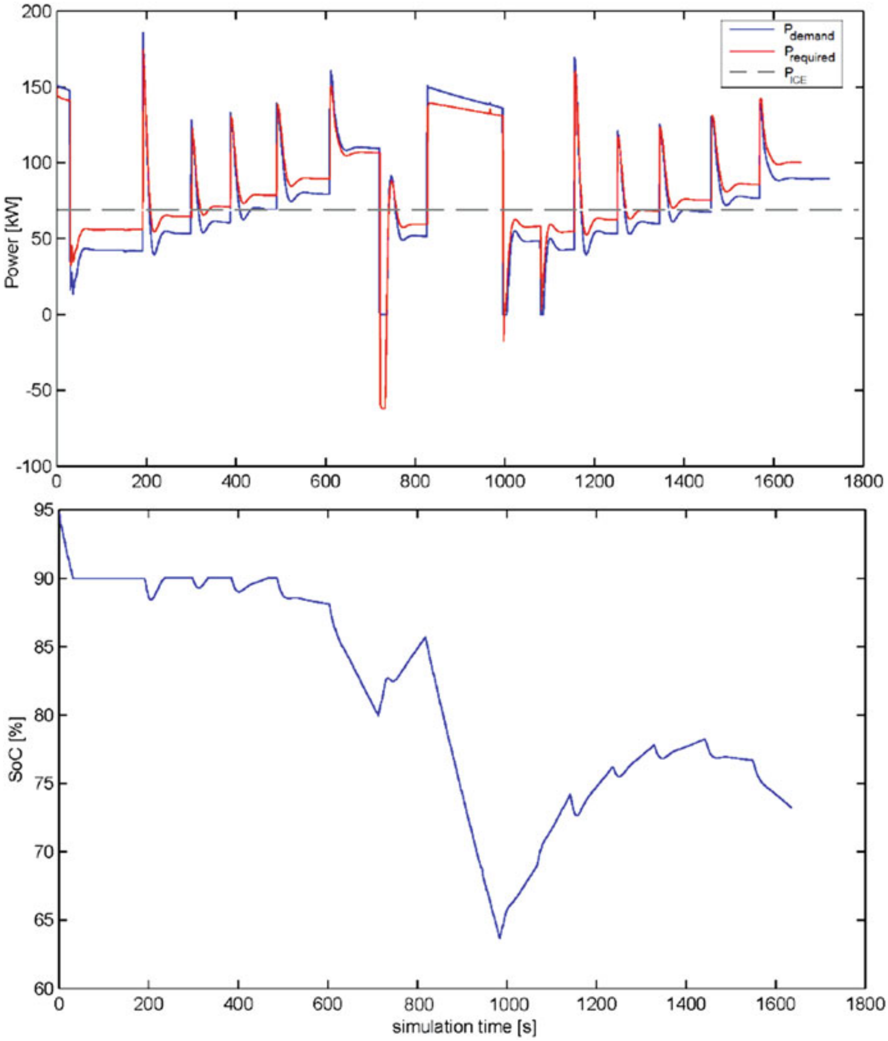


Fig. 18 Required power for flight and power demand to energy sources (*top*) and batteries SoC (*bottom*)

As Fig. 19 shows, when SoC reaches its lower limit the available power is instantaneously reduced to 80 kW and the aircraft speed decreases until the required power becomes lower than the available one. When this happens, the batteries begin charging and as soon as the SoC becomes higher than 4 % the available power is restored to the maximum value, which brings the autopilot to increase aircraft speed up to 80 m/s. Hence, batteries are discharged again and such cycle is repeated creating an oscillating behaviour which can have negative consequences on both batteries health and flight dynamics.

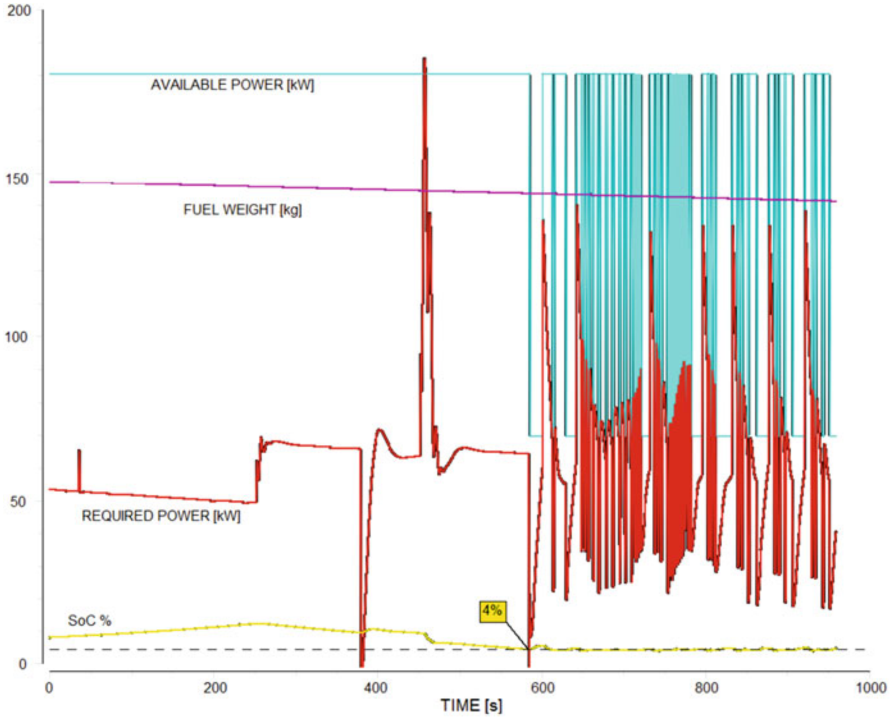


Fig. 19 Simulation of fully discharged batteries

It has been observed that such oscillations can be avoided by adding a second SoC threshold of 7 %, below which the battery charge is not activated. The introduction of this additional threshold changes the power profile as illustrated in Fig. 20, in which oscillations can be still observed but the frequency is much lower and the effects on flight dynamics are reduced.

7 Conclusions

The activities presented in this paper have been part of the European project called *HYPSTAIR*, concerning the development and validation of hybrid propulsion system components and sub-systems for electrical aircraft.

In particular, the development of a hybrid plane simulator, called `HyPSim`, has been described focusing on the software architecture and the functionality of such a simulation tool.

The main modules which compose the simulator are a commercial *Flight Simulator* (X-Plane), with which the aircraft geometry and aerodynamics have been modelled, a *Performance Module* developed in Simulink and used to simulate

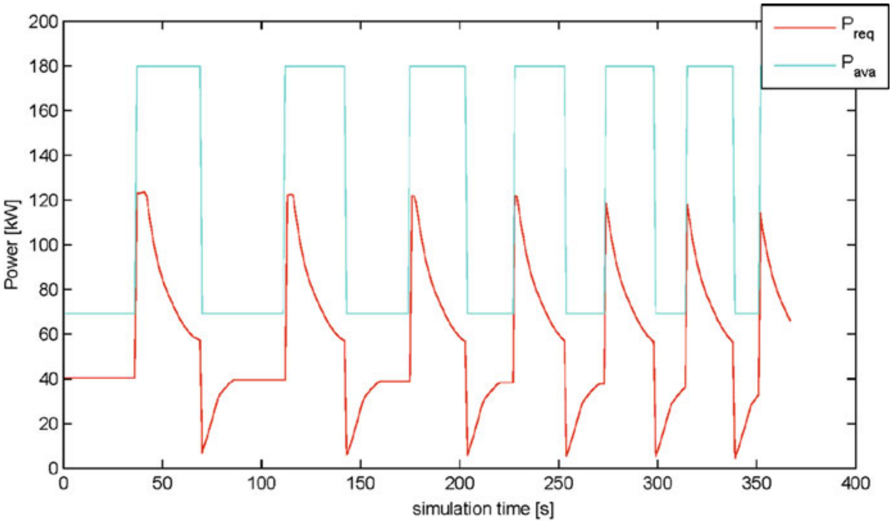


Fig. 20 Power function after the introduction of a second threshold on SoC

the hybrid powertrain, calculate the energy consumption, and predict the flight endurance, and a in-house developed *Flight Planner* which allows to define the mission profile, select the flight mode (manual, autopilot, etc.), and allow the data exchange between all the modules. The simulator, in addition, can provide the main output using the HMI developed in the *HYPSTAIR* project.

The accuracy of the simulator in evaluating the energy consumption has been verified by comparing the required power for flight with experimental results provided by the aircraft manufacturer. In addition, specific mission profiles have been given as input and positive results on the reliability of the power demand evaluation have been achieved.

Finally, it has been observed that for some peculiar conditions, such as the case of fully discharged batteries, additional control logics must be implemented in order to avoid divergence phenomena.

As a general conclusion, `HyPSim` is a simulation tool able to achieve the several purposes for which it has been conceived, allowing to simulate any kind of mission profile taking also the human-in-the-loop factor into account. Moreover, the simulator is a practical tool for dissemination purposes.

Further development can be focused on the following aspects:

- implementation of additional control logics for off-design conditions;
- deeper and more complete implementation of the HMI module in the simulator;
- integration with haptic input devices developed within the *HYPSTAIR* project [6];
- development of more detailed models for the powertrain simulation.

Acknowledgements The research has been carried out within project ‘Development and validation of hybrid propulsion system components and sub-systems for electrical aircraft’ (HYPSTAIR), which was co-funded by the European Union’s Seventh Framework Programme for research, technological development, and demonstration under grant agreement no. 605305.

References

1. Oliviero, F., Cipolla, V.: HYPSTAIR Project Deliverable D2.1: preliminary design of a serial hybrid aircraft. Internal Report of the Hypstair Project, University of Pisa (2014)
2. Oliviero, F., Cipolla, V.: HYPSTAIR Project Deliverable D2.2: preliminary design of a serial hybrid aircraft. Internal Report, University of Pisa (2015)
3. Laminar Research ind.: Plane maker for X-Plane 10 App Manual (2013)
4. Rotax Aircraft Engines: Operator’s Manual for Rotax 912 (2014)
5. Ferracci, M., Barlocchetti, S.: HYPSTAIR Project Deliverable D3.1: HMI design. Internal Report, MB Vision (2014)
6. Hace, A., Golob, M.: HYPSTAIR Project Deliverable D3.3: Haptic Interface. Internal Report, University of Maribor (2015)

Evolutionary and Heuristic Methods Applied to Problems in Optimal Control

Bruce A. Conway

Abstract About two decades ago years researchers began to apply a new approach, using evolutionary algorithms or metaheuristics, to solve continuous optimal control problems. The evolutionary algorithms use the principle of "survival of the fittest" applied to a population of individuals representing candidate solutions for the optimal trajectories. Metaheuristics optimize by iteratively acting to improve candidate solutions, often using stochastic methods. Because of certain compromises that are usually necessary when transcribing the problem for solution by these methods it has been thought that they were not capable of yielding accurate solutions. However that is a misconception as is demonstrated by examples in this work.

Keywords Optimal control • Evolutionary algorithms • Metaheuristic algorithms

1 Introduction

The subject of optimization of a continuous dynamical system has a long and interesting history. The first, and perhaps archetypal example is the *Brachistochrone* problem posed by Galileo and later by Bernoulli (and solved by Newton in 1696) [1]. The problem can be simply stated as the determination of a trajectory that satisfies specified initial and terminal conditions (and possibly constraints on the path), while satisfying the system governing equations, and minimizing some quantity of importance. We use the term "trajectory" here as representing a path or time history of the system state variables. Of course in the field of spacecraft and aircraft trajectory optimization the trajectories are literal, but the theory can be applied to any system whose governing equations can be reduced to a system of (piecewise) continuous first-order ODEs.

There are two common objectives to minimize in most of the dynamical systems with which we are acquainted. Either some function related to the amount of control

B.A. Conway (✉)
Department of Aerospace Engineering, 306 Talbot Laboratory, University of Illinois, Urbana, IL 61801, USA
e-mail: bconway@uiuc.edu

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_5

authority that is to be applied to the system or the amount of time allowed to satisfy the desired terminal conditions, or possibly a combination of the two, is to be minimized. The latter case would be an example of multi-objective optimization. Such problems are common in the optimization of continuous dynamical systems where there is often a cost both to using control and to using time. In such problems there is no unique solution that optimizes both objectives. *Pareto optimal* or *non-dominated* solutions may exist in which one objective cannot be improved without worsening the other objective(s). There is then a multiplicity of possible solutions. An obvious example is that of finding optimal interplanetary spacecraft trajectories. If some practical upper bound for the final time is not provided the optimizer will trade time for use of control (propellant). A Pareto front of solutions exists and represents a set of choices that will minimize the propellant required for any feasible time of flight. This paper will not consider multi-objective optimization in its analysis or examples, though some of the numerical methods presented would be useful tools for determining the Pareto optimal solutions. Only a single objective function will be employed, however that function may be a linear combination of objectives, e.g., a weighted sum of the control effort employed and the time required.

Except in very special (integrable) cases, which reduce naturally to parameter optimization, the problem is a continuous optimization problem of an especially complicated kind. The complications include the following: (i) The dynamical system may be nonlinear. (ii) The optimal trajectory may include discontinuities in the state variables. (iii) The terminal conditions, initial or final or both, may not be known explicitly. (iv) There may be time-dependent perturbations, i.e., the right-hand side of the system governing equations may change qualitatively with time. (v) There may be bounds on the state and control variables that are effective on, i.e., are a feature of, the optimal trajectory.

Another complication is that the basic structure of the optimal trajectory may not be a priori specified but is instead itself subject to optimization. An obvious example is the travelling salesman problem where the order of visitation of the cities must be solved for simultaneously with the optimal paths between the cities [2]. Or, for an interplanetary transfer, the optimal number of impulses or the optimal number of planetary flybys (or even the planets to use for the flybys) may not be known [3, 4].

Necessary conditions for optimality for all types of trajectory optimization problems may be derived using the calculus of variations (CoV) [1]. Unfortunately, analytical solution of the resulting system of equations and boundary conditions, which is possible for a problem such as the *Brachistochrone* (which does not have any of the pathologies (ii)–(iv) described in a previous paragraph), is seldom feasible for any significant, “real-world” problems. The vast majority of researchers and analysts today use numerical optimization for such problems. Numerical optimization methods for continuous optimal control problems are generally divided into two types. *Indirect* solutions are those using the analytical necessary conditions from the calculus of variations. This requires the addition of the costate variables (or “adjoint” variables or Lagrange multipliers) of the problem, equal in number to the state variables, and their governing equations. This instantly doubles the size of the dynamical system, which alone of course makes it more difficult to solve.

Direct solutions, of which there are many types, transcribe the continuous optimal control problem into a parameter optimization problem [5–7]. Satisfaction of the system equations is accomplished by integrating them stepwise using either implicit or explicit (for example, Runge–Kutta) rules; in either case the effect is to generate nonlinear constraint equations that must be satisfied by the parameters, that are the discrete representations of the state and control time histories. The problem is thus converted into a nonlinear programming (NLP) problem. There is a comprehensive survey paper by Betts [8] that describes direct and indirect optimization, the relation between these two approaches, and the development of these two approaches.

In just the decade since the publication of Betts' survey paper [8] there has been considerable advancement of direct numerical solutions for optimal control problems [9]. There has also been even more development and improvement, in relative terms, of a qualitatively different approach to solving such problems, one using evolutionary algorithms and metaheuristics [10, 11]. While these methods have existed since the 1960s, they were for a long time applied primarily to parameter optimization problems. Their application to optimization of continuous dynamic systems is comparatively recent. A survey paper by Conway [9] discusses methods for direct optimization and also how these methods may be aided (or even supplanted) by the evolutionary or metaheuristic algorithms. The best known of the evolutionary algorithms are the genetic algorithms (GA) [10]. There are many evolutionary algorithms but their common characteristic is that they apply the principle of "survival of the fittest" in the determination of the optimal solution.

Metaheuristic methods are qualitatively similar to evolutionary algorithms but "evolve" differently. Evolutionary algorithms discard poorly performing individuals (i.e., solutions) and proceed with the survivors and new individuals created from the survivors. In metaheuristic methods, the best known of which may be particle swarm optimization (PSO) [12], the entire population takes steps with the intention of improving the fitness of all the individuals. (Of course since the process is partly stochastic not every individual is improved at every iteration.)

The evolutionary algorithms and metaheuristic methods have two principal advantages over other methods; they are comparatively simple and thus easy to program and use and they are generally more likely, in comparison to conventional optimizers, to locate global minima. Perhaps their most compelling feature is that they require no initial guess of the solution be provided; their initial populations are initialized randomly, within bounds for the solution provided by the user. This is in great contrast to the direct methods that transcribe the continuous problem into a discrete problem, actually an NLP problem. The NLP problem solvers such as MATLAB's *fmincon* or SNOPT [13] require an initial guess of the solution vector, which is typically a time history of all of the discrete states and controls. (Even the indirect solution methods require an initial guess, usually of the initial values of the system costates or Lagrange multipliers, whether the solution is being obtained by shooting or by conversion to an NLP problem.) This can be problematic for a number of reasons; an approximate solution, especially one that satisfies the system governing equations, may be difficult to generate. Without a "reasonable" initial guess, the NLP problem solver may fail to converge (this is lack of "robustness").

However, even when convergence to a solution occurs, the initial guess may in some cases prejudice the solver, so that it finds a solution "near" to the guess, which may only be locally optimal. The evolutionary and metaheuristic methods are not similarly prejudiced by an initial guess and with a sufficiently large search space, a sufficiently large population, and a sufficient number of iterations, they can search the solution space comprehensively and discard local minima in favor of the global minimum. An obvious question is how does one determine what constitutes "sufficient" for each of these three things? This will be described in more detail later, but a brief answer is that there is no formula; experience helps, sometimes trial-and-error is required.

2 The Optimal Control Problem

Before describing the numerical, and necessarily approximate, solutions to the optimal control problem, it will be useful to consider the analytical necessary conditions. The classical, calculus of variations approach, resulting in the Euler–Lagrange equations, i.e., the necessary conditions for a local extremum, will be illustrated for a generic (but typical) problem whose dynamics are governed by a second-order ODE, such as Newton's second law, and in which the controls appear linearly, i.e.,

$$\ddot{r} = \mathrm{g}(r) + u \tag{1}$$

where u represents a vector of continuous control variables. After the second-order equations are converted into a set of first-order ODEs the system governing equations are

$$\dot{x} = f = \begin{bmatrix} \dot{r} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v \\ \mathrm{g}(r) + u \end{bmatrix} \tag{2}$$

where x represents the system state vector.

The objective (to minimize) may be written in the Bolza form as:

$$J = \phi\,[x(T),\ T] + \int_0^T L\,[x,\ u,\ t]\,dt \tag{3}$$

where ϕ is a terminal cost function and T is the final time while the integral expresses a cost incurred during the entire trajectory. Terminal boundary functions are given as the vector:

$$\Psi\,[x(T),\ T] = 0. \tag{4}$$

The system equations are "adjoined" to the L function through continuous Lagrange multipliers (or adjoint variables or costates) to form the system Hamiltonian H,

$$H = L + \lambda^T f = L + \lambda_r{}^T \mathrm{v} + \lambda_v^T \left[g(r) + u\right]. \tag{5}$$

The necessary conditions for an optimal trajectory then become [1]:

$$\begin{aligned} \dot{\lambda} &= -\left(\tfrac{\partial H}{\partial x}\right)^T \textit{with boundary condition} \\ \lambda(T) &= \left[\left(\tfrac{\partial \phi}{\partial x}\right) + \nu^T \left(\tfrac{\partial \Psi}{\partial x}\right)\right]^T_{t=T} \end{aligned} \tag{6}$$

or for this example:

$$\begin{aligned} &\dot{\lambda}_{\mathrm{r}}^{\mathrm{T}} = -\tfrac{\partial H}{\partial r} = -\lambda_{\mathrm{v}}^{\mathrm{T}}\, G(r), \text{ where } G(r) = \tfrac{\partial g}{\partial r}, \text{ a symmetric 3x3 matrix} \\ &\dot{\lambda}_{\mathrm{v}}^{\mathrm{T}} = -\tfrac{\partial H}{\partial \mathrm{v}} = -\lambda_{\mathrm{r}}^{\mathrm{T}}, \text{ with} \\ &\lambda_{\mathrm{r}}^{\mathrm{T}}\left(t_f\right) = \tfrac{\partial \phi}{\partial \mathrm{r}} + \nu^T \left(\tfrac{\partial \psi}{\partial \mathrm{r}}\right)_{t_f} \\ &\lambda_{\mathrm{v}}^{\mathrm{T}}\left(t_f\right) = \tfrac{\partial \phi}{\partial \mathrm{v}} + \nu^T \left(\tfrac{\partial \psi}{\partial \mathrm{v}}\right)_{t_f}. \end{aligned} \tag{7}$$

along with (2), (4), (6) and a method for choosing the optimal control that will be described in the next paragraph. This system of equations constitutes a two-point-boundary-value problem (TPBVP); some boundary conditions on the states are specified at the initial time and some boundary conditions on the states and adjoints are specified at the terminal time. In addition, if the terminal time is unspecified (i.e., free to be optimized) as is often the case, an additional scalar equation (the transversality condition) obtains

$$\left[\frac{\partial \phi}{\partial t} + \nu^T \left(\frac{\partial \Psi}{\partial t}\right) + \left(\frac{\partial \phi}{\partial x} + \nu^T \left(\frac{\partial \Psi}{\partial x}\right)\right) f + L\right]_{t=T} = 0 \tag{8}$$

There are additional necessary conditions for optimality for common cases including (i) when the initial time (and thus possibly some of the initial states) is unspecified, (ii) cases with boundary functions on the initial states, (iii) cases where a state constraint function obtains at an intermediate time (such as a dynamic pressure constraint for a launch vehicle), and (iv) cases involving singular arcs. Discussion of these cases is more appropriately the job of a textbook on optimal control theory [1, 14].

For all but the most elementary optimal control problems the solution of this TPBVP is challenging and numerical solutions are required. Despite this, even for systems that are by no means elementary, several very useful observations regarding the optimal control may be made. The optimal control is chosen according to Pontryagin's minimum principle [1], that at any time on the optimal trajectory the control variables are chosen in order to minimize the Hamiltonian. (For a problem with unbounded control this results in a simple requirement that the Hamiltonian be stationary with respect to small changes in the control, i.e., $\partial H/\partial u = 0$.) Thus the first simple observation for the example above is that the control u should be chosen to be parallel to the opposite of the adjoint (to the velocity) vector, i.e., $-\lambda_v(t)$. (In space flight mechanics problems this (adjoint) vector is referred to as the *primer vector* [15, 16].) The minimum principle then also requires that the magnitude of the control be chosen to minimize the Hamiltonian. This generally yields a "switching function," i.e., the coefficient of the control u in the Hamiltonian. The switching function for the Hamiltonian (5) does not have a general form since the function L may involve the control in an arbitrary way. The optimal strategy is thus to choose the control u at its most negative bound if the switching function is positive and choose u at its positive bound if the switching function is negative. The adjoint vector $\lambda_v(t)$ would obviously appear in the switching function for this system and is governed by the system Eqs. (6) and (7) with the Hamiltonian (4).

These (first-order) necessary conditions identify stationary solutions. To determine if the stationary solution is optimal additional tests are required. A necessary condition for optimality is the second-order Jacobi no-conjugate-point condition [1]. There are a number of sufficient condition tests available depending on the form of the problem. Wood [17] derived new sufficient conditions for a weak local minimum of the Bolza problem. Jo and Prussing [18] derived a sufficient condition test based on the work of Wood but having advantages with regard to computational simplicity. A procedure for verifying minimality of a singular extremal (arc) was derived by Kelley et al. [19]. Minimizing singular subarcs are not "common" in trajectory optimization problems but can certainly arise even in unsophisticated problems; the most famous example probably being the optimal thrust program for a rocket launched from the Earth's surface posed and solved by Goddard [20].

The solution of the TPBVP resulting from (or constituting) the necessary conditions becomes quite difficult for sophisticated problems, particularly those with path constraints (typically on the state variables or on functions of the state variables) or constraints on total control authority available (e.g., total fuel available). Many methods have been developed to solve the TPBVP numerically [5–9]. The long-recognized difficulty of this "indirect" approach to determining the optimal trajectory is that the initial costate variables of the TPBVP are unknown and further that the nonlinearity of the problem means that the vector flow is very sensitive to some or all of these initial costate variables [1]. This may well have been one of the motivations for the (generally) more robust direct solutions for the optimal control which, as previously mentioned, transcribe the problem into an NLP problem and do not require the system costate variables.

Normally, when using evolutionary or metaheuristic methods to solve for the optimal program, the analytical necessary conditions of the problem are not explicitly considered. (There will be one example to follow where this is not true, where the solution by PSO does satisfy the system Euler–Lagrange Eqs. (2), (3), (4), (5), (6), and (7).) However it is nonetheless useful to have some knowledge of the types of solutions that can obtain, particularly bang-bang solutions that result from Pontryagin's principle and solutions having singular arcs, so that when something similar results from using an evolutionary or metaheuristic algorithm, that included none of the analytical necessary conditions, it will be better understood.

3 Evolutionary Algorithms and Metaheuristics

> "Evolutionary computation has as its objective to mimic processes from natural evolution, where the main concept is survival of the fittest: the weak must die." A. Engelbrecht [11].

A qualitatively different approach, long used for parameter optimization problems and more recently applied to dynamical system optimization, is the use of "evolutionary" algorithms (EA). The best known of the EA's is the genetic algorithm (GA) [10]. Evolutionary algorithms use the principle of "survival of the fittest" applied to a population of individuals representing candidate solutions for the optimal trajectories. To this end, they employ mechanisms inspired by nature: selection, reproduction, and mutation. Metaheuristic methods are qualitatively similar to evolutionary algorithms but "evolve" differently. Evolutionary algorithms discard poorly performing individuals (i.e., solutions) and proceed with the survivors and new individuals created from the survivors. In metaheuristic methods, the best known of which may be particle swarm optimization (PSO) [12], the entire population takes steps with the intention of improving the fitness of all the individuals.

EAs and metaheuristics are numerical optimizers that determine an optimal set of discrete parameters that has been used to characterize the problem solution. They are similar in many respects, particularly with regard to what they require to operate (an objective, constraints, and bounds on the parameters) and in what they do not require (gradients, Jacobian, Hessian of the system). These methods have two principal advantages over all of the direct and indirect solution methods previously described; they require no initial "guess" of the solution (in fact they generate a population of initial solutions randomly) and they are more likely than other methods to locate a global minimum in the search space rather than be attracted to a local minimum.

The EAs and metaheuristics require that the problem solution be capable of being described by a relatively "small" set of discrete parameters, i.e., small in comparison to the thousands or tens of thousands of parameters that may comprise the vector of parameters of a nonlinear program. This can be accomplished, for dynamic system optimization problems, in a number of ways:

(i) If the trajectory can naturally be described by a finite set, e.g., if the control inputs are applied impulsively or in a bang-bang fashion so that the parameters will be such things as times, magnitudes, directions, and durations of inputs. In this case a small number of parameters will "naturally" suffice to completely describe the solution.

(ii) If the trajectory contains non-integrable arcs it is still the case that much of the trajectory can be described with a small number of parameters such as times of significant events or initial or final states that are to be optimized. Quantities that must be known continuously, such as a control time history, can be described using a comparatively small number of parameters. For example, the time history can be represented by B-splines [21, 22], or polynomials in time [23] or Fourier series [24]. Such time histories can be placed in a sequence so that the entire time interval is spanned by several continuous functions. Then the additional parameters are a small number of polynomial coefficients or B-spline coefficients or the amplitudes, frequencies, and phases of trigonometric functions, as appropriate. Of course if this approach is taken the resulting solution will be sub-optimal because the control time history has been constrained.

(iii) If the analytical necessary conditions of the problem are derived, as shown in the example above, then the problem can be reduced to a TPBVP where the unknowns are primarily the values of all or some of the initial adjoint variables and possibly some of the initial or final state variables, and also possibly the final (or even the initial) time. The states are then found by numerical integration and the control is found using Pontryagin's principle. In any event the continuous problem has been reduced to a finite parameter optimization problem.

There are no analytical necessary conditions that must be satisfied by a solution obtained through the use of evolutionary computation or by a metaheuristic, so there is no guarantee that even if the iterative solution process converges to a single point that point represents a local (much less a global) minimum [25]. Conditions under which convergence to a global minimum is guaranteed can be found but such conditions are not satisfied by EAs or metaheuristics in general [26, 27].

Among the best known and most often employed EAs and metaheuristics are

Genetic algorithms (GA) that model genetic evolution [10, 11].

Differential evolution algorithms (DE) similar to GA but for continuous-valued problems; also the mutation operator is dependent on the current population [25, 28].

Simulated annealing (SA) is a probabilistic metaheuristic inspired by the physical process of annealing in metallurgy [29].

Particle swarm algorithms (PSO) that model cooperative behavior of a swarm; e.g., a flock of birds [12].

Ant colony algorithms (ACO) that model the foraging behavior of ants.

In the following sections brief introductions will be given to GA, DE, and PSO, primarily because these are the EAs and metaheuristics that seem to be most often used for dynamic optimization problems (and some will be used in the examples to follow). This survey cannot describe any of the algorithms (and their many variants) in detail, but this is unnecessary as there is a vast literature on these methods.

4 Genetic Algorithms

In the simplest form of the genetic algorithm the set of parameters describing the solution is written as a string or sequence of numbers [10, 11]. Suppose that this sequence is converted to binary form; it is then similar to a chromosome, but

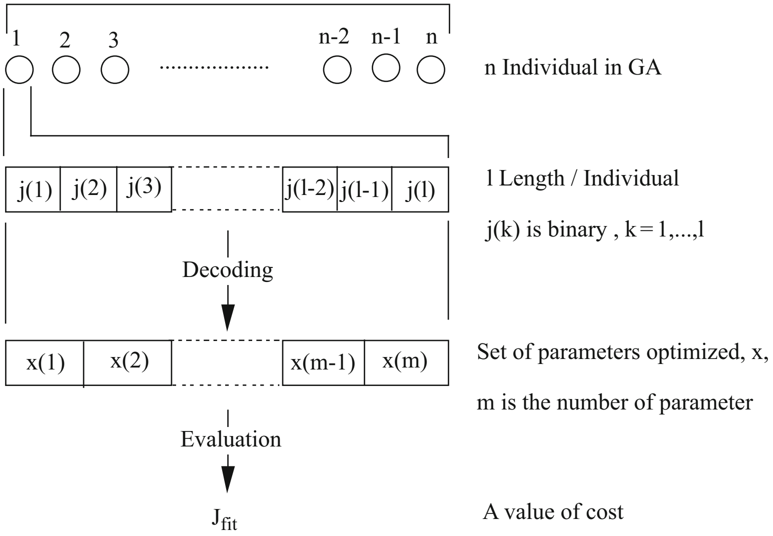


Fig. 1 Illustrating the process of decoding a GA individual

consisting only of two possible variables, a 1 or a 0. (There are also real GAs but the binary algorithm is the easiest to describe.) Every sequence can be "decoded" to yield a trajectory, whose cost or objective value can be determined. The first step in the GA is the generation of a "population" of sequences using a random process. Figure 1 illustrates how a sequence is decoded to yield the value of the cost or objective function. The great majority of these randomly generated sequences will have very large costs; many may even be infeasible. The population is then improved using three natural processes; *selection*, *combination*, and *mutation*. Selection removes the worst sequences and may also, via *elitism*, guarantee that the best sequence survives into the next generation unchanged. Following selection, remaining sequences are used as "parents"; i.e., partial sequences from two parents are combined to form new individuals. Finally, mutation changes a randomly chosen bit (from 0 to 1 or vice versa) in a small fraction of the population.

The process is then repeated; the cost of every individual in the new generation is determined. Since the best individual from the previous generation was retained, the objective may improve but cannot worsen. In practice there is generally rapid improvement in the early generations; if the process locates the global minimum then of course improvement will cease. Termination of the algorithm is usually done after either a fixed number of generations or after the objective has reached a plateau. Of course neither of these termination conditions guarantees that a minimum has been found, nor are there necessary conditions for optimality with this method. Additional shortcomings are that there is no way to enforce satisfaction of boundary conditions; normally a "penalty function" approach is taken in which unsatisfied boundary conditions are added to the cost (as will be described in the next section), and that the solution will be less accurate than a typical direct solution (and even less accurate than a indirect solution). Nevertheless, the method has been very useful when applied to optimizing space trajectories, either for finding approximate solutions or when used to provide an initial guess for more accurate methods, e.g., collocation with NLP. Betts [8] notes that one significant advantage of the GA in comparison to all other solution methods is how straightforward it is to use. There are many GA routines available (a commonly used one is found in MATLAB) so the user need only provide a subroutine for decoding the sequence to evaluate the cost, which for space trajectory problems can be as simple as a routine that integrates the system equations of motion, provide bounds on the parameters, and then provide values for certain constant parameters that control the evolutionary processes.

5 Particle Swarm Optimization

In particle swarm optimization (PSO), some number (say 100) of particles are randomly distributed in an N-dimensional decision parameter space. The objective value is determined for the solution vector corresponding to each particle. Taking an anthropomorphic view, it is then assumed that the particles can communicate so that all know the objective value for all the others [11]. Let $x_i(n)$ denote the position of particle i at the nth time step. At the next iteration, the particles take a step $v_i\,(n+1)$ in the parameter space so that the new position of particle i becomes:

$$x_i\,(n+1) = x_i(n) + v_i\,(n+1) \tag{9}$$

with (in one form of the PSO)

$$v_{ij}\,(n+1) = v_{ij}(n) + c_1 r_{1j}(n)\left[y_{ij}(n) - x_{ij}(n)\right] + c_2 r_{2j}(n)\left[\widehat{y}_j(n) - x_{ij}(n)\right] \tag{10}$$

where $v_{ij}(n)$ is the velocity (step) for component j of particle i at time step n, $x_{ij}(n)$ is the jth component of the position of particle i at the nth time step, $r_{1j}(n)$ and $r_{2j}(n) \subset U\,(0,1)$ are random values in the range [0, 1] sampled from a uniform distribution. $y_i(n)$ is the "personal best" position, the best position located by the ith particle since the first time step; $\hat{y}_j(n)$ is the "global best" position, the best position located by the any particle of the swarm since the first time step. The step described in Eq. (10) thus has three components. The first is an "inertia" that causes the particle to move in the direction it had previously been moving, the second "nostalgia" or "cognitive" component reflects a tendency for the particle to move toward its own most satisfactory position and the third "social" component draws the particle toward the best position found by any of its colleagues. The c's are constants that weight the importance of the three components and the r's provide stochasticity to the system.

As with the GA, the process can be terminated after a fixed number of iterations or when the "best" solution has not changed for several iterations. This method has proven quite robust, is also very simple to use, and is particularly good in locating global minima when the solution space contains many local minima.

6 Differential Evolution

Differential evolution (DE), first developed by Storn and Price [28], is a population-based evolutionary algorithm that, unlike PSO, does not require a concept of velocity within the decision space. Differential evolution comes in many "strategies," and the specific one used here is called *best/2/bin*. Consider an n-dimensional vector space, and generate a population P of vectors within that space. Then, four vectors $\overrightarrow{u}_1$ through $\overrightarrow{u}_4$ are selected randomly from the population and combined via:

$$\overrightarrow{d} = \overrightarrow{u}_1 - \overrightarrow{u}_2 + \overrightarrow{u}_3 - \overrightarrow{u}_4. \tag{11}$$

The resulting difference vector is then multiplied by a scaling factor F and added to the best known vector in the population $\overrightarrow{u}_{best}$ to create the *trial vector* $\overrightarrow{u}_*$. Finally a fifth vector $\overrightarrow{u}_5$ is randomly selected from the population. The fitnesses of $\overrightarrow{u}_5$ and $\overrightarrow{u}_*$ are then evaluated, and the more-fit vector becomes an element of the population for the next generation. The less fit vector is discarded. In total, a fraction (the crossover ratio) of the population will be perturbed and replaced in this manner. CR is defined as the crossover ratio. This process is repeated for a set number of generations or until some convergence criterion is reached.

Differential evolution has been used in the determination of a number of sophisticated (multiple flyby, multiple impulse) space trajectory optimizations [25, 30–32]. A technique that has also been employed is to alternate between two EAs; running a population through some small fixed number of generations with one optimizer before switching and doing the same with the other. In particular, results from alternating between a GA and a DE have been found in some problems to be better than using either EA alone [31, 32].

7 Including Constraints

All the heuristic algorithms use penalty function methods for incorporating constraints. There are many versions of such methods in the EA literature, e.g., in [10, 11, 33, 34]. Typically if the constraints are a combination of inequality and equality constraints of the form:

$$\begin{aligned} g_j\left(\overrightarrow{x}_i\right) &\leq 0, \quad \text{for } 1 \leq j \leq q \\ h_j\left(\overrightarrow{x}\right) &= 0, \quad \text{for } q+1 \leq j \leq m \end{aligned} \tag{12}$$

Then the objective function or “fitness” of the solution can be defined as:

$$fitness\left(\overrightarrow{x}\right) = \begin{cases} f\left(\overrightarrow{x}\right), & if\ \overrightarrow{x} \in F \\ f\left(\overrightarrow{x}\right) + penalty\left(\overrightarrow{x}\right), & otherwise \end{cases} \tag{13}$$

where F represents the family of solutions satisfying all of the constraints. The penalty function is

$$penalty\left(\overrightarrow{x}\right) = \sum_{j=1}^{m} w_j \cdot v_j\left(\overrightarrow{x}\right) \text{ where } v_j\left(\overrightarrow{x}\right) = \begin{cases} max\left(0,\ g_j\left(\overrightarrow{x}_i\right)\right), & 1 \le j \le q \\ \left|h_j\left(\overrightarrow{x}\right)\right|, & \text{if } q+1 \le j \le m \end{cases} \tag{14}$$

One of the principal distinctions among the various penalty function methods for including constraints is whether the weighting coefficients w for the constraints are constants or possibly vary dynamically with time or according to the convergence of the solution. There is undoubtedly no single penalty function method that is best for all problems. Typically users experiment with various choices and keep what works the best on a given problem.

8 Allowing the EA to Determine the Optimal Solution Structure

As mentioned previously, if a heuristic method is to be used for a problem in which the continuous optimal control history needs to be provided (and optimized) this history needs to be described in some way that requires only a relatively small number of parameters. This can be done, for example, by modeling the control as a polynomial function in time [23], or with a Fourier series [24] or using a sum of B-spline basis functions $B_{i,p}$ with distinct interior knot points [35],

$$\overline{u}(t) = \sum_{i=1}^{p+L} \alpha_i B_{i,p}(t) \tag{15}$$

where p is the degree of the splines and L is the number of subintervals in the domain. The ith B-spline of degree p is defined recursively by the Cox-de Boor formula [36],

$$\begin{aligned} B_{i,0}(t) &= \begin{cases} 1 & \text{if } t_i \le t \le t_{i+1} \\ 0 & \text{otherwise} \end{cases} \\ B_{i,p}(t) &= \tfrac{t-t_i}{t_{i+p}-t_i} B_{i,p-1}(t) + \tfrac{t_{i+p+1}-t}{t_{i+p+1}-t_{i+1}} B_{i+1,p-1}(t) \end{aligned} \tag{16}$$

Now, before parameterizing the control function, the problem of selecting the shape of the latter must be addressed; should the search be conducted in the space of straight lines or higher-degree curves or oscillatory functions? Clearly, the actual control history is known only a posteriori. For instance, in the case of the well-known Brachistochrone problem, the optimal control is linear in time [1], and an attempt to capture it using, say, a sum of higher-degree B-spline basis functions would yield, even after optimization, a sub-optimal solution. For many problems of course one has some intuition regarding what the likely form of the optimal control will be and this can guide the selection of a form for parameterizing the control history. It is also the case that a poor choice can be improved by trial-and-error.

One advantage of the evolutionary algorithms is that it is possible to allow the details of the parameterization to be described by decision parameters and then allow the EA optimizer to "simultaneously" determine both the optimal form of the control parameterization and the optimal trajectory [21]. For example, if the B-spline method is chosen, it may be left to the EA to choose the appropriate B-spline degree p in addition to the coefficients α_{i} so as to minimize the objective function and satisfy all of the boundary conditions. This approach proves particularly useful in multi-phase trajectory optimization problems where controls in different phases may need to be described by different types of functions. Precisely the same technique can be used to allow the EA to choose the optimal degree of a polynomial (in time) approximation of the control history, etc.

9 Application to Hybrid Optimal Control Problems

Evolutionary algorithms and metaheuristics are also beneficial in the solution of hybrid optimal control problems (HOCP). These are problems involving both continuous variables and categorical (discrete) variables. Perhaps the best-known HOCP is the travelling salesman problem [2]. Each selection of a sequence of cities to be visited by the salesman represents a set of categorical variables. For each such sequence there is a corresponding continuous optimal trajectory problem to solve to determine the optimal, minimum-time, or minimum-distance path.

Our research group and others have had success solving such problems using a nested loop approach where the outer loop finds the optimal sequence of categorical variables, using typically a genetic algorithm, while the inner loop determines the corresponding optimal cost for the continuous system using either an EA or a direct transcription method. Perhaps the most sophisticated HOCPs yet solved with this approach are those for optimal space mission planning [32, 37, 38]. In these problems the categorical variables are the planets to be used for hyperbolic flybys, whose number is not even known a priori. An archetypical example would be trajectory planning for the Galileo or Cassini missions. There may be other possible ways to solve HOCPs, but the use of a GA to solve (at least) for the categorical variables is natural and efficient.

10 Advantages and Disadvantages of EA

Evolutionary algorithms and metaheuristics thus have the following advantages and disadvantages for trajectory optimization of continuous systems:

10.1 Advantages

Straightforward to code (possibly the most-straightforward extant method to code).

Don't need to know optimal control theory; don't need possibly difficult analytical differentiation.

Requires no initial guess; the initial population is chosen randomly.

More likely than other methods to locate the global minimum.

Capable of determining the optimal solution structure, e.g., the optimal switching structure, at runtime.

10.2 Disadvantages

The problem needs to be parameterized by a (relatively) small number of variables.

The methods depend on a number of user-selectable parameters and it is not a priori clear how these are chosen for a successful or efficient solution.

Likely to need explicit numerical integration of the EOM, which can be time-consuming.

The solution will not be as accurate as that of the CoV necessary conditions or a DT solution.

Constraints need to be included via a penalty function method and this is especially problematic for equality constraints.

There is no guarantee that an optimal or even near-optimal solution will be found by a given method. It may be necessary for the user to try more than one optimizer on a problem.

If a solution is found there is no guarantee of optimality, i.e., there are no "necessary conditions" for optimality as obtain with deterministic CoV-based methods.

11 Examples

The following examples are of course not a proof of the assertions of the previous sections but are intended to illustrate the ease of use and the success obtained when an EA or metaheuristic is applied to dynamic system optimization problems. Only a few examples of the ways in which a continuous optimization problem can be converted into a few-parameter problem for solution via evolutionary algorithms are explicitly shown.

11.1 Manned Asteroid Sample Return Mission with Time Constraint

This example is taken from the M.S. thesis of Ms. Aishwarya Stanley [39]. The objective is to minimize the sum of the velocity impulses (i.e., the total ΔV) required for a mission that will rendezvous with an asteroid, remain in the vicinity of the asteroid for 8–25 days, and then return to Earth. There is an upper bound on the total flight time of 365 days. The asteroid target is not specified a priori but is chosen from a catalog of near-Earth asteroids (NEA) satisfying the following criteria:

1. Allow Earth departure dates between 2025 and 2035
2. Are at least 30 m in diameter
3. Allow 365 day round trip missions

   and *are not* limited by the following considerations:
4. Location uncertain and/or limited Earth-based observation opportunity to improve prior to the human mission,
5. Few departure opportunities,
6. Likely too small based on estimated albedo (albedo assumed to be between 0.05 and 0.25).

This is what we have termed a "naturally discrete" problem because the solution of the continuous problem depends on only four quantities; the dates of Earth departure, asteroid interception, asteroid departure, and Earth arrival. With those dates specified, the ΔV's required can be found using Lambert's method [40]. The PSO method is used for the optimization with the constraints of a 365 day total flight time and the 8–25 day stay time enforced through a penalty function. Some of the results obtained are shown in Table 1. Note that for some of the optimal trajectories the various flight time constraints are effective; i.e., one or more of the flight times are at an upper or lower bound (this implies that loosening the constraints could result in a lower cost).

11.2 Two-Burn Low-Thrust Spacecraft Rendezvous

This example solves for a minimum-fuel low-thrust transfer from a circular orbit of radius 1 to rendezvous with a target that is already in circular orbit of radius 3 (in normalized units). A solution of the form coast-thrust-coast-thrust is assumed (which in fact is the optimal structure). The governing equations are

$$\begin{aligned}
&\frac{dr}{dt} = v_r, \quad r(0) = 1 \\
&\frac{d\theta}{dt} = v_\theta / r, \quad \theta(0) = 0 \\
&\frac{dv_r}{dt} = {v_\theta}^2/r + g(r) + sa\sin\beta, \quad v_r(0) = 0 \\
&\frac{dv_\theta}{dt} = -v_\theta v_r / r + sa\cos\beta, \quad v_\theta(0) = 1 \\
&\tfrac{da}{dt} = sa^2/c, \quad a(0) = 0.1
\end{aligned} \tag{17}$$

Table 1 Optimal asteroid sample return trajectories found using PSO

Asteroid name	Pop, Gen	Lower bounds	Upper bounds	Desired epoch date, t	Total delta-V (km/s)	Optimal t_launch, days	Optimal launch date	Optimal t_flight1, t_flight2 days	Optimal t_wait, days	Total mission duration, days
99942 Apophis	100, 2000	[30,90, 8,100]	[3000,782,25, 782]	Jan 1, 2023	7.4688	2757.2	July 20, 2030	198.7 138.9	25	362.6
99942 Apophis	100, 2000	[30,90, 8,100]	[3000,782,25, 782]	Sep 30, 2022	7.6139	2850.2	July 20, 2030	202.9 146.3	12.7	361.9
2011 AA 37	100, 2000	[60,90, 8,90]	[1000,782,25, 782]	Jan 18, 2028	7.2843	88.2058	Apr 15, 2028	173.6769 175.6233	14.6220	363.9222
2011 AA 37	100, 2000	[60,90, 8,90]	[1000,782,25, 782]	Apr 18, 2027	7.1332	361.2868	Apr 13, 2028	184.2333 172.6753	8.0165	364.9251
2007 YF	100, 2000	[30,80, 8,80]	[4000,782,25, 782]	Apr 18, 2021	6.1038	3464.8	Oct 13, 2030	154 83.7	17.6	255.3
2007 YF	100, 2000	[30,80, 8,80]	[800,782,25, 782]	Sep 1, 2029	5.9626	407.6332	Oct 14, 2030	156.2895 90.8209	8.0005	255.1109
2009 CV	100, 2000	[30,90, 8,100]	[4000,782,25, 782]	Apr 18, 2023	9.1055	606.6053	Dec 15, 2024	190.7190 134.5491	24.9987	350.2668
2009 CV	100, 2000	[30,90, 8,100]	[300,782,25, 782]	June 1, 2024	9.1052	196.9495	Dec 15, 2024	190.7593 134.7199	24.9998	350.479

As previously described, the control, β, must somehow be described by a tractable number of parameters in the thrust arcs where it is a continuous quantity. There are many ways to do this but we have chosen to describe β using a cubic polynomial in time. This problem is solved with a GA. It is transcribed into a 11 parameter problem requiring 58 binary bits:

1: initial coast duration (six binary bits)
2: 1st thrust arc duration (six bits)
3: 2nd thrust arc duration (six bits)
4–7: 1st polynomial coefficients (centered about 1st thrust time of flight) (five bits each)
8–11: 2nd polynomial coefficients (centered about 2nd thrust time of flight) (five bits each)

A population size of 50 is used; the probabilities for crossover and mutation are 50 and 0.5 %, respectively, and the GA is run for 10,000 generations. The GA employs elitism, i.e., the best solution found in a given iteration is always propagated unchanged into the next generation.

The resulting trajectory is shown in Fig. 2; the time history of the optimal control is shown in Fig. 3. The exact optimal solution has been determined, using a direct transcription method and NLP, for comparison. Table 2 shows that despite the approximation of the control using a polynomial, the GA finds the solution quite accurately. In particular, the total thrust time, which determines fuel required, is accurate to a small fraction of 1 %.

11.3 Max-Radius Orbit Transfer Using Solar Sail

The following example is drawn from the Ph.D. thesis of Mr. Pradipto Ghosh [21, 41]. The objective is to determine the solar sail orientation history (the control) in order to transfer the vehicle from a specified initial circular orbit to the largest possible co-planar circular orbit in a fixed time (of 450 days in this case). The system equations are

$$\begin{aligned}
\dot{r} &= v_r, \qquad r(0) = 1\\
\dot{\theta} &= \tfrac{v_\theta}{r}\\
\dot{v}_r &= \tfrac{v_\theta^2}{r} - \tfrac{\mu}{r^2} + a\tfrac{\cos^3\alpha}{r^2}\\
\dot{v}_\theta &= -\tfrac{v_r v_\theta}{r} + a\tfrac{\sin\alpha\cos^2\alpha}{r^2}
\end{aligned} \tag{18}$$

The problem becomes

$$\min_{\alpha(\cdot)} J\left[x(\cdot),\ \alpha(\cdot),\ t_f\right] = -r\left(t_f\right) \tag{19}$$

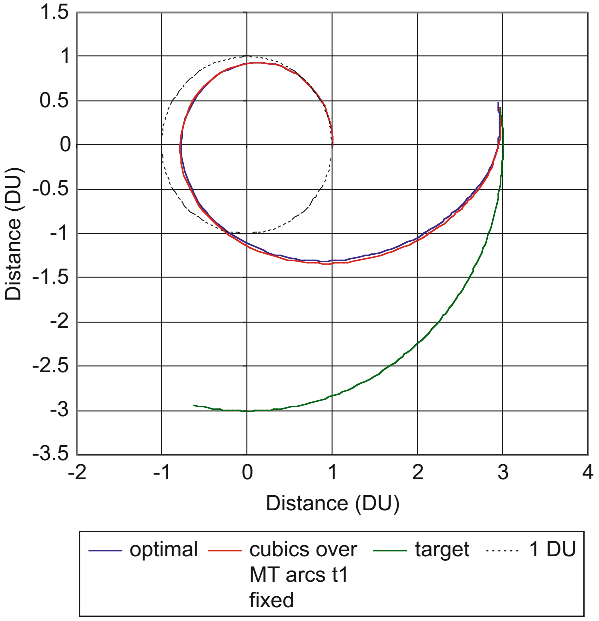


Fig. 2 Trajectory for optimal rendezvous

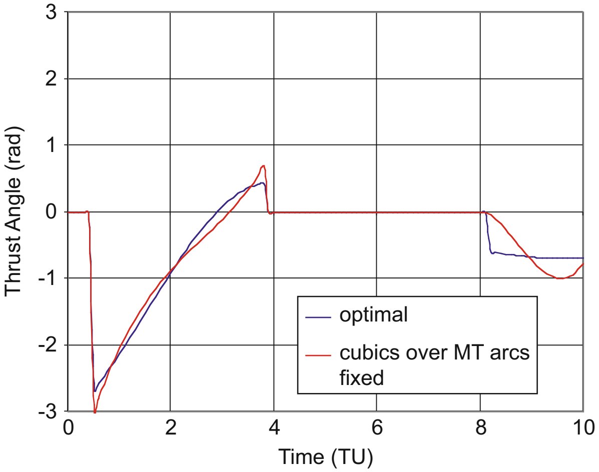


Fig. 3 Illustrating how the GA approximates the optimal control

Table 2 Comparison of GA and true solutions

Solution using GA	Exact solution
First coast arc $\Delta t = 0.4576$	First coast arc $\Delta t = 0.4560$
First thrust arc $\Delta t = 3.404$	First thrust arc $\Delta t = 3.404$
Second coast arc $\Delta t = 4.299$	Second coast arc $\Delta t = 4.316$
Second thrust arc $\Delta t = 1.839$	Second thrust arc $\Delta t = 1.824$
Total thrust time $= 5.243$	Total thrust time $= 5.228$

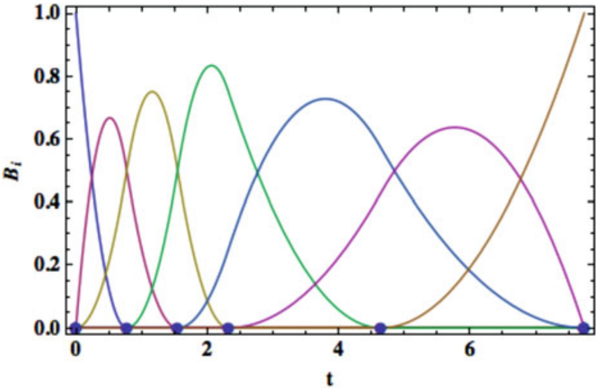


Fig. 4 B-spline curves from which solution is constructed

The condition of a final circular orbit requires

$$\begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} v_\theta\left(t_f\right) - \sqrt{\mu/r\left(t_f\right)} \\ v_r\left(t_f\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{20}$$

The solution for the solar sail pointing angle α (t) was approximated as a sum of 7 quadratic splines.

The degree of the spline $p_{s,k}$ is determined according to the "degree-parameter" $m_{s,k}$, which decides the B-spline degree of the **s**th control in the **k**th phase in the following fashion:

$$p_{s,k} = \begin{cases} 1 \text{ if } -2 \leq m_{s,k} < -1 \\ 2 \text{ if } \quad -1 \leq m_{s,k} < 0 \\ 3 \text{ if } \quad\; 0 \leq m_{s,k} < 1 \end{cases} \tag{21}$$

The B-splines from which the solution is constructed are shown in Fig. 4.

The optimal values of the 7 coefficients are found using PSO and are shown in Table 3. The PSO solver uses a population of 200 and iterates only 50 times before converging.

As in the previous example, a more accurate solver using direct transcription and NLP is used to find an "exact" solution of the problem for comparison. Figure 5

Table 3 Optimal B-spline coefficients

Parameter	Value
α_1	0.4377
α_2	0.3879
α_3	0.1665
α_4	1.096
α_5	0.9228
α_6	0.7494
α_7	0.6798
$m_{1,1}$	−0.3364

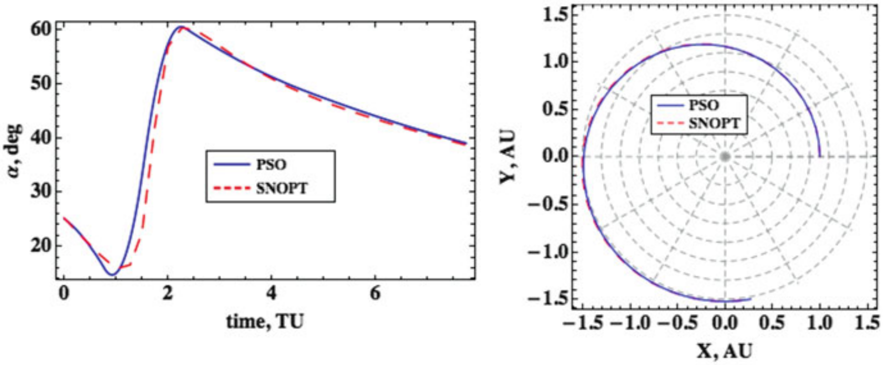


Fig. 5 Solar sail control history and optimal trajectories

shows the time histories of the optimal control. Figure 5 shows the PSO and “exact” trajectories. The objective final radius, to be maximized, is 1.527 for the PSO solution and 1.525 for the “exact” solution, a negligible difference. The terminal constraints of Eq. (20) are satisfied to 2.2E-8 and 6E-3, respectively.

11.4 Optimal Deflection of Earth-Approaching Asteroid with Consideration of Future Events

This work [42] was done as part of the NEO-Shield initiative and has application to the AIDA mission that is scheduled for 2020. It was done by the author and collaborators Dr. Siegfried Eggl and Dr. Daniel Hestroffer at the Observatory of Paris. The concept is to have a spacecraft impact a small near-Earth asteroid (NEA) and then measure the subsequent deflection from the asteroid’s original orbit. However, one criterion that may be enforced for the mission is that the deflection must not make any future (i.e., next 125 years) encounters with Earth closer than they would have been absent the deflection. The problem can be formulated as an optimization problem:

(i) Choose launch date and impact date (from which impact location, direction, and relative velocity can be determined). There is an upper bound on departure ΔV corresponding to the capability of the rocket the spacecraft is launched with.
(ii) Considering impact's change in asteroid momentum, integrate system EOM forward to find change in close approach distances to Earth at first and future encounters.
(iii) Then want to maximize first deflection (Δ_1), but with all future encounter distances *only increased*.

The problem has been solved for a proof-of-concept case (not a real NEA), using PSO, demonstrating that the method is capable of being applied to whichever NEA is ultimately selected for the AIDA mission.

In order to make the constraint, of not worsening any future Earth encounter distances, effective, an "artificial" asteroid in a 5/4 orbit period commensurability with Earth is the target. This causes the original close approach distance to periodically recur; thus, absent the deflection impulse the close approach would neither worsen nor improve. (Many NEAs are in near-commensurabilities with Earth so that this is not an unlikely initial condition.) The asteroid and Earth orbits are shown in Fig. 6.

A PSO optimizer (MATLAB's *fmincon*) is used to solve the problem. The objective function is a linear combination of the close approach distances at the

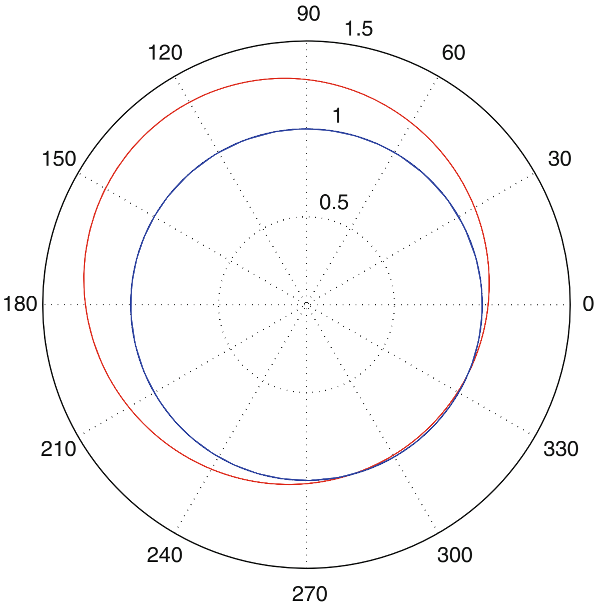


Fig. 6 Earth orbit and asteroid in 5/4 orbit commensurability

Table 4 Variation of deflection result with different cost functions

Objective function	$\Delta_1 + \Delta_2$	$\Delta_1 + \Delta_2/25$	$\Delta_1 + \Delta_2/125$
Δ_1 (R_{Earth})	0.52	0.68	0.74
Δ_2 (R_{Earth})	18.7	17.1	14.1
Beta	3.067 = 175.7°	3.461 = 198.2°	3.775 = 216.3°

Table 5 Variation of deflection result with different cost functions and models

Objective function	None	None	Δ_1	$\Delta_1 + \Delta_2/5$
Yarkovsky?	No	Yes	Yes	Yes
Δ_1 (R_{Earth})	0.09	0.07	0.74	0.40
Δ_2 (R_{Earth})	0.65	29.6	17.8	48.3
Beta (rad)	NA	NA	3.922	0.016

first encounter after the deflection, Δ_1, and subsequent encounter, Δ_2. In Table 4 one sees in the first column that when these distances are equally weighted the first deflection is of 0.52 Earth radii and the second much larger at 18.7 Earth radii. (The second close approach distance Δ_2 is naturally much larger than Δ_1 because the deflection velocity impulse has had an additional 5 years to act on changing the position of the asteroid, from what it would otherwise have been.) Making Δ_1 more important changes the results only a small amount. Table 5 shows, in the first two columns, the close approach distances absent a deflection impulse but without or with a consideration of the Yarkovsky effect. Of particular note is that when the Yarkovsky effect is included, Δ_2 is 29.6. However, an optimization that maximizes only Δ_1, which is useful for observing the more immediate effect of the spacecraft impact, reduces the second close approach distance from 29.6 to 17.8, i.e., this violates the mission requirement that future encounter distance only be increased as a result of the initial deflection impulse. The problem can be remedied, as seen in the last column, by optimizing a linear combination of the first and second encounters; then the second close approach distance increases to 48.3. This result for the proof-of-concept example shows that PSO can effectively, and simply, solve a seemingly difficult problem.

11.5 Optimal Low-Thrust Transfer Between Arbitrary Elliptic Orbits

This example, taken from a paper by Pontani and Conway [43], shows an alternative way of transcribing an infinite dimensional (continuous) optimal control problem into a few-parameter system for solution by a metaheuristic method. The problem is a low-thrust transfer between co-planar elliptic orbits; the starting position and the time of flight are free parameters. The spacecraft mass and the thrust magnitude are assumed constant; thus the only control is the thrust pointing angle. The objective is to accomplish the transfer using minimum fuel, but since the engine is always on this is equivalent to minimizing the transfer time t_f. The state variables of the problem are radial and azimuthal velocity, and radial and azimuthal position, i.e., $x = [x_1\ x_2\ x_3\ x_4]^T = [v_r\ v_\theta\ r\ \xi]^T$. The equations of motion are then:

$$\dot{v}_r = -\tfrac{\mu - r v_\theta^2}{r^2} + \tfrac{T}{m}\sin\delta\ \dot{v}_\theta = -\tfrac{v_r v_\theta}{r} + \tfrac{T}{m}\cos\delta\ \dot{r} = v_r\ \dot{\xi} = \tfrac{v_\theta}{r} \tag{22}$$

where dimensionless variables are employed.

In order to transcribe the problem into one depending on only a small number of parameters, the calculus of variations necessary conditions are derived. There are 7 system constraints representing initial and final conditions. The system Hamiltonian and the function Φ that adjoins these constraints to the cost function are

$$\begin{aligned}
&H\underline{\underline{\Delta}}\,\lambda_1\left[\tfrac{\mu - x_3 x_2^2}{x_3^2} + \tfrac{T}{m}\sin\delta\right] + \lambda_2\left[-\tfrac{x_1 x_2}{x_3} + \tfrac{T}{m}\cos\delta\right] + \lambda_3 x_1 + \lambda_4 \tfrac{x_2}{x_3}\\
&\Phi\underline{\underline{\Delta}}\,v_1\left[x_{10} - \sqrt{\tfrac{\mu}{a_0\left(1-e_0^2\right)}}e_0\sin f_o\right] + v_2\left[x_{20} - \sqrt{\tfrac{\mu}{a_0\left(1-e_0^2\right)}}\left(1 + e_0\cos f_o\right)\right]\\
&+v_3\left[x_{30} - \tfrac{a_0\left(1-e_0^2\right)}{1+e_0\cos f_0}\right] + v_4\left[x_{40} - f_0\right] + v_5\left[x_{1f} - \sqrt{\tfrac{\mu}{a_f\left(1-e_f^2\right)}}e_f\sin f_\mathsf{f}\right]\\
&+v_6\left[x_{2f} - \sqrt{\tfrac{\mu}{a_f\left(1-e_f^2\right)}}\left(1+e_f\cos f_\mathsf{f}\right)\right] + v_7\left[x_{3f} - \tfrac{a_f\left(1-e_f^2\right)}{1+e_f\cos f_\mathsf{f}}\right] + v_8\left[x_{4f} - \phi - f_\mathsf{f}\right] + t_f
\end{aligned} \tag{23}$$

In these expressions a, e, and f are conventional elliptic elements. Initial and final conditions are represented by subscripts 0 and f, respectively. ϕ is the angle between the major axes of the initial and final orbits. Note that true anomaly f_f is completely determined if transfer (final) time t_f is known.

It can be shown [42] that the system Euler–Lagrange equations yield the optimal control solely as a function of the costate variables:

$$\cos\ u^* = -\frac{\lambda_2^*}{\sqrt{\left(\lambda_1^*\right)^2 + \left(\lambda_2^*\right)^2}} \quad \text{and} \quad \sin\ u^* = -\frac{\lambda_1^*}{\sqrt{\left(\lambda_1^*\right)^2 + \left(\lambda_2^*\right)^2}} \tag{24}$$

where necessary conditions on these costates are

$$\begin{aligned}
\dot{\lambda}_1^* &= -\lambda_3^* + \frac{x_2^*\lambda_2^*}{x_3^*},\ \dot{\lambda}_2^* = -\lambda_3^* + \frac{-2x_2^*\lambda_1^* + x_1^*\lambda_2^* - \lambda_4^*}{x_3^*},\\
\dot{\lambda}_3^* &= \frac{\left(x_2^*\right)^2\lambda_1^* - x_1^*x_2^*\lambda_2^* + \lambda_4^*x_2^*}{\left(x_3^*\right)^2} - \frac{2\mu\lambda_1^*}{\left(x_3^*\right)^2},\ \dot{\lambda}_4^* = 0
\end{aligned} \tag{25}$$

with

$$\begin{array}{llll}
\lambda_{10}^* = -v_1 & \lambda_{20}^* = -v_2 & \lambda_{30}^* = -v_3 & \lambda_{40}^* = -v\ {}_4\\
\lambda_{1f}^* = -v_5 & \lambda_{2f}^* = -v_6 & \lambda_{3f}^* = -v_7 & \lambda_{4f}^* = -v_8
\end{array} \tag{26}$$

It can also be shown that the (constant) costate variable λ_4 is not independent of the other costates and that the solution thus depends only on 5 parameters $[\lambda_{10}\ \lambda_{20}\ \lambda_{30} f_0\ t_f]$. The PSO solution must thus find a solution for these 5 parameters that satisfies the system and costate EOM (22) and (25), the costate terminal boundary conditions (26) and the state terminal conditions:

$$\begin{aligned} d_1 &= v_r\left(t_f\right) - \sqrt{\frac{\mu}{a_f\left(1-e_f^2\right)}} e_f \sin f_{\mathrm{f}} & d_2 &= v_\theta\left(t_f\right) - \sqrt{\frac{\mu}{a_f\left(1-e_f^2\right)}}\left(1 + e_f \cos f_{\mathrm{f}}\right) \\ d_3 &= r\left(t_f\right) - \frac{a_f\left(1-e_f^2\right)}{1+e_f \cos f_{\mathrm{f}}} & d_4 &= \xi\left(t_f\right) - \left(\phi + f_{\mathrm{f}}\right) \end{aligned} \tag{27}$$

with the optimal control determined using (24). Thus the objective function is only

$$\tilde{J} = \sum_{k-1}^{4} |d_k| \tag{28}$$

having a desired value $J = 0$.

To satisfy the state and costate EOM, the PSO calls a numerical integration routine (for each particle) from initial conditions determined by the 5 PSO parameters. Thus the PSO, for this problem, acts in some respects like a conventional "shooting" method, but of course without any gradient information and without the need for an initial guess. As an example, here is an optimal transfer between two quite different orbits:

$$\mathrm{a}_0 = 1.8,\ e_0 = 0.6,\ a_f = 3.5,\ e_f = 0.8,\ \phi = 120^o,\ T/m = 0.03$$

The PSO uses 100 particles and runs for 3000 iterations. The optimal value of the transfer time is 19.718 and the objective $J = O\left(10^{-6}\right)$, i.e., the constraints are very well satisfied. The transfer trajectory is shown in Fig. 7 and the thrust pointing angle (δ) time history is shown in Fig. 7.

12 Conclusions

Methods in which the continuous problem (in direct or indirect form) is discretized and converted to an NLP problem have become extremely competent and useful and there is now a large literature of solutions to sophisticated problems using such methods. However recently, methods using evolutionary computation methods and metaheuristics have been applied to continuous problems, also with great success. Some observations can definitely be made, that will hopefully be of value to those who are not yet experienced in solving such problems and are in the process of choosing a potentially good solution approach:

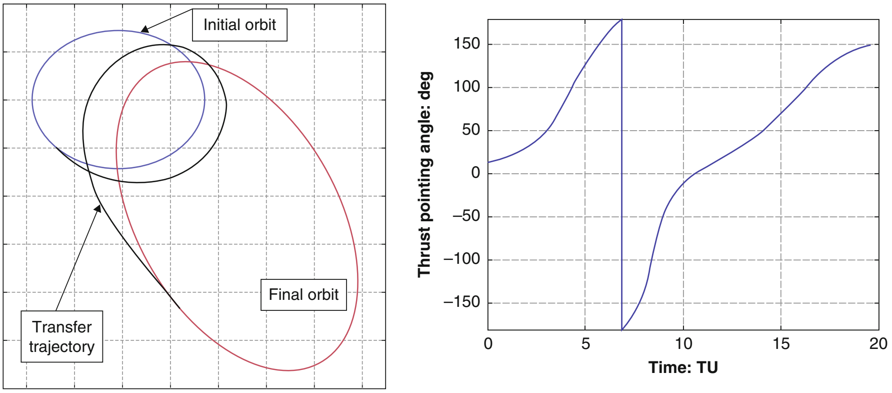


Fig. 7 Optimal trajectory and optimal control for orbit transfer

(i) For many problems, if extreme accuracy is not (or at least is not initially) required, an EA method such as GA or a metaheuristic such as PSO or DE might give a sufficiently good answer with a minimum of programming, and no work at all required for the initial guess!

(ii) For problems requiring greater accuracy, an EA or metaheuristic method can provide a good initial guess, more likely than other methods to be in the vicinity of the global optimum, for a more accurate CoV-based or NLP solution. The EA solution guess/NLP transcription solution refinement is synergistic!

(iii) Because the EA or metaheuristic solution requires parameterization of the solution with a small number of parameters, numerical solutions using the CoV (which are known to generally be quite challenging to solve) may become useful or tractable again. That is, some (most likely many) of the solution parameters can be initial values of the costate (adjoint) variables, which are comparatively small in number. Solutions will then be quite accurate and have guaranteed (local) optimality.

(iv) An EA method is perhaps the only good choice for mission planning problems that are hybrid optimal control problems, i.e., that are continuous optimal control problems that depend in part on the proper selection of a set of discrete decision parameters.

References

1. Bryson, A.E., Ho, Y.-C.: Applied Optimal Control. Hemisphere Publishing Corporation, New York (1975)
2. von Stryk, O., Glocker, M.: Numerical mixed-integer optimal control and motorized travelling salesmen problems. Eur. J. Control **35**, 519–533 (2001)
3. D’Amario, L., et al.: Galileo 1989 VEEGA trajectory design. J. Astronaut. Sci. **37**, 281–306 (1989)

4. Englander, J., Conway, B.A., Williams, T.: Automated mission planning via evolutionary algorithms. J. Guid. Control Dyn. **35**, 1878–1887 (2012)
5. Hargraves, C.R., Paris, S.W.: Direct trajectory optimization using nonlinear programming and collocation. J. Guid. Control. Dyn. **10**, 338–342 (1987)
6. Conway, B.A., Paris, S.W.: Spacecraft Trajectory Optimization Using Direct Transcription and Nonlinear Programming. In: Conway, B.A. (ed.) Spacecraft Trajectory Optimization. Cambridge University Press, Cambridge (2011)
7. Fahroo, F., Ross, I.M.: Direct trajectory optimization by a chebyshev pseudospectral method. J. Guid. Control. Dyn. **25**, 160–166 (2002)
8. Betts, J.T.: Survey of numerical methods for trajectory optimization. J. Guid. Control. Dyn. **21**, 193–207 (1998)
9. Conway, B.A.: Invited paper: a survey of methods available for the numerical optimization of continuous dynamic systems. J. Optim. Theory Appl. **152**(2), 271–306 (2011)
10. Goldberg, D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, New York (1989)
11. Engelbrecht, A.P.: Computational Intelligence, 2nd edn. John Wiley & Sons, New York (2007)
12. Kennedy, J., Eberhart, R.: Swarm Intelligence. Academic, San Diego (2001)
13. Gill, P., Murray, W., Saunders, M.A.: SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev. **47**, 99–131 (2005)
14. Hull, D.G.: Optimal Control Theory for Applications. Springer, New York (2003)
15. Lawden, D.F.: Optimal Trajectories for Space Navigation. Butterworths, London (1963)
16. Prussing, J.E.: Primer Vector Theory and Applications. In: Conway, B.A. (ed.) Spacecraft Trajectory Optimization. Cambridge University Press, Cambridge (2011)
17. Wood, L.J.: Second-order optimality conditions for the Bolza problem with both endpoints variable. J. Aircr. **11**, 212–222 (1974)
18. Jo, J.-W., Prussing, J.E.: Procedure for applying second-order conditions in optimal control problems. J. Guid. Control Dyn. **23**, 241–251 (2000)
19. Kelley, H.J., et al.: Singular Extremals. In: Leitmann, G. (ed.) Topics in Optimization. Academic, New York, NY (1967)
20. Goddard, R.H.: A method of reaching extreme altitudes. Smithson. Inst. Misc. Collect. **71** (1919)
21. Ghosh, P., Conway, B.A.: Numerical trajectory optimization with swarm intelligence and dynamic assignment of solution structure. J. Guid. Control Dyn. **35**, 1178–1192 (2012)
22. Ghosh, P., Conway, B.A.: A direct method for trajectory optimization using the particle. Swarm Approach, Paper AAS 11-155, 21st AAS/AIAA Spaceflight Mechanics Meeting, New Orleans (2011)
23. Wall, B.J., Conway, B.A.: Near-optimal low-thrust earth-mars trajectories found via a genetic algorithm. J. Guid. Control Dyn. **28**, 1027–1032 (2005)
24. Wall, B.J.: Technology for the solution of hybrid optimal control problems in astronautics. Ph.D. thesis, University of Illinois at Urbana (2007)
25. Vasile, M., Minisci, E., Locatelli, M.: An inflationary differential evolution algorithm for space trajectory optimization. IEEE Trans. Evol. Comput. **15**(2), 267–281 (2011)
26. Pinter, J.: Convergence properties of stochastic optimization procedures. Math. Operationsforsch. Statist. Ser. Optim. **15**, 405–427 (1984)
27. Rudolph, G.: Convergence of evolutionary algorithms in general search space. In: Proceedings of the IEEE International Conference on Evolutionary Computation, Nagoya, Japan (1996)
28. Storn, R., Price, K.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. **11**, 341–359 (1997)
29. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
30. Vinkó, T., Izzo, D.: Global Optimisation Heuristics and Test Problems for Preliminary Spacecraft Trajectory Design, European Space Agency, the Advanced Concepts Team, ACT technical report (GOHTPPSTD). http://www.esa.int/gsp/ACT/doc/INF/pub/ACT-TNT-INF-2008-GOHTPPSTD.pdf (2008)

31. Izzo, D.: Global Optimization and Space Pruning for Spacecraft Trajectory Design. In: Conway, B.A. (ed.) Spacecraft Trajectory Optimization. Cambridge University Press, Cambridge (2011)
32. Englander, J., Conway, B.A.: Optimal autonomous mission planning via evolutionary algorithms. Paper AAS 11-159, 21st AAS/AIAA Spaceflight Mechanics Meeting, New Orleans (2011)
33. Hu, X., Eberhart, R.: Solving constrained nonlinear optimization problems with particle swarm optimization. In: Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, FL (2002)
34. Sedlaczek, K., Eberhard, P.: Using augmented Lagrangian particle swarm optimization for constrained problems in engineering. Struct. Multidiscip. Optim. **32**, 277–286 (2006)
35. Martin, C.S., Conway, B.A.: Optimal low-thrust trajectories to the moon with manifolds. Paper AAS 10-105, AAS/AIAA Space Flight Mechanics Meeting, San Diego, CA (2010)
36. de Boor, C.: A Practical Guide to Splines. Springer, New York (2001)
37. Chilan, C.M.: Automated design of multiphase space missions using hybrid optimal control. Ph.D. thesis, University of Illinois at Urbana (2009)
38. Chilan, C., Conway, B.A.: Automated design of multiphase space missions using hybrid optimal control. J. Guid. Control Dyn. **36**, 1410–1424 (2013)
39. Stanley, A.: Identifying near-earth asteroid targets for human exploration using particle swarm optimization. M. S. thesis, University of Illinois at Urbana (2013)
40. Prussing, J.E., Conway, B.A.: Orbital Mechanics, 2nd edn. Oxford University Press, New York, NY (2013)
41. Ghosh, P.: New numerical methods for open-loop and feedback solutions to dynamic optimization problems. Ph.D. thesis, University of Illinois at Urbana (2013)
42. Eggl, S., Conway, B.A., Hestroffer, D.: NEOSHIELD: Finding safe harbors in asteroid deflection missions. Paper IAA-PDC-15-P-83, 4th IAA Planetary Defense Conference – PDC 2015, Frascati, Roma, 13–17 April 2015
43. Pontani, M., Conway, B.A.: Optimal low-thrust orbital maneuvers via indirect swarming method. J. Optim. Theory Appl. **162**, 272–292 (2014)

Composite Thin-Walled Beams by Γ-Convergence: From Theory to Application

Cesare Davini, Lorenzo Freddi, and Roberto Paroni

Abstract In two previous papers we deduced the asymptotic models for fully anisotropic inhomogeneous linear elastic thin-walled beams within the general framework of Γ-convergence. Here we establish a bridge between those mathematical results and their implementation to real problems. In particular, we determine the relationship between the energy densities and the stresses of the asymptotic models with those of the real problem under study.

1 Introduction

The importance of thin-walled beams is widely acknowledged in engineering because of their high structural performances and low weight, that make them attractive for advanced technological applications such as in aeronautics and turbo machinery. In these fields peculiar aspects concerning dynamics, stability, and control lead engineers to explore new areas and novel problems. To face these unconventional applications, it would be desirable to have a unitary theory and clearly deduced models. Unfortunately this is not the case, because in the field there is a vast offer of models based on 'ad hoc' hypotheses and often hardly comparable, as is emphasized in [13].

When dealing with the equilibrium of thin-walled beams, the geometric features of the body naturally lead to the study of problems characterized by two smallness parameters—the diameter of the cross section and the thickness of the wall. Therefore, Γ-convergence theory offers a quite general framework to get asymptotic

C. Davini (✉)
University of Udine, Via Parenzo 17, 33100 Udine, Italy (private address)
e-mail: davini@uniud.it

L. Freddi
DIMA,University of Udine, via delle Scienze 206, 33100 Udine, Italy
e-mail: freddi@dimi.uniud.it

R. Paroni
DADU, Università degli Studi di Sassari, Palazzo del Pou Salit, Piazza Duomo,
07041 Alghero, Italy
e-mail: paroni@uniss.it

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_6

models by means of a minimal and nowadays canonical set of assumptions. This way has been followed by various authors, starting from the simplest instance of thin-walled beams with rectangular [5, 7] or pluri-rectangular [6] isotropic cross-section. In these papers the problem was faced within the linear elasticity framework; thin-walled beams within the three-dimensional theory of nonlinear elasticity have been studied in [4, 8, 9]. Following the same approach, in [2, 3] we considered a fully anisotropic and inhomogeneous linear elastic thin-walled beam and found exhaustive results that cover the cases of beams with both closed and open cross-section. In particular, we got an assessment of the kinematic assumptions that stand at the foundations of Vlasov's theory [10, 12]. For simplicity, here we focus on thin-walled beams with open section.

The analysis of [2, 3] is rather complex, because it passes through different rescalings of the various components of the displacement field and because it calls upon analytical details that could be hardly appreciated by those who are mainly interested in applications. One purpose of the present paper is to make the mathematical results more easily readable, and to delineate the way they should be applied. By this reason, the mathematical details are kept at a minimum, even if we keep on providing the general threads that underlie the analysis.

Our starting point is a well-specified problem for a thin-walled beam with given dimensions and material properties, hereafter called the "real problem," which may represent a generic problem encountered in engineering applications. Since in the Γ-convergence approach one is led to study the limit of a sequence of problems, the paper aims at establishing a connection between the sequence of problems considered in the Γ-convergence analysis and the real problem. We follow ideas put forth in [11] and construct a sequence of problems consistent with the one used in [2, 3] by starting from the real problem. Clearly, there is some degree of arbitrariness in the construction of the sequence, and for the same real problem different choices can produce different asymptotic models, each of which may have a particular mechanical interest. Here, we consider the sequence obtained by "rescaling" the real problem in the most natural way and show that with such a choice we indeed obtain the Vlasov's model.

Because of various rescalings of the domains involved in the asymptotic analysis the Γ-limit has an integration domain different from the real domain. One of the central aims of the paper is to rewrite the asymptotic limit, with all its kinematical and mechanical quantities, over the real domain. In doing so, one discovers that a fastidious dependence of the limit problem on what we called the slenderness parameter drops. Furthermore, the analysis provides two asymptotic models that apply to what we called "thick" thin-walled beams and "regular" thin-walled beams, respectively, together with a quantitative criterion to decide which one to use in a specific problem. The paper also clarifies the reasons for regarding the asymptotic model as an approximation of the real problem. This is not a formal statement of convergence of course, because no error estimate is attempted, but it is at least a clue to believe that this is true.

The paper is organized as follows. In Sect. 2 we establish the real problem and anticipate our achievments. In Sect. 3, we outline the statement of the problem within the Γ-convergence scheme and summarize, without proofs, the outcome of our analysis. We also briefly discuss a simple instance of loads to which the conclusions apply. In Sect. 4 we discuss the dependence of the asymptotic analysis on the chosen reference domain and, on this basis, in Sect. 5 we construct a sequence of problems suitable to define an approximation of the real problem. In Sect. 6 the limit problem is rewritten on the real domain and in Sect. 7 we make explicit how it approximates the real problem. Finally, in Sect. 8 we show that the asymptotic model provides an approximation of the stresses in the real problem.

We conclude this introduction by fixing a bit of notation. Hereafter, we denote by (e_1, e_2, e_3) an orthonormal basis of $\mathbb{R}^3$, and identify $\mathbb{R}^2$ with the subspace $\mathbb{R}^2 \times \{0\}$ of $\mathbb{R}^3$ when necessary. We also distinguish between the domains occupied by the beam in the physical space and those in the space of coordinates used to parametrize them, denoting the former by a letter with a superposed circumflex accent. Likewise, we denote by the same letter, but with or without circumflex accent, fields that describe the same physical quantity, in the physical and the coordinate domain, respectively, e.g., $\hat{u}$ and u for the displacement field. The two are obviously correlated and the rules to pass from one to another are given in Sect. 3. Further information on the used notation is directly supplied in the text.

2 The Real Problem

We consider the equilibrium of a linearly elastic beam that in the natural configuration occupies the cylindrical region

$$\hat{\Omega}^r := \hat{\omega}^r \times (0, L),$$

where L is the length of the beam and the cross-section

$$\hat{\omega}^r := \left\{(\hat{x}_1^r, \hat{x}_2^r) = \gamma^r(x_1^r) + x_2^r n^r(x_1^r) : \quad x_1^r \in (0, \ell^r),\, x_2^r \in (-h^r/2, h^r/2)\right\} \qquad (1)$$

is the two-dimensional tube of thickness h^r generated by a smooth planar curve $\gamma^r : (0, \ell^r) \to \mathbb{R}^2$ of length ℓ^r whose tangent and normal unit vectors at $x_1^r \in (0, \ell^r)$ are, respectively, $t^r(x_1^r)$ and $n^r(x_1^r)$. It is assumed that

$$h^r \ll \ell^r \ll L,$$

so that the beam is slender: $\ell^r/L \ll 1$, and the walls are thin: $h^r/\ell^r \ll 1$.

We assume that the beam is made of an inhomogeneous linear hyper-elastic material with an elasticity tensor $\hat{\mathbb{C}}^r$ uniformly positive definite, that is, such that there exists $\hat{c}^r > 0$ for which the inequality

$$\hat{\mathbb{C}}^r(\hat{x}^r)E \cdot E \geq \hat{c}^r |E|^2$$

holds true for almost every $\hat{x}^r \in \hat{\Omega}^r$ and all symmetric matrices E. We also assume that the components of $\hat{\mathbb{C}}^r$ are essentially bounded functions over $\hat{\Omega}^r$, that is $\hat{\mathbb{C}}^r_{ijkl} \in L^\infty(\hat{\Omega}^r)$, while no assumption is made on the symmetry group of the material.

Finally, assuming that the beam is clamped at $\hat{x}^r_3 = 0$, we introduce the following function space for the admissible displacement fields

$$H^1_{dn}(\hat{\Omega}^r;\mathbb{R}^3) := \left\{\hat{u} \in H^1(\hat{\Omega}^r;\mathbb{R}^3) : \hat{u}_{|\hat{x}^r_3=0} = 0\right\},$$

where the subscript *dn* reminds that the admissible displacements $\hat{u}$ satisfy Dirichlet boundary conditions at one base and Neumann conditions at the other.

Then, under mild assumptions on the applied loads, it follows that the equilibrium problem amounts to finding the unique minimizer of the total energy functional $\hat{\mathscr{F}}^r : H^1_{dn}(\hat{\Omega}^r;\mathbb{R}^3) \to \mathbb{R}$ defined by

$$\hat{\mathscr{F}}^r(\hat{u}) := \frac{1}{2}\int_{\hat{\Omega}^r} \hat{\mathbb{C}}^r E\hat{u} \cdot E\hat{u}\, d\hat{x}^r - \hat{\mathscr{L}}_{\hat{\Omega}^r}(\hat{u}), \tag{2}$$

where

$$E\hat{u} := \mathrm{sym}(\nabla\hat{u}) := \frac{\nabla\hat{u} + \nabla\hat{u}^T}{2}$$

denotes the linearized strain tensor associated with the displacement field $\hat{u}$, and $\mathscr{L}^r_{\hat{\Omega}^r}(\hat{u})$ is the work done by the loads. The minimizer

$$\hat{u}^r = \underset{\hat{u}\in H^1_{dn}(\hat{\Omega}^r;\mathbb{R}^3)}{\mathrm{argmin}} \; \hat{\mathscr{F}}^r(\hat{u}). \tag{3}$$

is the solution to the "real" problem that we would like to approximate by means of the minimizers of suitable simpler problems.

Notice that this problem can be equivalently reformulated as a minimization problem

$$u^r = \underset{u\in H^1_{dn}(\Omega^r;\mathbb{R}^3)}{\mathrm{argmin}} \; \mathscr{F}^r(u) \tag{4}$$

for the functional

$$\mathscr{F}^r(u) := \hat{\mathscr{F}}^r(u \circ (\chi^r)^{-1})$$

over the admissible set

$$H^1_{dn}(\Omega^r;\mathbb{R}^3) := \left\{u \in H^1(\Omega^r;\mathbb{R}^3) : u_{x^r_3=0} = 0\right\}$$

of the displacement fields $u : \Omega^r \to \mathbb{R}^3$ defined on the reference slab

$$\Omega^r = (0, \ell^r) \times (0, h^r) \times (0, L),$$

with $\chi^r : \Omega^r \to \hat{\Omega}^r$ given by

$$\hat{x}^r = \chi^r(x^r) = \gamma^r(x_1^r) + x_2^r n^r(x_1^r) + x_3^r e_3 \in \hat{\Omega}^r. \tag{5}$$

On the basis of [2] and [11] hereafter we present two theories for the approximation of u^r. Both theories can be applied to thin-walled beams. The first theory is adequate to describe thin-walled beams with relatively thick walls, characterized by the condition $\ell^r/L \ll h^r/\ell^r$. The second theory is appropriate when dealing with "regular" thin-walled beams, characterized by the condition that h^r/ℓ^r and ℓ^r/L have the same order. Later on we'll give a criterion to decide which one of the two cases applies.

The theory offers two approximation formulas for the displacement field over Ω^r

$$u^r \approx u^{\min}, \qquad n^r \cdot H^r u^r t^r \approx \vartheta^{\min} \quad \text{where } H^r u^r := (\nabla \hat{u}^r) \circ \chi^r, \tag{6}$$

as well as an approximation for the stress field. In (6) the pair $(u^{\min}, \vartheta^{\min})$ is the minimizer of the effective energy

$$\mathscr{F}_0^r(u, \vartheta) := \int_{\Omega^r} f_{00}^r(x^r, \partial_3 \vartheta, \partial_3 u_3)\, dx^r - \mathscr{L}_0^r(u, \vartheta), \tag{7}$$

in an admissible set $\mathcal{A}^r(\mathfrak{s})$ that depends on a slenderness parameter $\mathfrak{s}$, with the choice to take $\mathfrak{s} = 0$ or $\mathfrak{s} \neq 0$ depending on whether the section is considered thick or regular. The effective strain energy density $f_{00}^r(x^r, \cdot, \cdot)$ is a quadratic function constructed by taking certain averages of

$$\mathbb{C}^r = \hat{\mathbb{C}}^r \circ \chi^r$$

over the transverse fiber $\{x_1^r\} \times (0, h^r) \times \{x_3^r\}$ passing through x^r, based on a recipe given in [2] and characterized in (55) later on. As is made explicit in what follows, the elements of $\mathcal{A}^r$ depend on functions of the variable x_3 only, so that the functional in (7) can be written as an integral over $(0, L)$ and the minimization problem is one-dimensional.

3 A Summary of the Γ-Convergence Results for Thin-Walled Beams

In this section we summarize the results achieved in [2]. We consider a sequence of thin-walled beams that in the reference configuration occupy the regions $\hat{\Omega}_\varepsilon$ described by

$$\hat{x} = \varepsilon \gamma(x_1) + \delta_\varepsilon x_2 n(x_1) + x_3 e_3 \tag{8}$$

with $x = (x_1, x_2, x_3)$ in the domain

$$\Omega = (0, \ell) \times (-h/2, h/2) \times (0, L).$$

Here, $\gamma = \gamma(x_1)$ is a simple open curve of length ℓ and arc length x_1, with trace contained in the plane x_1x_2, and sufficiently regular. We denote by $t := \partial_1\gamma$ the unit tangent and by $n := e_3 \wedge t$ the unit normal. Let $\kappa := \partial_1 t \cdot n$, with $\partial_1 = \frac{\partial}{\partial x_1}$, be the curvature of γ. Then, $\partial_1 t = \kappa n$ and $\partial_1 n = -\kappa t$. Equation (8) describes a cylinder whose cross-section is a tubular neighborhood of the curve $\varepsilon\gamma$ with wall thickness $\delta_\varepsilon h$.

It is assumed that

$$\lim_{\varepsilon\to 0} \frac{\delta_\varepsilon}{\varepsilon} = 0,$$

so that the walls are eventually thin, and that the quantity $\varepsilon^2/\delta_\varepsilon$ has a finite limit

$$\mathfrak{s} := \lim_{\varepsilon\to 0} \frac{\varepsilon^2}{\delta_\varepsilon}, \quad \mathfrak{s} \in [0, +\infty).$$

The value of this slenderness parameter characterizes in fact different asymptotic kinematics:

- the case $\mathfrak{s} \in (0, +\infty)$ corresponds to "regular" thin-walled beams,
- the case $\mathfrak{s} = 0$ corresponds to "thick" thin-walled beams.

These two cases were fully analyzed in [2], while the case of a beam with "ultra thin" walls, i.e., $\mathfrak{s} = +\infty$, is still open.

Equation (8) gives a representation of the domains $\hat{\Omega}_\varepsilon$ onto the ε-independent set of parameters Ω. By introducing the map

$$\chi_\varepsilon : x \in \Omega \mapsto \hat{x} \in \hat{\Omega}_\varepsilon \tag{9}$$

defined by (8), all the relevant fields of our problem can be represented as fields on Ω by composition. For instance, for the displacement field we write

$$u = \hat{u} \circ \chi_\varepsilon. \tag{10}$$

Similarly, the displacement gradient $\nabla\hat{u}$ and the linearized strain $E\hat{u} = \frac{1}{2}(\nabla\hat{u}+\nabla\hat{u}^T)$ are represented on Ω and denoted by

$$H^\varepsilon u := (\nabla\hat{u}) \circ \chi_\varepsilon \quad \text{and} \quad E^\varepsilon u := (E\hat{u}) \circ \chi_\varepsilon. \tag{11}$$

Through χ_ε, the variables (x_1, x_2, x_3) define a curvilinear system of coordinates on $\hat{\Omega}_\varepsilon$. It is convenient to introduce the (rescaled) local basis

$$g_1^\varepsilon := \frac{1}{\varepsilon}\frac{\partial \chi_\varepsilon}{\partial x_1} = (1 - \frac{\delta_\varepsilon}{\varepsilon} x_2 \kappa)t, \quad g_2^\varepsilon := \frac{1}{\delta_\varepsilon}\frac{\partial \chi_\varepsilon}{\partial x_2} = n, \quad g_3^\varepsilon := \frac{\partial \chi_\varepsilon}{\partial x_3} = e_3,$$

and its dual, in order to express the components of vectors and tensors defined on $\hat{\Omega}_\varepsilon$.

We consider the following sequence of energy functionals

$$\mathscr{F}_\varepsilon(u) := \frac{1}{2}\int_\Omega \mathbb{C}E^\varepsilon u \cdot E^\varepsilon u \sqrt{g^\varepsilon}\, dx - \mathscr{L}_\varepsilon(u) \tag{12}$$

where the fourth order tensor field $\mathbb{C}$ on Ω is essentially bounded and uniformly positive definite, and

$$\sqrt{g^\varepsilon} := g_1^\varepsilon \wedge g_2^\varepsilon \cdot g_3^\varepsilon = \frac{1}{\varepsilon\delta_\varepsilon} \det D\chi_\varepsilon = 1 - (\delta_\varepsilon/\varepsilon) x_2 \kappa.$$

For simplicity, we postpone for the moment to specify the sequence of loads $\mathscr{L}_\varepsilon$, because they play a minor role in our analysis.

As in the real problem, we assume that the beam is clamped at $x_3 = 0$ and denote by

$$H^1_{dn}(\Omega;\mathbb{R}^3) := \left\{u \in H^1(\Omega;\mathbb{R}^3) : u_{|x_3=0} = 0\right\}$$

the domain of $\mathscr{F}_\varepsilon$.

By means of (10) and (11) we may rewrite

$$\mathscr{F}_\varepsilon(u) = \frac{1}{\varepsilon\delta_\varepsilon}\Big(\frac{1}{2}\int_{\hat{\Omega}_\varepsilon} \mathbb{C}\circ\chi_\varepsilon^{-1} E\hat{u}\cdot E\hat{u}\, d\hat{x} - \hat{\mathscr{L}}_{\hat{\Omega}_\varepsilon}(\hat{u})\Big), \tag{13}$$

where

$$\hat{\mathscr{L}}_{\hat{\Omega}_\varepsilon}(\hat{u}) := \varepsilon\delta_\varepsilon \mathscr{L}_\varepsilon(u). \tag{14}$$

The expression within parentheses is the total energy of a beam, with elasticity tensor $\mathbb{C}\circ\chi_\varepsilon^{-1}$, that in the reference configuration occupies the region $\hat{\Omega}_\varepsilon$:

$$\hat{\mathscr{F}}_{\hat{\Omega}_\varepsilon}(\hat{u}) := \frac{1}{2}\int_{\hat{\Omega}_\varepsilon} \mathbb{C}\circ\chi_\varepsilon^{-1} E\hat{u}\cdot E\hat{u}\, d\hat{x} - \hat{\mathscr{L}}_{\hat{\Omega}_\varepsilon}(\hat{u}). \tag{15}$$

The results of [2] are based on certain compactness properties of the sequences of displacements u^ε with equi-bounded rescaled strain energy, or equivalently, since the elasticity tensor $\mathbb{C}(x)$ is uniformly positive definite, such that

$$\sup_{\varepsilon} \frac{1}{\delta_\varepsilon} \| E^\varepsilon u^\varepsilon \|_{L^2(\Omega)} < +\infty. \tag{16}$$

Namely, for given u^ε, let

$$\bar{v}^\varepsilon := \frac{u^\varepsilon - u_3^\varepsilon e_3}{\delta_\varepsilon / \varepsilon} \qquad \text{and} \qquad v_3^\varepsilon := \frac{u_3^\varepsilon}{\delta_\varepsilon}. \tag{17}$$

Then, up to a subsequence, the bound (16) implies that

$$\frac{E^\varepsilon u^\varepsilon}{\delta_\varepsilon} \rightharpoonup E \quad \text{in } L^2(\Omega; \mathbb{R}^{3\times 3})$$

for some $E \in L^2(\Omega; \mathbb{R}^{3\times 3})$. Furthermore, it is shown that

$$\bar{v}^\varepsilon \rightharpoonup \bar{v}, \qquad v_3^\varepsilon \rightharpoonup v_3, \qquad g_2^\varepsilon \cdot H^\varepsilon u^\varepsilon g_1^\varepsilon \rightharpoonup \vartheta \quad \text{in } H^1(\Omega; \mathbb{R}^3), \tag{18}$$

where the pair $(v = \bar{v} + v_3 e_3, \vartheta)$ belongs to the set

$$\begin{aligned} \mathcal{A}_{\mathfrak{s}} := \{ (v, \vartheta) \in H^1(\Omega; \mathbb{R}^3) \times H^{1+\hat{\mathfrak{s}}}_{dn}(0, L) \, : \, \exists\, \bar{m} \in H^2_{dn}(0, L; \mathbb{R}^3), \, \exists\, m_3 \in H^1_{dn}(0, L) \\ \text{such that } v = \bar{v} + v_3 e_3, \ \text{where } \bar{v} = \bar{m} + \mathfrak{s} \vartheta e_3 \wedge (\gamma - \gamma_G), \\ \text{and } v_3 = m_3 - \partial_3 \bar{m} \cdot (\gamma - \gamma_G) + \mathfrak{s} \partial_3 \vartheta \textstyle\int_0^{x_1} (\gamma - \gamma_G) \cdot n \, ds \}, \end{aligned} \tag{19}$$

with the slenderness parameter $\mathfrak{s} \in [0, +\infty)$ and

$$\hat{\mathfrak{s}} = \begin{cases} 0 & \text{if } \mathfrak{s} = 0, \\ 1 & \text{if } \mathfrak{s} \neq 0. \end{cases}$$

Here, γ_G denotes the center of mass of the curve γ.

If the strain tensor field is represented in the local basis $(t(x_1), n(x_1), e_3)$, it is also shown that the following partial characterization of the limit strain

$$E_{11} = \eta_1 + x_2 \eta_3, \quad E_{13} = -x_2 \partial_3 \vartheta + \eta_2, \quad E_{33} = \partial_3 v_3, \tag{20}$$

holds true, with $\eta_i = \eta_i(x_1, x_3)$ scalar functions in $L^2((0, \ell) \times (0, L))$.

By using the same basis, the elastic energy density of the beam

$$f(x, M) := \frac{1}{2} \mathbb{C}(x) M \cdot M, \tag{21}$$

can be written as a quadratic form of the components of the strain tensor M. Therefore, by denoting with different letters the components on which we have no information, in [2] we first introduced a function f_0 defined by

$$f_0(x, M_{11}, M_{13}, M_{33}) := \min_{A_{ij}} f(x, M_{11}, A_{12}, M_{13}, A_{22}, A_{23}, M_{33}). \tag{22}$$

The minimization provides the optimal A_{ij} as linear functions of (M_{11}, M_{13}, M_{33}). Then, by substituting such functions into $f(x,\cdot)$, f_0 takes the form

$$f_0(x, M_{11}, M_{13}, M_{33}) = \frac{1}{2}\begin{pmatrix} \mathbb{c}_{11}(x) & \mathbb{c}_{12}(x) & \mathbb{c}_{13}(x) \\ \mathbb{c}_{12}(x) & \mathbb{c}_{22}(x) & \mathbb{c}_{23}(x) \\ \mathbb{c}_{13}(x) & \mathbb{c}_{23}(x) & \mathbb{c}_{33}(x) \end{pmatrix}\begin{pmatrix} M_{11} \\ M_{13} \\ M_{33} \end{pmatrix} \cdot \begin{pmatrix} M_{11} \\ M_{13} \\ M_{33} \end{pmatrix},$$

where the reduced elasticity constants $\mathbb{c}_{ij}$ can be easily computed in terms of the original constants $\mathbb{C}_{ijhk}$ through (22).

Afterwards, we implemented the characterization (20) and studied the minimization of the strain energy with respect to the η_i in $L^2((0,\ell)\times(0,L))$

$$\min_{\eta_1,\eta_2,\eta_3} \int_{\Omega} f_0(x, \eta_1 + x_2\,\eta_3, -x_2\,\partial_3\vartheta + \eta_2, \partial_3 v_3)\,dx.$$

It turns out that the optimal η_i are found by solving the algebraic system of equations

$$\begin{pmatrix} \langle \mathbb{c}_{11}\rangle & \langle \mathbb{c}_{12}\rangle & \langle x_2\,\mathbb{c}_{11}\rangle \\ \langle \mathbb{c}_{12}\rangle & \langle \mathbb{c}_{22}\rangle & \langle x_2\,\mathbb{c}_{12}\rangle \\ \langle x_2\,\mathbb{c}_{11}\rangle & \langle x_2\,\mathbb{c}_{12}\rangle & \langle x_2^2\,\mathbb{c}_{11}\rangle \end{pmatrix}\begin{pmatrix} \eta_1^{\text{opt}} \\ \eta_2^{\text{opt}} \\ \eta_3^{\text{opt}} \end{pmatrix} = \begin{pmatrix} \langle x_2\,\mathbb{c}_{12}\rangle a - \langle \mathbb{c}_{13}\rangle b \\ \langle x_2\,\mathbb{c}_{22}\rangle a - \langle \mathbb{c}_{23}\rangle b \\ \langle x_2^2\,\mathbb{c}_{12}\rangle a - \langle x_2\,\mathbb{c}_{13}\rangle b \end{pmatrix}, \tag{23}$$

where

$$\langle\cdot\rangle := \fint_{-h/2}^{h/2} \cdot\, dx_2$$

denotes the average over the x_2 variable and $a, b \in \mathbb{R}$. It follows that the minimum of the strain energy is given by

$$\int_{\Omega} f_0(x, \eta_1^{opt} + x_2\,\eta_3^{opt}, -x_2\,\partial_3\vartheta + \eta_2^{opt}, \partial_3 v_3)\,dx,$$

where the $\eta_i^{opt} = \eta_i^{opt}(x_1, x_3, a, b)$ are evaluated at $a = \partial_3\vartheta$ and $b = \partial_3 v_3$. Therefore, if we define the reduced strain energy density function $f_{00} : \Omega\times\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ by

$$f_{00}(x, a, b) := f_0(x, \eta_1^{\text{opt}} + x_2\eta_3^{\text{opt}}, -x_2 a + \eta_2^{\text{opt}}, b), \tag{24}$$

we have that

$$\min_{\eta_1,\eta_2,\eta_3} \int_{\Omega} f_0(x, \eta_1 + x_2\,\eta_3, -x_2\,\partial_3\vartheta + \eta_2, \partial_3 v_3)\,dx = \int_{\Omega} f_{00}(x, \partial_3\vartheta, \partial_3 v_3)\,dx.$$

The following Γ-convergence result is proved in [2, Theorems 7.1 and 7.2].

Theorem 3.1 ($\boldsymbol{\Gamma}$-Limit). *Let $\mathscr{L}_0$ be the continuous limit of $\frac{1}{\delta_\varepsilon^2}\mathscr{L}_\varepsilon$ with respect to the convergence* (18)*. Then, the Γ-limit of the sequence of functionals $\frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon$, with respect to the convergence* (18)*, is the functional $\mathscr{F}_0 : \mathcal{A}_{\mathfrak{s}} \to \mathbb{R}$ given by*

$$\mathscr{F}_0(v,\vartheta) := \int_\Omega f_{00}(x, \partial_3\vartheta, \partial_3 v_3)\, dx - \mathscr{L}_0(v,\vartheta). \tag{25}$$

With a very mild assumption on the sequence $\mathscr{L}_\varepsilon$, the following convergence result for the sequence of minimizers is the consequence of a variational property of Γ-convergence (see [1]).

Theorem 3.2 (Convergence of the Minimizers). *Let*

$$u^{\min,\varepsilon} := \underset{u\in H^1_{dn}(\Omega;\mathbb{R}^3)}{\operatorname{argmin}} \frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon(u),$$

and

$$(v^{\min}, \vartheta^{\min}) := \underset{(v,\vartheta)\in\mathcal{A}_{\mathfrak{s}}}{\operatorname{argmin}} \mathscr{F}_0(v,\vartheta).$$

Then,

i) $\dfrac{u^{\min,\varepsilon} - u_3^{\min,\varepsilon} e_3}{\delta_\varepsilon/\varepsilon} \rightharpoonup v_1^{\min} e_1 + v_2^{\min} e_2$ *in* $H^1_{dn}(\Omega;\mathbb{R}^3)$,

ii) $\dfrac{u_3^{\min,\varepsilon}}{\delta_\varepsilon} \rightharpoonup v_3^{\min}$ *in* $H^1_{dn}(\Omega)$,

iii) $g_2^\varepsilon \cdot H^\varepsilon u^{\min,\varepsilon} g_1^\varepsilon \rightharpoonup \vartheta^{\min}$ *in* $H^1(\Omega)$.

Furthermore, we have the convergence of the minima

$$\frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon(u^{\min,\varepsilon}) \to \mathscr{F}_0(v^{\min}, \vartheta^{\min}). \tag{26}$$

Loads play a minor role in the analysis of [2]. Nevertheless, to give an example of loadings for which the above theorems apply let us consider the class of loads specified below.

Let the work of the loads be given by

$$\mathscr{L}_\varepsilon(u^\varepsilon) = \int_\Omega b^\varepsilon \cdot u^\varepsilon \sqrt{g^\varepsilon}\, dx, \tag{27}$$

with

$$b_\gamma^\varepsilon = \delta_\varepsilon \varepsilon\, b_\gamma \quad \gamma = 1,2, \quad \text{and} \quad b_3^\varepsilon = \delta_\varepsilon\, b_3, \tag{28}$$

and $b_\gamma, b_3 \in L^2(\Omega)$. Indeed, by taking (17) into account we get

$$\mathscr{L}_\varepsilon(u^\varepsilon) = \delta_\varepsilon^2 \int_\Omega \big(\frac{\varepsilon}{\delta_\varepsilon}(b_1 u_1^\varepsilon + b_2 u_2^\varepsilon) + \frac{1}{\delta_\varepsilon} b_3 u_3^\varepsilon\big)\sqrt{g^\varepsilon}\, dx = \delta_\varepsilon^2 \int_\Omega b \cdot v^\varepsilon \sqrt{g^\varepsilon}\, dx.$$

Hence, by observing that $g^\varepsilon \to 1$ uniformly, the functionals $\frac{1}{\delta_\varepsilon^2}\mathscr{L}_\varepsilon(u^\varepsilon)$ continuously converge to

$$\mathscr{L}_0(v, \vartheta) = \int_\Omega b \cdot v\, dx, \tag{29}$$

with respect to the convergence (18).

The class of loads described by (28) is possibly less general than that one could conceive, but it seems adequate for many of the problems commonly encountered.

4 Changes of the Reference Domain

Following [11], in order for the Γ-convergence results to be applicable to the real problem the sequence of Sect. 3 is to be constructed so that: (1) the real problem be equivalent to some problem of the sequence; and, (2) that problem be a far ahead element of the sequence, so that it is reasonable to assume that its limit is a good approximation of the real problem. Clearly, the latter is a loose statement unless one discusses the error estimate explicitly, but this is not done here.

To accomplish the scope requires that we address two issues: first, how the choice of a reference domain influences the analysis; and, second, how the sequence is to be chosen for the given real problem.

We observe that there are equivalent ways to describe the same sequence of problems. Indeed, the maps

$$\chi_{\tilde{\varepsilon}}: \quad \hat{x} = \tilde{\varepsilon}\tilde{\gamma}(\tilde{x}_1) + \tilde{\delta}_{\tilde{\varepsilon}}\tilde{x}_2\tilde{n}(\tilde{x}_1) + \tilde{x}_3 e_3, \quad \tilde{x} \in \tilde{\Omega} := (0, \tilde{\ell}) \times (-\frac{\tilde{h}}{2}, \frac{\tilde{h}}{2}) \times (0, L),$$

describe the same sequence of domains in the physical space as χ_ε provided that

$$\begin{cases} \tilde{x}_1 = \beta x_1,\ \tilde{\ell} = \beta\ell,\ \tilde{\varepsilon} = \frac{1}{\beta}\varepsilon;\ \tilde{x}_2 = \alpha x_2,\ \tilde{h} = \alpha h,\ \tilde{\delta}_{\tilde{\varepsilon}} = \frac{1}{\alpha}\delta_\varepsilon;\ \tilde{x}_3 = x_3; \\ \tilde{\gamma}(\tilde{x}_1) = \beta\gamma(\frac{1}{\beta}\tilde{x}_1). \end{cases} \tag{30}$$

In particular, under (30), the slenderness parameter changes according to

$$\tilde{\mathfrak{s}} := \lim_{\tilde{\varepsilon}} \frac{\tilde{\varepsilon}^2}{\tilde{\delta}_{\tilde{\varepsilon}}} = \frac{\alpha}{\beta^2}\mathfrak{s}. \tag{31}$$

Let us denote by $\chi := \tilde{\chi}_{\tilde{\varepsilon}}^{-1} \circ \chi_\varepsilon$ the map from Ω to $\tilde{\Omega}$ defined by (30). Namely,

$$\chi : \quad x \to \tilde{x} = (\beta x_1, \alpha x_2, x_3). \tag{32}$$

We note that, with the above choice, we have that $\tilde{t}(\tilde{x}_1) = t(x_1)$, $\tilde{n}(\tilde{x}_1) = n(x_1)$ and $\tilde{\kappa}(\tilde{x}_1) = \frac{1}{\beta}\kappa(x_1)$. Furthermore, the local basis vectors coincide

$$\tilde{g}_i^{\tilde{\varepsilon}}(\tilde{x}) = g_i^\varepsilon(x).$$

It follows that the components of corresponding tensors, when represented as fields on the two domains, are equal. That is, e.g.,

$$(\tilde{H}^{\tilde{\varepsilon}}\tilde{u})_{ij} = (H^\varepsilon u)_{ij} \circ \chi^{-1}, \quad \text{with } u := \hat{u} \circ \chi_\varepsilon \text{ and } \tilde{u} := \hat{u} \circ \tilde{\chi}_{\tilde{\varepsilon}}.$$

and the same is also true for the strain components

$$(\tilde{E}^{\tilde{\varepsilon}}\tilde{u})_{ij} = (E^\varepsilon u)_{ij} \circ \chi^{-1}. \tag{33}$$

As in (17), for any given u^ε let us set

$$\begin{aligned}
\bar{\tilde{v}}^{\tilde{\varepsilon}} &:= \tfrac{\tilde{u}^\varepsilon - \tilde{u}_3^\varepsilon e_3}{\tilde{\delta}_{\tilde{\varepsilon}}/\tilde{\varepsilon}} = \tfrac{\alpha}{\beta}\tfrac{u^\varepsilon - u_3^\varepsilon e_3}{\delta_\varepsilon/\varepsilon} \circ \chi^{-1} = \tfrac{\alpha}{\beta}\bar{v}^\varepsilon \circ \chi^{-1},\\
\tilde{v}_3^{\tilde{\varepsilon}} &:= \tfrac{\tilde{u}_3^{\tilde{\varepsilon}}}{\tilde{\delta}_{\tilde{\varepsilon}}} = \alpha \tfrac{u_3^\varepsilon}{\delta_\varepsilon} \circ \chi^{-1} = \alpha v_3^\varepsilon \circ \chi^{-1}.
\end{aligned}$$

It follows that the limit functions of converging sequences $\{u^\varepsilon\}$ are related by

$$\bar{\tilde{v}} = \frac{\alpha}{\beta}\, \bar{v} \circ \chi^{-1} \quad \text{and} \quad \tilde{v}_3 = \alpha\, v_3 \circ \chi^{-1}.$$

Thus, the change of domain induces an isomorphism

$$\psi : (v, \vartheta) \mapsto (\tilde{v}, \tilde{\vartheta}), \qquad \tilde{\vartheta} = \vartheta \circ \chi^{-1}, \; \bar{\tilde{v}} = \frac{\alpha}{\beta}\bar{v} \circ \chi^{-1}, \; \tilde{v}_3 = \alpha v_3 \circ \chi^{-1},$$

between the limit sets $\mathcal{A}_{\mathfrak{s}}$ and $\tilde{\mathcal{A}}_{\tilde{\mathfrak{s}}}$, with $\tilde{\mathfrak{s}} = \frac{\alpha}{\beta^2}\mathfrak{s}$.

In order for the problems described by the two sequences to be the same, we have to assume that the elasticity tensor $\tilde{\mathbb{C}}$ associated with the new reference domain is given by

$$\tilde{\mathbb{C}} = \mathbb{C} \circ \chi^{-1}, \tag{34}$$

and also that the body forces are the same, e.g., that their densities per unit physical volume are related to each other by

$$\tilde{b}^{\tilde{\varepsilon}} = b^\varepsilon \circ \chi^{-1}. \tag{35}$$

Let us repeat the deduction of the reduced strain energy density $\tilde{f}_{00}$ for the reference domain $\tilde{\Omega}$ using the same notations as in Sect. 3.

We first note that, because of (33) and (34), the strain energy density is given by

$$\tilde{f}(\tilde{x}, M_{ij}) = \frac{1}{2}\mathbb{C} \circ \chi^{-1}(\tilde{x})M \cdot M = f(\chi^{-1}(\tilde{x}), M_{ij}), \tag{36}$$

that is, functions $\tilde{f}(\tilde{x}, \cdot)$ and $f(\chi^{-1}(\tilde{x}), \cdot)$ coincide. Accordingly, the minimization

$$\tilde{f}_0(\tilde{x}, M_{11}, M_{13}, M_{33}) := \min_{A_{ij}} \tilde{f}(\tilde{x}, M_{11}, A_{12}, M_{13}, A_{22}, A_{23}, M_{33}).$$

yields that

$$\tilde{f}_0(\tilde{x}, M_{11}, M_{13}, M_{33}) = f_0(\chi^{-1}(\tilde{x}), M_{11}, M_{13}, M_{33}). \tag{37}$$

Then, by the representation formula of $f_0(\chi^{-1}(\tilde{x}), M_{11}, M_{13}, M_{33})$ and (34), it follows that

$$\tilde{f}_0(\tilde{x}, M_{11}, M_{13}, M_{33}) = \frac{1}{2}\begin{pmatrix} \tilde{\mathfrak{C}}_{11}(\tilde{x}) & \tilde{\mathfrak{C}}_{12}(\tilde{x}) & \tilde{\mathfrak{C}}_{13}(\tilde{x}) \\ \tilde{\mathfrak{C}}_{12}(\tilde{x}) & \mathfrak{C}^r_{22}(\tilde{x}) & \tilde{\mathfrak{C}}_{23}(\tilde{x}) \\ \tilde{\mathfrak{C}}_{13}(\tilde{x}) & \tilde{\mathfrak{C}}_{23}(\tilde{x}) & \tilde{\mathfrak{C}}_{33}(\tilde{x}) \end{pmatrix}\begin{pmatrix} M_{11} \\ M_{13} \\ M_{33} \end{pmatrix} \cdot \begin{pmatrix} M_{11} \\ M_{13} \\ M_{33} \end{pmatrix},$$

where

$$\tilde{\mathfrak{C}}_{ij} = \mathfrak{C}_{ij} \circ \chi^{-1}. \tag{38}$$

As in (24), for $\tilde{a}, \tilde{b} \in \mathbb{R}$ the reduced energy density $\tilde{f}_{00} : \tilde{\Omega} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined by

$$\tilde{f}_{00}(\tilde{x}, \tilde{a}, \tilde{b}) := \tilde{f}_0(\tilde{x}, \tilde{\eta}_1^{\text{opt}} + \tilde{x}_2\tilde{\eta}_3^{\text{opt}}, -\tilde{x}_2\tilde{a} + \tilde{\eta}_2^{\text{opt}}, \tilde{b}), \tag{39}$$

where the $\tilde{\eta}_i^{\text{opt}}$ are the solutions of the system

$$\begin{pmatrix} \langle \tilde{\mathfrak{C}}_{11} \rangle & \langle \tilde{\mathfrak{C}}_{12} \rangle & \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{11} \rangle \\ \langle \tilde{\mathfrak{C}}_{12} \rangle & \langle \tilde{\mathfrak{C}}_{22} \rangle & \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{12} \rangle \\ \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{11} \rangle & \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{12} \rangle & \langle (\tilde{x}_2)^2 \tilde{\mathfrak{C}}_{11} \rangle \end{pmatrix}\begin{pmatrix} \tilde{\eta}_1^{\text{opt}} \\ \tilde{\eta}_2^{\text{opt}} \\ \tilde{\eta}_3^{\text{opt}} \end{pmatrix} = \begin{pmatrix} \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{12} \rangle \tilde{a} - \langle \tilde{\mathfrak{C}}_{13} \rangle \tilde{b} \\ \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{22} \rangle \tilde{a} - \langle \tilde{\mathfrak{C}}_{23} \rangle \tilde{b} \\ \langle (\tilde{x}_2)^2 \tilde{\mathfrak{C}}_{12} \rangle \tilde{a} - \langle \tilde{x}_2 \tilde{\mathfrak{C}}_{13} \rangle \tilde{b} \end{pmatrix}. \tag{40}$$

With an abuse of notation, in the above formula the average over the thickness of the wall in the reference domain $\tilde{\Omega}$ is still denoted by

$$\langle \cdot \rangle := \fint_{-\tilde{h}/2}^{\tilde{h}/2} \cdot \, d\tilde{x}_2.$$

System of Eq. (40) allows us to find the relationship between $\tilde{f}_{00}$ and f_{00}.

By taking into account the definition of χ and (38) we find

$$
\begin{aligned}
\langle\, \tilde{\mathbb{C}}_{\alpha\beta}\rangle &= \fint_{-\tilde{h}/2}^{\tilde{h}/2} \tilde{\mathbb{C}}_{\alpha\beta}\, d\tilde{x}_2 = \fint_{-h/2}^{h/2} \mathbb{C}_{\alpha\beta}\, dx_2 = \langle\, \mathbb{C}_{\alpha\beta}\rangle,\\
\langle \tilde{x}_2\, \tilde{\mathbb{C}}_{\alpha\beta}\rangle &= \fint_{-\tilde{h}/2}^{\tilde{h}/2} \tilde{x}_2\, \tilde{\mathbb{C}}_{\alpha\beta}\, d\tilde{x}_2 = \alpha \fint_{-h/2}^{h/2} x_2\, \mathbb{C}_{\alpha\beta}\, dx_2 = \alpha \langle x_2\, \mathbb{C}_{\alpha\beta}\rangle,\\
\langle (x_2^r)^2\, \tilde{\mathbb{C}}_{\alpha\beta}\rangle &= \fint_{-\tilde{h}/2}^{\tilde{h}/2} x_2^{r2}\, \tilde{\mathbb{C}}_{\alpha\beta}\, d\tilde{x}_2 = \alpha^2 \fint_{-h/2}^{h/2} x_2^2\, \mathbb{C}_{\alpha\beta}\, dx_2 = \alpha^2 \langle x_2^2\, \mathbb{C}_{\alpha\beta}\rangle,
\end{aligned}
$$

and hence system (40) becomes

$$
\begin{pmatrix}
\langle\, \mathbb{C}_{11}\rangle & \langle\, \mathbb{C}_{12}\rangle & \alpha\langle x_2\, \mathbb{C}_{11}\rangle\\
\langle\, \mathbb{C}_{12}\rangle & \langle\, \mathbb{C}_{22}\rangle & \alpha\langle x_2\, \mathbb{C}_{12}\rangle\\
\alpha\langle x_2\, \mathbb{C}_{11}\rangle & \alpha\langle x_2\, \mathbb{C}_{12}\rangle & \alpha^2\langle x_2^2\, \mathbb{C}_{11}\rangle
\end{pmatrix}
\begin{pmatrix}
\tilde{\eta}_1^{\mathrm{opt}}\\ \tilde{\eta}_2^{\mathrm{opt}}\\ \tilde{\eta}_3^{\mathrm{opt}}
\end{pmatrix} =
$$

$$
= \begin{pmatrix}
\alpha\langle x_2\, \mathbb{C}_{12}\rangle\tilde{a} - \langle\, \mathbb{C}_{13}\rangle\tilde{b}\\
\alpha\langle x_2\, \mathbb{C}_{22}\rangle\tilde{a} - \langle\, \mathbb{C}_{23}\rangle\tilde{b}\\
\alpha^2\langle x_2^2\, \mathbb{C}_{12}\rangle\tilde{a} - \alpha\langle x_2\, \mathbb{C}_{13}\rangle\tilde{b}
\end{pmatrix},
$$

which can be equivalently written as

$$
\begin{pmatrix}
\langle\, \mathbb{C}_{11}\rangle & \langle\, \mathbb{C}_{12}\rangle & \langle x_2\, \mathbb{C}_{11}\rangle\\
\langle\, \mathbb{C}_{12}\rangle & \langle\, \mathbb{C}_{22}\rangle & \langle x_2\, \mathbb{C}_{12}\rangle\\
\langle x_2\, \mathbb{C}_{11}\rangle & \langle x_2\, \mathbb{C}_{12}\rangle & \langle x_2^2\, \mathbb{C}_{11}\rangle
\end{pmatrix}
\begin{pmatrix}
\tilde{\eta}_1^{\mathrm{opt}}\\ \tilde{\eta}_2^{\mathrm{opt}}\\ \alpha\tilde{\eta}_3^{\mathrm{opt}}
\end{pmatrix} =
\begin{pmatrix}
\langle x_2\, \mathbb{C}_{12}\rangle\alpha\tilde{a} - \langle\, \mathbb{C}_{13}\rangle\tilde{b}\\
\langle x_2\, \mathbb{C}_{22}\rangle\alpha\tilde{a} - \langle\, \mathbb{C}_{23}\rangle\tilde{b}\\
\langle x_2^2\, \mathbb{C}_{12}\rangle\alpha\tilde{a} - \langle x_2\, \mathbb{C}_{13}\rangle\tilde{b}
\end{pmatrix}. \tag{41}
$$

By comparing (41) and (23) it follows that

$$
\tilde{\eta}_1^{\mathrm{opt}} = \eta_1^{\mathrm{opt}}(x_1, x_2, \alpha\tilde{a}, \tilde{b}), \quad \tilde{\eta}_2^{\mathrm{opt}} = \eta_2^{\mathrm{opt}}(x_1, x_2, \alpha\tilde{a}, \tilde{b}),
$$

$$
\tilde{\eta}_3^{r,opt} = \tfrac{1}{\alpha}\eta_3^{\mathrm{opt}}(x_1, x_2, \alpha\tilde{a}, \tilde{b}).
$$

These identities, the definitions (39), (30) and (24), the identity of the functions $\tilde{f}_0(\tilde{x}, \cdot)$ and $f_0(\chi^{-1}(\tilde{x}), \cdot)$ imply that

$$
\begin{aligned}
\tilde{f}_{00}(\tilde{x}, \tilde{a}, \tilde{b}) &= \tilde{f}_0(\tilde{x}, \tilde{\eta}_1^{\mathrm{opt}} + \tilde{x}_2\tilde{\eta}_3^{\mathrm{opt}}, -\tilde{x}_2\tilde{a} + \tilde{\eta}_2^{\mathrm{opt}}, \tilde{b})\\
&= \tilde{f}_0(\tilde{x}, \eta_1^{\mathrm{opt}} + x_2\eta_3^{\mathrm{opt}}, -x_2\alpha\tilde{a} + \eta_2^{\mathrm{opt}}, \tilde{b})\\
&= f_0(x, \eta_1^{\mathrm{opt}} + x_2\eta_3^{\mathrm{opt}}, -x_2\alpha\tilde{a} + \eta_2^{\mathrm{opt}}, \tilde{b})\\
&= f_{00}(x, \alpha\tilde{a}, \tilde{b}),\\
&= \alpha^2 f_{00}(x, \tilde{a}, \tilde{b}/\alpha),
\end{aligned}
$$

where in the last identity we have made use of the fact that $f_{00}(x, \cdot, \cdot)$ is a homogeneous function of degree 2 in the last two variables. Thus, for $\tilde{a} = a$ and $\tilde{b} = \alpha b$, we get

$$f_{00}(x, a, b) = \frac{1}{\alpha^2}\tilde{f}_{00}(\chi(x), a, \alpha b), \tag{42}$$

for every $a, b \in \mathbb{R}$ and for every $x \in \Omega$.

By integration and a change of variables, from (42) it follows that

$$\int_{\Omega} f_{00}(x, \partial_3\vartheta, \partial_3 v_3)\, dx = \frac{1}{\alpha^3\beta}\int_{\tilde{\Omega}} \tilde{f}_{00}(\tilde{x}, \partial_3\tilde{\vartheta}, \partial_3\tilde{v}_3)\, d\tilde{x}, \tag{43}$$

where we have taken into account that det $\frac{\partial x}{\partial \tilde{x}} = \frac{1}{\alpha\beta}$.

Likewise, if one assumes that the loads rescale according to (28), it can be shown that a similar relation holds true for the work done.

To see this, let us recall (27) and write, by a change of variables and taking (35) into account,

$$\mathscr{L}_{\varepsilon}(u) = \int_{\Omega} b^{\varepsilon} \cdot u \sqrt{g^{\varepsilon}}\, dx = \frac{1}{\alpha\beta}\int_{\tilde{\Omega}} \tilde{b}^{\tilde{\varepsilon}} \cdot \tilde{u} \sqrt{g^{\tilde{\varepsilon}}}\, d\tilde{x} = \frac{1}{\alpha\beta}\tilde{\mathscr{L}}_{\tilde{\varepsilon}}(\tilde{u}), \tag{44}$$

with

$$\tilde{\mathscr{L}}_{\tilde{\varepsilon}}(\tilde{u}) := \int_{\tilde{\Omega}} \tilde{b}^{\tilde{\varepsilon}} \cdot \tilde{u} \sqrt{g^{\tilde{\varepsilon}}}\, d\tilde{x}. \tag{45}$$

Hence, by using (28) and repeating the calculations thereafter, we get

$$\begin{aligned}\tilde{\mathscr{L}}_{\tilde{\varepsilon}}(\tilde{u}) &= \int_{\tilde{\Omega}} \left(\varepsilon\delta_{\varepsilon}\, b_{\gamma} \circ \chi^{-1}\tilde{u}_{\gamma} + \delta_{\varepsilon} b_3 \circ \chi^{-1}\tilde{u}_3\right) \sqrt{g^{\tilde{\varepsilon}}}\, d\tilde{x} \\ &= \int_{\tilde{\Omega}} \left(\tilde{\varepsilon}\tilde{\delta}_{\tilde{\varepsilon}}\, \tilde{b}_{\gamma}\tilde{u}_{\gamma} + \tilde{\delta}_{\tilde{\varepsilon}}\tilde{b}_3\tilde{u}_3\right) \sqrt{g^{\tilde{\varepsilon}}}\, d\tilde{x},\end{aligned}$$

where the sum over γ, with $\gamma = 1, 2$, is assumed and the components of $\tilde{b}$ are defined by

$$\tilde{b}_{\gamma} := \frac{\varepsilon\delta_{\varepsilon}}{\tilde{\varepsilon}\tilde{\delta}_{\tilde{\varepsilon}}}\, b_{\gamma} \circ \chi^{-1} = \alpha\beta\, b_{\gamma} \circ \chi^{-1} \quad \text{and} \quad \tilde{b}_3 := \frac{\delta\varepsilon}{\tilde{\delta}_{\tilde{\varepsilon}}} b_3 \circ \chi^{-1} = \alpha\, b_3 \circ \chi^{-1}.$$

It follows that

$$\tilde{\mathscr{L}}_{\tilde{\varepsilon}}(\tilde{u}) = \tilde{\delta}_{\tilde{\varepsilon}}^2 \int_{\tilde{\Omega}} \left(\tilde{b}_{\gamma}\frac{\alpha}{\beta}(\frac{\varepsilon}{\delta_{\varepsilon}}\tilde{u}_{\gamma}) + \tilde{b}_3\alpha(\frac{1}{\delta_{\varepsilon}}\tilde{u}_3)\right) \sqrt{g^{\tilde{\varepsilon}}}\, d\tilde{x} \tag{46}$$

Thus, by taking into account that $\frac{\tilde{\delta}^2_{\tilde{\varepsilon}}}{\delta^2_{\varepsilon}} = \frac{1}{\alpha^2}$, for any converging sequence of displacements $\{u^\varepsilon\}$ and the corresponding one $\{\tilde{u}^\varepsilon\}$ in (44) one calculates that

$$\frac{1}{\delta^2_\varepsilon}\mathscr{L}_\varepsilon(u^\varepsilon) = \frac{1}{\alpha^3\beta}\frac{1}{\tilde{\delta}^2_{\tilde{\varepsilon}}}\tilde{\mathscr{L}}_{\tilde{\varepsilon}}(\tilde{u}^{\tilde{\varepsilon}}).$$

By passing to the limit, we get

$$\mathscr{L}_0(v,\vartheta) = \frac{1}{\alpha^3\beta}\tilde{\mathscr{L}}_0(\tilde{v},\tilde{\vartheta}), \tag{47}$$

with

$$\tilde{\mathscr{L}}_0(\tilde{v},\tilde{\vartheta}) := \int_{\tilde{\Omega}} \tilde{b}\cdot\tilde{v}\,d\tilde{x}$$

and

$$\tilde{v}:\quad \tilde{v}_\gamma = \frac{\alpha}{\beta}v_\gamma\circ\chi^{-1},\ \tilde{v}_3 = \alpha v_3\circ\chi^{-1} \text{ and } \tilde{\vartheta} = \vartheta\circ\chi^{-1}.$$

From (43) and (47) it follows that

$$\begin{aligned}\mathscr{F}_0(v,\vartheta) &= \int_\Omega f_{00}(x,\partial_3\vartheta,\partial_3 v_3)\,dx - \mathscr{L}_0(v,\vartheta) \\ &= \frac{1}{\alpha^3\beta}\left(\int_{\tilde{\Omega}}\tilde{f}_{00}(\tilde{x},\partial_3\tilde{\vartheta},\partial_3\tilde{v}_3)\,d\tilde{x} - \tilde{\mathscr{L}}_0(\tilde{v},\tilde{\vartheta})\right) = \frac{1}{\alpha^3\beta}\tilde{\mathscr{F}}_0(\tilde{v},\tilde{\vartheta}),\end{aligned} \tag{48}$$

where $\tilde{\mathscr{F}}_0$ is the functional that appears within brackets.

Thus, in particular, if $(v^{\min},\vartheta^{\min})$ is the minimizer of $\mathscr{F}_0$, then the minimizer $(\tilde{v}^{\min},\tilde{\vartheta}^{\min})$ of $\tilde{\mathscr{F}}_0$ is given by

$$\tilde{v}^{\min}_\gamma = \frac{\alpha}{\beta}v^{\min}_\gamma\circ\chi^{-1},\ \tilde{v}^{\min}_3 = \alpha v^{\min}_3\circ\chi^{-1} \text{ and } \tilde{\vartheta}^{\min} = \vartheta^{\min}. \tag{49}$$

Since ϑ and $\tilde{\vartheta}$ depend upon the third coordinate only, and $\tilde{x}_3 = x_3$ under χ^{-1}, we simply write $\tilde{\vartheta} = \vartheta$ in the following.

Remark 4.1. Formula (48) shows that changing the reference domain, together with the assumptions (34) and (35), yields an isomorphism between the respective limit sets $\mathcal{A}_{\mathfrak{s}}$ and $\tilde{\mathcal{A}}_{\tilde{\mathfrak{s}}}$, with $\tilde{\mathfrak{s}} = \frac{\alpha}{\beta^2}\mathfrak{s}$, that maps the minimizers of the limit functionals $\mathscr{F}_0$ and $\tilde{\mathscr{F}}_0$ into one another. In other words, it shows that up to an isomorphism the Γ-lim does not depend on the choice of the reference domain. So, for instance, one may choose Ω to be the cube

$$\ell = h := L.$$

Notice that the sequence of domains Ω_ε keeps on describing thin-walled beams for small ε in virtue of the assumption that $\delta_\varepsilon/\varepsilon \to 0$.

It is also clear that it's the limit value of $\frac{\varepsilon^2}{\delta_\varepsilon}$ that matters for the Γ-lim. So, without loss of generality, for $\mathfrak{s} \neq 0$ one may assume a quadratic rescaling of the wall thickness with ε

$$\delta_\varepsilon = \frac{1}{\mathfrak{s}}\varepsilon^2.$$

We make these two choices in the following.

5 Embedding of the Real Problem

In order to apply the Γ-convergence results to the real problem, we have to ask that the sequence considered in Sect. 3 be such that

$$\hat{\Omega}_{\varepsilon^r} = \hat{\Omega}^r,$$
$$\hat{\mathscr{F}}_{\hat{\Omega}_{\varepsilon^r}} = \hat{\mathscr{F}}^r,$$

for some "small" ε^r, cf. [11].

To get $\hat{\Omega}_{\varepsilon^r} = \hat{\Omega}^r$, set

$$\begin{cases} x_1^r = \varepsilon^r x_1, \quad x_2^r = \delta_{\varepsilon^r} x_2, \quad x_3^r = x_3;\ \varepsilon^r := \ell^r/\ell, \quad \delta_{\varepsilon^r} := h^r/h, \\ \\ \gamma(x_1) := \frac{1}{\varepsilon^r}\gamma^r(\varepsilon^r x_1), \end{cases} \tag{50}$$

where $\gamma^r(x_1^r)$ is the middle line of the beam's cross-section in the real problem, ℓ^r its length, and h^r the thickness of the walls. Thus, by (1) and (8),

$$\hat{x}^r = \gamma^r(x_1^r) + x_2^r n^r(x_1^r) + x_3^r e_3 = \varepsilon^r \gamma(x_1) + \delta_{\varepsilon^r} x_2 n_(x_1) + x_3 e_3.$$

The first equality defines the map $\chi^r : x^r \in \Omega^r \mapsto \hat{x}^r \in \hat{\Omega}^r$, with

$$\Omega^r := (0, \ell^r) \times (-\frac{h^r}{2}, \frac{h^r}{2}) \times (0, L);$$

the second one, $\chi_{\varepsilon^r} : x \in \Omega \mapsto \hat{x}^r \in \hat{\Omega}^r$.

The diffeomorphism

$$\chi : \quad x \mapsto x^r = (\varepsilon^r x_1, \delta_{\varepsilon^r} x_2, x_3) \tag{51}$$

maps Ω onto Ω^r. We note that

$$\chi^r = \chi_{\varepsilon^r} \circ \chi^{-1}.$$

To get $\hat{\mathscr{F}}_{\hat{\Omega}_{\varepsilon^r}} = \hat{\mathscr{F}}^r$, let us define

$$\mathbb{C} := \hat{\mathbb{C}}^r \circ \chi_{\varepsilon^r}, \tag{52}$$

and, with reference to (28),

$$b^{\varepsilon^r} := \hat{b}^r \circ \chi_{\varepsilon^r}, \tag{53}$$

where $\hat{\mathbb{C}}^r$ and $\hat{b}^r$ are the elasticity tensor and the density of the body forces in the real problem. Then, from (12)–(15) it follows that

$$\begin{aligned}\hat{\mathscr{F}}_{\hat{\Omega}_{\varepsilon^r}}(u \circ \chi_{\varepsilon^r}^{-1}) &= \varepsilon^r \delta_{\varepsilon^r}\left(\frac{1}{2}\int_{\Omega} \mathbb{C}E^{\varepsilon^r}u \cdot E^{\varepsilon^r}u\sqrt{g^{\varepsilon^r}}\,dx - \mathscr{L}_{\varepsilon^r}(u)\right) \\ &= \frac{1}{2}\int_{\Omega^r} \mathbb{C} \circ \chi^{-1}E^{\varepsilon^r}u \circ \chi^{-1} \cdot E^{\varepsilon^r}u \circ \chi^{-1}\sqrt{g^{\varepsilon^r}}\,dx^r - \varepsilon^r\delta_{\varepsilon^r}\mathscr{L}_{\varepsilon^r}(u) \\ &= \frac{1}{2}\int_{\hat{\Omega}^r} \hat{\mathbb{C}}^r E\hat{u} \cdot E\hat{u}\,d\hat{x}^r - \hat{\mathscr{L}}_{\hat{\Omega}^r}(\hat{u}) = \hat{\mathscr{F}}^r(\hat{u}).\end{aligned} \tag{54}$$

6 Γ-Limit on Ω^r

Let us look at (50) as a change of the reference domain of the type described in Sect. 4, with

$$\mathbb{C}^r := \mathbb{C} \circ \chi^{-1} \quad \text{and} \quad b^r := b^{\varepsilon^r} \circ \chi^{-1}.$$

In particular, by comparing (30) and (50) we have that

$$\beta \equiv \varepsilon^r \quad \text{and} \quad \alpha \equiv \delta_{\varepsilon^r},$$

and, by using (34), (35), (52) and (53) and noting that $\chi_{\varepsilon^r} \circ \chi^{-1}(x^r) = \hat{x}^r$,

$$\mathbb{C}^r(x^r) = \hat{\mathbb{C}}^r(\hat{x}^r) \quad \text{and} \quad b^r(x^r) = \hat{b}^r(\hat{x}^r).$$

By applying the results of Sect. 4 with the due adjustments, the reduced strain energy density is calculated in accordance with definition (39) and, by (42), satisfies

$$f_{00}(x, a, b) = \frac{1}{\delta_{\varepsilon^r}^2} f_{00}^r(x^r, a^r, b^r), \quad \text{with} \quad a^r = a \text{ and } b^r = \delta_{\varepsilon^r} b. \tag{55}$$

Therefore, formula (48) becomes

$$\begin{aligned}\mathscr{F}_0(v,\vartheta) &= \int_\Omega f_{00}(x,\partial_3\vartheta,\partial_3 v_3)\,dx - \mathscr{L}_0(v,\vartheta) \\ &= \frac{1}{\delta_{\varepsilon^r}^3\varepsilon^r}\left(\int_{\Omega^r} f_{00}^r(x^r,\partial_3\vartheta,\partial_3 u_3)\,dx^r - \mathscr{L}_0^r(u,\vartheta)\right) = \frac{1}{\delta_{\varepsilon^r}^3\varepsilon^r}\mathscr{F}_0^r(u,\vartheta)\end{aligned} \tag{56}$$

where ϑ and u in the functional of the second line are functions on Ω^r and u is given by

$$\bar{u} = \frac{\delta_{\varepsilon^r}}{\varepsilon^r}\bar{v}\circ\chi^{-1} \qquad \text{and} \qquad u_3 = \delta_{\varepsilon^r} v_3\circ\chi^{-1}. \tag{57}$$

Recalling the definition of $\mathcal{A}_\mathfrak{s}$, cf. (19), and assuming that $\delta_\varepsilon = \frac{1}{\mathfrak{s}}\varepsilon^2$ when $\mathfrak{s}\neq 0$, as specified in Remark 4.1, calculations yield that the pair (u,ϑ) belongs to the set $\mathcal{A}^r$ given by

$$\bar{u} = \bar{\xi} + \begin{cases} 0 & \text{if } \mathfrak{s}_r = 0, \\ \vartheta e_3 \wedge (\gamma^r - \gamma_G^r) & \text{if } \mathfrak{s}_r \neq 0, \end{cases} \tag{58}$$

and

$$u_3 = \xi_3 - \partial_3\bar{\xi}\cdot(\gamma^r - \gamma_G^r) + \begin{cases} 0 & \text{if } \mathfrak{s}_r = 0, \\ \partial_3\vartheta \displaystyle\int_0^{x_1^r}(\gamma^r - \gamma_G^r)\cdot n^r\,ds & \text{if } \mathfrak{s}_r \neq 0, \end{cases} \tag{59}$$

where

$$\mathfrak{s} = \mathfrak{s}_r = \frac{\varepsilon^{r2}}{\delta_{\varepsilon^r}}.$$

In this formula $\bar{\xi}$, ξ_3 and ϑ are functions of x_3 only. Thus, the problem of finding the minimum of $\mathscr{F}_0^r(u,\vartheta)$ in $\mathcal{A}^r$, or equivalently of $\mathscr{F}_0(v,\vartheta)$ in $\mathcal{A}_\mathfrak{s}$, is one-dimensional.

In particular, if

$$(u^{\min},\vartheta^{\min}) := \underset{(u,\vartheta)\in\mathcal{A}^r}{\operatorname{argmin}}\mathscr{F}_0^r(u,\vartheta),$$

then, by (57),

$$\bar{u}^{\min} := \frac{\delta_{\varepsilon^r}}{\varepsilon^r}\bar{v}^{\min}\circ\chi^{-1} \qquad \text{and} \qquad u_3^{\min} := \delta_{\varepsilon^r} v_3^{\min}\circ\chi^{-1}, \tag{60}$$

where the bar denotes the component of the displacement in the plane of the cross-section and $(v^{\min},\vartheta^{\min})$ is the minimizer of $\mathscr{F}_0(v,\vartheta)$.

7 Asymptotic Approximation of the Real Problem

The application of Γ-convergence theory to a specific problem is based on the ansatz that the asymptotic model is close to the problem at hand. Hereafter, we illustrate a motivation for this belief in the case discussed above.

If $u^{\min,\varepsilon}$ are the minimizers of $\frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon$ (equivalently, of $\mathscr{F}_\varepsilon$), then

$$u^{\min,\varepsilon} \rightharpoonup v^{\min}, \qquad n \cdot H^\varepsilon u^{\min,\varepsilon} t \rightharpoonup \vartheta^{\min}$$

in the sense stated in Theorem 3.2. Then, from the fact that

$$\varepsilon^r = \frac{\ell^r}{L} \ll 1,$$

one is led to accept that

$$u^r \approx v^{\min} \circ \chi^{-1}, \qquad n^r \cdot H^r u^r t^r \approx \vartheta^{\min} \circ \chi^{-1},$$

where $u^r := u^{\min,\varepsilon^r} \circ \chi^{-1}$ is the minimizer of $\mathscr{F}^r$ (defined on Ω^r). By recalling the notion of convergence from Theorem 3.2, it follows that we may write

$$\frac{u^r - u_3^r e_3}{\delta_{\varepsilon^r}/\varepsilon^r} \approx \bar{v}^{\min} \circ \chi^{-1}, \quad \frac{u_3^r}{\delta_{\varepsilon^r}} \approx v_3^{\min} \circ \chi^{-1}, \quad n^r \cdot H^r u^r t^r \approx \vartheta^{\min} \circ \chi^{-1}.$$

Thus, by (60),

$$\begin{cases} \bar{u}^r := u^r - u_3^r e_3 \approx \dfrac{\delta_{\varepsilon^r}}{\varepsilon^r} \bar{v}^{\min} \circ \chi^{-1} = \bar{u}^{\min}, \\[2ex] u_3^r \approx \delta_{\varepsilon^r} v_3^{\min} \circ \chi^{-1} = u_3^{\min}, \\[2ex] \vartheta^r \approx \vartheta^{\min} \circ \chi^{-1}. \end{cases} \tag{61}$$

We recall that, with the notation introduced in Sect. 5, the minimizer of the physical problem is given by

$$\hat{u}^r(\hat{x}^r) = u^r(x^r).$$

Statement (61) legitimates us to look at the problem

$$\text{Find}(u^{\min}, \vartheta^{\min}) \quad \text{s.t.} \quad (u^{\min}, \vartheta^{\min}) = \underset{(u,\vartheta) \in \mathcal{A}^r}{\text{argmin}} \mathscr{F}_0^r(u, \vartheta)$$

in order to approximate the solution of the real problem. To do so, the theory provides two asymptotic models, defined by (58) and (59), depending on the value of $\mathfrak{s}_r$. With the choice: $h = \ell := L$ and $\delta_\varepsilon = \varepsilon^2/\mathfrak{s}$, cf. Remark 4.1, we have

$$\mathfrak{s}_r = \frac{(\varepsilon^r)^2}{\delta_{\varepsilon^r}} = \frac{(\ell^r)^2}{h^r L}.$$

Thus, as a rule of thumb, one might choose the model with $\mathfrak{s}_r = 0$ when $(\ell^r)^2/(h^r L) \simeq 0$, and the other one otherwise. The former corresponds to the case $\frac{\ell^r}{L} \ll \frac{h^r}{\ell^r}$, the latter to $\frac{\ell^r}{L} \sim \frac{h^r}{\ell^r}$.

Remark 7.1. As a final remark, we notice that by the convergence of the minima and the definition of $\mathscr{F}_\varepsilon$ we have that

$$\left(\frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon(u^{\min,\varepsilon}) =\right) \frac{1}{\delta_\varepsilon^3 \varepsilon}\mathscr{F}_{\hat{\Omega}_\varepsilon}(\hat{u}^{\min,\varepsilon}) \to \frac{1}{\delta_{\varepsilon^r}^3 \varepsilon^r}\mathscr{F}_0^r(u^{\min}, \vartheta^{\min}).$$

In the same spirit as above, for $\varepsilon = \varepsilon^r$ this yields

$$\left(\mathscr{F}_{\hat{\Omega}_\varepsilon^r}(\hat{u}^{\min,\varepsilon^r}) =\right) \mathscr{F}^r(u^r) \approx \mathscr{F}_0^r(u^{\min}, \vartheta^{\min}).$$

Functional $\mathscr{F}^r$ is the total energy of the real problem, while $\mathscr{F}_0^r$ is that of the asymptotic model. Therefore, the latter also provides an approximation of the energy.

8 Approximation of the Stress in the Real Problem

For given $a, b \in \mathbb{R}$ and $x \in \Omega$, let $E^{\mathrm{opt}}(a,b) = E^{\mathrm{opt}}(x,a,b) \in \mathbb{R}^{3\times 3}$ be the tensor field for which we have

$$f_{00}(x,a,b) = f(x, E^{\mathrm{opt}}(x,a,b)),$$

where f is defined in (21) and f_{00} in (24). In the sequel the dependence on x will be omitted to be consistent with the notation of [2]

Let

$$E^{\min,\varepsilon} := E^\varepsilon u^{\min,\varepsilon}$$

be the strain calculated at the minimum of $\mathscr{F}_\varepsilon$, and

$$E^{\mathrm{opt}} := E^{\mathrm{opt}}(\partial_3 \vartheta^{\min}, \partial_3 v_3^{\min})$$

the strain defined above and evaluated in $a = \partial_3 \vartheta^{\min}$ and $b = \partial_3 v_3^{\min}$, with $(v^{\min}, \vartheta^{\min})$ the unique minimizer of $\mathscr{F}_0$.

By replicating an argument used in [2, Theorem 7.2], see CLAIM 1 and 2, it can be proved that, as $\varepsilon \to 0$,

$$\frac{1}{\delta_\varepsilon} E^{\min,\varepsilon} \rightharpoonup E^{\text{opt}} \quad \text{in } L^2(\Omega;\mathbb{R}^{3\times 3}). \tag{62}$$

Indeed, the proof of that theorem implied strong convergence, but this fact was not explicitly stated. Here we briefly give a proof of it.

By the positiveness of $\mathbb{C}$, for some $c > 0$ we have that

$$\begin{aligned}
&c\|\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} - E^{\text{opt}}\|^2_{L^2(\Omega)} \leq \\
&\leq \frac{1}{2}\int_\Omega \mathbb{C}(x)(\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} - E^{\text{opt}}) \cdot (\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} - E^{\text{opt}})\sqrt{g^\varepsilon}\, dx \\
&= \frac{1}{2}\int_\Omega \mathbb{C}(x)\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} \cdot \frac{1}{\delta_\varepsilon}E^{\min,\varepsilon}\sqrt{g^\varepsilon}\, dx - \frac{1}{2}\int_\Omega \mathbb{C}(x)E^{\text{opt}} \cdot \frac{1}{\delta_\varepsilon}E^{\min,\varepsilon}\sqrt{g^\varepsilon}\, dx \\
&\quad + \frac{1}{2}\int_\Omega \mathbb{C}(x)E^{\text{opt}} \cdot (E^{\text{opt}} - \frac{1}{\delta_\varepsilon}E^{\min,\varepsilon})\sqrt{g^\varepsilon}\, dx.
\end{aligned}$$

Hence,

$$\begin{aligned}
c\|\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} - E^{\text{opt}}\|^2_{L^2(\Omega)} \leq &\left[\frac{1}{2}\int_\Omega \mathbb{C}(x)\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} \cdot \frac{1}{\delta_\varepsilon}E^{\min,\varepsilon}\sqrt{g^\varepsilon}\, dx - \frac{1}{\delta_\varepsilon^2}\mathscr{L}_\varepsilon(u^{\min,\varepsilon})\right] \\
&-\left[\frac{1}{2}\int_\Omega \mathbb{C}(x)E^{\text{opt}} \cdot \frac{1}{\delta_\varepsilon}E^{\min,\varepsilon}\sqrt{g^\varepsilon}\, dx - \frac{1}{\delta_\varepsilon^2}\mathscr{L}_\varepsilon(u^{\min,\varepsilon})\right] \\
&+\frac{1}{2}\int_\Omega \mathbb{C}(x)E^{\text{opt}} \cdot (E^{\text{opt}} - \frac{1}{\delta_\varepsilon}E^{\min,\varepsilon})\sqrt{g^\varepsilon}\, dx.
\end{aligned}$$

The first term in the square parentheses on the right-hand side is $\frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon(u^{\min,\varepsilon})$, while the second tends to $\mathscr{F}_0(v^{\min}, \vartheta^{\min})$ and the third to zero when $\varepsilon \to 0$, thanks to (62) and

$$\frac{1}{\delta_\varepsilon^2}\mathscr{L}_\varepsilon(u^{\min,\varepsilon}) \to \mathscr{L}_0(v^{\min}, \vartheta^{\min}).$$

Then, by passing to the limit on the two sides we get

$$c \limsup_{\varepsilon \to 0} \|\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} - E^{\text{opt}}\|^2_{L^2} \leq \limsup_{\varepsilon \to 0} \frac{1}{\delta_\varepsilon^2}\mathscr{F}_\varepsilon(u^{\min,\varepsilon}) - \mathscr{F}_0(v^{\min}, \vartheta^{\min}).$$

Hence, recalling (26), we conclude that

$$\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} \to E^{\text{opt}} \text{ in } L^2(\Omega;\mathbb{R}^{3\times 3}). \tag{63}$$

For given $a, b \in \mathbb{R}$ and $x^r \in \Omega^r$, let $E^{\mathrm{r,\,opt}}(a,b) = E^{\mathrm{r,\,opt}}(x^r, a, b) \in \mathbb{R}^{3\times 3}$ be the tensor for which we have

$$f_{00}^r(x^r, a, b) = f^r(x^r, E^{\mathrm{r,\,opt}}(x^r, a, b)),$$

with

$$f^r(x^r, M) := \frac{1}{2}\mathbb{C}^r(x^r)M \cdot M,$$

and f_{00}^r defined in accordance with (39). From the argument developed in Sect. 4 it is easily seen that, for $x = \chi^{-1}(x^r)$,

$$\begin{aligned} E^{r\,opt}(x^r, \partial_3\vartheta, \partial_3 u_3) &= E^{\mathrm{opt}}(x, \delta_{\varepsilon^r}\partial_3\vartheta, \partial_3 u_3 \circ \chi) \\ &= \delta_{\varepsilon^r}E^{\mathrm{opt}}(x, \partial_3\vartheta, \frac{1}{\delta_{\varepsilon^r}}\partial_3 u_3 \circ \chi) \\ &= \delta_{\varepsilon^r}E^{\mathrm{opt}}(x, \partial_3\vartheta, \partial_3 v_3), \end{aligned} \tag{64}$$

recalling that $v_3 = \frac{1}{\delta_{\varepsilon^r}} u_3 \circ \chi$, by (57). From equality (64) and (63) it follows that

$$\frac{1}{\delta_\varepsilon}E^{\min,\varepsilon} \to \frac{1}{\delta_{\varepsilon^r}}E^{r\,opt}(\partial_3\vartheta^{\min}, \partial_3 u_3^{\min}) \circ \chi \quad \text{in } L^2(\Omega; \mathbb{R}^{3\times 3}).$$

Thus, by the argument used in Sect. 7, we have that

$$\frac{1}{\delta_{\varepsilon^r}}E^{\min,\varepsilon} \approx \frac{1}{\delta_{\varepsilon^r}}E^{r\,opt}(\partial_3\vartheta^{\min}, \partial_3 u_3^{\min}),$$

where $E^{\min,\varepsilon^{\mathrm{r}}}$ is the strain at the minimizer of $\mathscr{F}^r$ and $E^{r\,opt}(\partial_3 u_3^{\min}, \partial_3\vartheta^{\min})$ that at the minimizer of $\mathscr{F}_0^r$. In the above expression we have regarded the terms on the two sides as fields on Ω^r. In particular we have

$$T^{r\varepsilon^r} := \mathbb{C}^r E^{\min,\varepsilon^r} \approx T_0^r := \mathbb{C}^r E^{r\,opt}(\partial_3\vartheta^{\min}, \partial_3 u_3^{\min}) \tag{65}$$

in the L^2 norm. Thus, we can estimate the stress in the real problem by looking at that of the asymptotic model.

We conclude by noticing that, since $T = \mathbb{C}E = \frac{\partial f}{\partial E}$, the expression on the right hand side of (65) yields that

$$T_{0ij}^r = \frac{\partial f}{\partial E_{ij}}_{\Big|E^{r\,opt}(\partial_3\vartheta^{\min}, \partial_3 u_3^{\min})}.$$

It follows that the stationarity conditions that provide the definition of function f_0^r, see (22) and (37), i.e., the requirement that the equalities

$$\frac{\partial f}{\partial E_{12}} = \frac{\partial f}{\partial E_{22}} = \frac{\partial f}{\partial E_{23}} = 0$$

hold true at $E = E^{r\,opt}(\partial_3 \vartheta^{\min}, \partial_3 u_3^{\min} \circ \chi^{-1})$, read as

$$T^r_{012} = T^r_{022} = T^r_{023} = 0,$$

which gives the mechanical meaning of the optimality conditions (22).

Acknowledgements CD gratefully acknowledges the support received from GNFM of the Italian Istituto Nazionale di Alta Matematica.

References

1. Dal Maso, G.: An Introduction to Γ-Convergence. Birkäuser, Boston (1993)
2. Davini, C., Freddi, L., Paroni, R.: Linear models for composite thin-walled beams by Γ-convergence. Part I: open cross-sections. SIAM J. Math. Anal. **46**(5), 3296–3331 (2014)
3. Davini, C., Freddi, L., Paroni, R.: Linear models for composite thin-walled beams by Γ-convergence. Part II: closed cross-sections. SIAM J. Math. Anal. **46**(5), 3332–3360 (2014)
4. Davoli, E.: Thin-walled beams with a cross-section of arbitrary geometry: derivation of linear theories starting from 3D nonlinear elasticity. Adv. Calc. Var. **6**, 33–91 (2013)
5. Freddi, L., Morassi, A., Paroni, R.: Thin-walled beams: the case of the rectangular cross-section. J. Elast. **76**, 45–66 (2004/2005)
6. Freddi, L., Morassi, A., Paroni, R.: Thin-walled beams: a derivation of Vlassov theory via Γ-convergence. J. Elast. **86**, 263–296 (2007)
7. Freddi, L., Murat, F., Paroni, R.: Anisotropic inhomogeneous rectangular thin-walled beams. SIAM J. Math. Anal. **40**, 1923–1951 (2008/2009)
8. Freddi, L., Mora, M.G., Paroni, R.: Nonlinear thin-walled beams with a rectangular cross-section—part I. Math. Models Methods Appl. Sci. **22**, 1150016, 34 (2012)
9. Freddi, L., Mora, M.G., Paroni, R.: Nonlinear thin-walled beams with a rectangular cross-section—part II. Math. Models Methods Appl. Sci. **23**, 743–775 (2013)
10. Kollbrunner, C.F., Basler, K.: Torsion in Structures. Springer, Berlin/Heidelberg (1969)
11. Paroni, R., Podio-Guidugli, P.: On variational dimension reduction in structure mechanics. J. Elast. **118**(1), 1–13 (2015)
12. Vlasov, V.Z.: Thin-Walled Elastic Beams. Israel Program for Scientific Translations, Jerusalem (1961)
13. Volovoi, V., Hodges, D., Cesnik, C., Popescu, B.: Assessment of beam modeling methods for rotor blade applications. Math. Comput. Model. **33**, 1099–1112 (2001)

The Variational Approach to Fracture: A Theoretical Model and Some Numerical Results

Gianpietro Del Piero

Abstract The evolution of the response of an elastic-plastic bar from the initial unstressed state to rupture is studied with a one-dimensional model based on incremental energy minimization. The model can reproduce both brittle and ductile fracture, as well as an intermediate fracture mode, called ductile–brittle, in which, due to an extreme localization of the plastic deformation, the bar suddenly breaks after a more or less protracted plastic regime. Numerical simulations obtained from the model's implementation are compared with the results of tensile tests on bars made of steel and of non-reinforced concrete. With an accurate choice of the analytical shapes of the plastic strain energy, not only the overall behavior, but also many details of the experimental response can be captured.

1 Introduction

This communication is an example of how some sophisticated aspects of material response can be captured with relatively simple mathematical tools. The energetic approach adopted here is based on the decomposition of the strain energy into the sum of elastic and inelastic parts. In the resulting model some peculiar aspects of plastic response, such as yield condition and elastic unloading, emerge as necessary conditions for an energy minimum. This renders the theory strictly related to plasticity. By consequence, other possible physical causes of fracture, for example, damage [8], are not properly described by the present model.

While many structural problems are reduced to the minimization of an energy functional under a given external load, the description of fracture requires the solution of *incremental problems*, which consist in minimizing the increment of the energy due to a given load increment, starting from a given initial equilibrium

G. Del Piero (✉)
Dipartimento di Ingegneria, Università di Ferrara, 44100 Ferrara, Italy

International Research Center M&MoCS, Cisterna di Latina, Italy
e-mail: dlpgpt@unife.it

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_7

configuration. Under a given load path, the overall response of the structure is then determined by solving a sequence of incremental problems, for each of which the initial configuration is provided by the solution of the preceding problem.

Two fundamental fracture modes are observed experimentally. In the first mode, the body deforms elastically and then suddenly breaks, without any premonitory sign. In the second mode, the initial elastic response is followed by a plastic regime of large deformations, in which the carrying capacity of the structure gradually reduces to zero. The two fracture modes are called *brittle* and *ductile*, respectively. There is also an intermediate mode, which may be called *ductile–brittle* [2, Lect. 1], in which a catastrophic fracture occurs in the regime of plastic softening, as a consequence of an extreme localization of the plastic deformation.

In the past, the study of plastic softening met with serious difficulties. In finite element solutions, the softening zone was usually represented by elements containing *fictitious cracks* [4]. Since the softening regime is intrinsically unstable [2, Lect. 5], the solutions obtained in this way were strongly dependent on the mesh size. This difficulty was overcome by the introduction of the *gradient plasticity* model [1], in which the energy contains a supplementary term depending on higher-order derivatives, whose effect is to stabilize the softening regime.

With this improved model, a more realistic representation of ductile and ductile–brittle fracture became possible. However, some points were not yet clear. Indeed, while for materials undergoing ductile fracture, such as concrete and other geomaterials, some satisfactory models were produced, in the study of materials undergoing brittle or ductile–brittle fracture, such as most metallic materials, some prejudice arose against the possibility of a re-enlargement of the plastic zone after localization. Rejecting this possibility as "nonphysical" [6] led to the "nonphysical" consequence of excluding ductile fracture altogether.

Though this problem was raised many years ago [4], the causes of the different fracture modes have not been adequately investigated. At my knowledge, a first step was made only recently in the paper [3], in which the fracture mode was related to the convexity–concavity properties of the derivative of the plastic energy density θ. A further step is made in the present communication, in which it is shown that the ductile–brittle fracture of metals may occur only if the second derivative θ'' becomes smaller than a critical value. This value is easily determined if fracture occurs before localization. If not, its determination is an open problem.

At the present stage of the theory, only the one-dimensional problem can be studied in detail. In Sect. 2 the form of the energy is specified, and the stationarity conditions are given in Sect. 3. In Sect. 4 the incremental problem is formulated, and a two-step minimization strategy is defined. In Sect. 5, the analysis of the boundary conditions leads to the recognition of two types of solutions, *full-size*, in which the plastic deformations spreads all over the bar, and *localized*, in which the plastic deformation concentrates on more restricted regions. For an incremental problem starting from a homogeneous configuration, both full-size and localized closed-form solutions are given in Sect. 6. Moreover, the stability analysis made in Sect. 7 shows that each type of solutions holds in a specific domain, determined by the initial data through a non-dimensional factor *kl*. For the slope of the response curve, an explicit dependence on the incremental energy minimizers is given in Sect. 8. In the homogeneous case, for which the minimizers are known, an explicit dependence on the initial data is deduced.

In Sect. 9, the evolution of the bar from the initial unstressed state is analyzed. Before the elastic limit, the deformation is elastic and homogeneous. When the elastic limit is attained, the subsequent evolution is governed by the factor kl mentioned above, and by a second non-dimensional factor Kl^2/α. They are the thresholds beyond which localization of plastic deformation and catastrophic fracture occur, respectively. The first factor varies during the evolution, while the second only depends on the initial data. Their interplay decides whether localization or fracture occurs at the onset of the plastic deformation or later, and which of the two occurs first. If fracture is preceded by localization it is impossible to follow the evolution with closed-form solutions, and in this case numerical procedures must be used.

The final Sect. 10 shows a comparison, taken from the papers [3] and [7], between some experimental response curves and the corresponding numerical simulations. The experiments show that the steel bars break according to the ductile–brittle fracture mode, while the bars made of non-reinforced concrete break according to the ductile mode. The simulations reproduce the experimental behavior with surprising accuracy, up to the smallest details, including the effects of the presence of stiffenings at the bar's ends and of notches of different dimension in the mid section.

This is due to the appropriate choice of the shape of the energy θ. At first sight, this may look like a sort of curve fitting. In reality, there are perspectives for rendering the proposed model predictive. Indeed, this would be the case if the correlations between the shape of the macroscopic energy θ and the microscopic structural properties of matter were known. But this is a problem which falls far beyond the domain of continuum mechanics, since it would require an interdisciplinary approach involving physics, chemistry, and materials science.

2 The One-Dimensional Equilibrium Problem

Consider a straight bar of length l, with constant cross section, free of external loads, and subjected to the axial displacements

$$u(0) = 0\,, \qquad u(l) = \beta l\,, \tag{1}$$

at the endpoints. The only kinematical variable is the axial displacement u, and its derivative u' is the measure of the axial deformation. The deformation is decomposed into the sum of an elastic and an inelastic part

$$u'(x) = \varepsilon(x) + \gamma(x)\,, \qquad x \in (0, l)\,, \tag{2}$$

and a strain energy of the form

$$E(\varepsilon,\gamma) = \int_0^l \big(w(\varepsilon(x)) + \theta(\gamma(x)) + \tfrac{1}{2}\alpha\gamma'^2(x)\big)\,dx\,, \tag{3}$$

is assumed. Here α is a positive material constant, and the elastic and inelastic energy densities w and θ are assumed to be C^2 functions such that

$$\begin{array}{lll} w(0)=0, & w'(0)=0, & w''(\varepsilon)\geq 0 \;\;\forall\varepsilon\in\mathbb{R}\,, \\ \theta(0)=0, & \lim\limits_{\gamma\to+\infty}\theta(\gamma)<+\infty, & \theta'(\gamma)>0 \;\;\forall\gamma\geq 0\,. \end{array} \tag{4}$$

The elastic part of the energy is supposed to be recoverable, and the inelastic part is supposed to be dissipative. The last term in (3) is non-local, since its value at x depends on the values taken by γ at the neighboring points of x, through the derivative $\gamma'(x)$. As we shall see below, this term is necessary to include in the model the softening response, which otherwise would be unstable.

In the absence of applied loads, the strain energy (3) is the total energy of the beam. Therefore, the total energy is a non-local function of the *configurations* (ε,γ) of the bar. We say that (ε,γ) is an *equilibrium configuration* if it is a stationary point for the energy, and that an equilibrium configuration is *stable* if it is a local energy minimizer.

A *stationary point* for the energy is a configuration (ε,γ) such that the *first variation* of E at (ε,γ)

$$\delta E(\varepsilon,\gamma,\delta\varepsilon,\delta\gamma) \;=\; \lim_{t\to 0^+}\frac{1}{t}\Big(E(\varepsilon+t\delta\varepsilon,\gamma+t\delta\gamma)-E(\varepsilon,\gamma)\Big)\,, \tag{5}$$

is non-negative for all admissible perturbations $(\delta\varepsilon,\delta\gamma)$. To be *admissible*, a perturbation must satisfy a boundary condition and a dissipation condition. The boundary condition requires that the perturbed configuration has the same length $l+\beta l$ of the unperturbed configuration

$$\int_0^l \delta u'(x)\,dx = \int_0^l (\delta\varepsilon(x)+\delta\gamma(x))\,dx = 0\,, \tag{6}$$

and the dissipation condition requires that

$$\delta\gamma(x)\geq 0 \qquad \forall x\in(0,l)\,. \tag{7}$$

That is, only perturbations which increase or leave unaltered the inelastic deformation are considered admissible.

3 Stationarity Conditions

Let us compute the energy of the perturbed configuration

$$E(\varepsilon + t\delta\varepsilon, \gamma + t\delta\gamma) = E(\varepsilon, \gamma) + t \int_0^l \big(w'(\varepsilon(x))\, \delta\varepsilon(x) + \theta'(\gamma(x))\, \delta\gamma(x) \\ + \alpha\, \gamma'(x)\, \delta\gamma'(x) \big)\, dx + o(t)\,. \tag{8}$$

After integration by parts and after replacing $\delta\varepsilon(x)$ with $(\delta u'(x) - \delta\gamma(x))$, the non-negativeness of the first variation (5) is expressed by

$$\int_0^l \big(w'(\varepsilon(x))\, \delta u'(x) + (\theta'(\gamma(x)) - w'(\varepsilon(x)) - \alpha\gamma''(x))\, \delta\gamma(x) \big)\, dx \\ + \Big[\alpha\gamma'(x)\, \delta\gamma(x) \Big]_0^l \geq 0\,. \tag{9}$$

In particular, for a purely elastic perturbation, $\delta\gamma = 0$, this condition reduces to the inequality

$$\int_0^l w'(\varepsilon(x))\, \delta u'(x)\, dx \geq 0\,. \tag{10}$$

Due to the boundary condition (6), this inequality is satisfied only if $w'(\varepsilon(x))$ is constant across the bar. We call σ this constant, which is the axial force in the bar. Moreover, w' being monotonic by assumption (4), $w'(\varepsilon(x))$ constant implies $\varepsilon(x)$ constant. So, a first consequence of stationarity is the almost trivial fact that in the absence of axial load the axial force in the bar is constant, and for a constant cross section the deformation is constant as well. Calling ε this constant, we have

$$\sigma = w'(\varepsilon), \qquad u'(x) = \varepsilon + \gamma(x)\,, \tag{11}$$

and, by integration over $(0, l)$,

$$\beta = \frac{1}{l} \int_0^l u'(x)\, dx = \varepsilon + \bar{\gamma}\,, \qquad \bar{\gamma} = \frac{1}{l} \int_0^l \gamma(x)\, dx\,. \tag{12}$$

By condition (6), for $w'(\varepsilon(x)) = \sigma$, the first term in the integrand function of (9) is zero. Then for arbitrary non-negative $\delta\gamma$ we have

$$\theta'(\gamma(x)) - \sigma - \alpha\gamma''(x) \geq 0 \tag{13}$$

at the interior points, and

$$\gamma'(0) \leq 0\,, \qquad \gamma'(l) \geq 0 \tag{14}$$

at the boundary. Inequality (13) is a *yield condition* of the form

$$\sigma \leq f(\gamma, x)\,, \qquad f(\gamma, x) = \theta'(\gamma(x)) - \alpha\gamma\,''(x)\,, \tag{15}$$

according to which the axial force cannot exceed the limit force $f(\gamma, x)$. This condition is non-local since, due to the presence of the derivative $\gamma''(x)$, the limit force at x depends on the values taken by γ at the neighboring points.

In conclusion, an *equilibrium configuration* is a pair (ε, γ) which satisfies the following stationarity conditions:

- the independence of ε on x,
- the boundary conditions (14),
- the yield condition (15).

In what follows, the prescribed elongation β of the bar will be called a *load*, and an equilibrium configuration (ε, γ) will be said to be *equilibrated with the load* $\beta = \varepsilon + \bar{\gamma}$.

The presence of a yield condition is the first typical aspect of plastic response which emerges from the energy minimization. Other aspects will emerge later. From here on, we call *plastic* the inelastic response defined by the decomposition (2).

4 The Incremental Problem

Let $\{t \mapsto \beta_t\,,\ t \geq 0\}$ be a load process, and let $(\varepsilon_0, \gamma_0)$ be a stable equilibrium configuration equilibrated with the initial load β_0. A *quasi-static evolution from* $(\varepsilon_0, \gamma_0)$ is a deformation process $t \mapsto (\varepsilon_t, \gamma_t)$ such that every configuration $(\varepsilon_t, \gamma_t)$ is a stable equilibrium configuration, equilibrated with the corresponding load β_t. The *incremental problem* consists in finding a stable equilibrium configuration equilibrated with the load $\beta_{\tau+t}$ starting from a given stable equilibrium configuration equilibrated with β_τ, in the limit for $t \to 0^+$. The solution of this problem is a basic step for determining a quasi-static evolution. Indeed, a quasi-static evolution can be approximated by a sequence of incremental problems on finite intervals (τ_k, τ_{k+1}), in which the solution of each problem provides the initial condition for the next one. For $\tau = 0$, consider the expansions

$$\beta_t = \beta_0 + t\dot{\beta} + o(t)\,, \quad \varepsilon_t = \varepsilon_0 + t\,\dot{\varepsilon} + o(t)\,, \quad \gamma_t(x) = \gamma_0(x) + t\dot{\gamma}(x) + o(t)\,, \tag{16}$$

where, by the dissipation condition (7) and by differentiation of $(12)_1$,

$$\dot{\gamma}(x) \geq 0\,, \quad \dot{\beta} = \dot{\varepsilon} + \bar{\dot{\gamma}}\,, \tag{17}$$

and, by the boundary conditions (14),

$$\dot{\gamma}\,'(0) \leq 0 \quad \text{if} \;\; \gamma_0'(0) = 0\,, \qquad \dot{\gamma}\,'(l) \geq 0 \quad \text{if} \;\; \gamma_0'(l) = 0\,. \tag{18}$$

Then the energy admits the expansion

$$E(\varepsilon_t,\gamma_t) = E(\varepsilon_0,\gamma_0) + t\,I(\dot\gamma) + \tfrac{1}{2}\,t^2 J(\dot\gamma) + o(t^2)\,, \tag{19}$$

with

$$\begin{aligned} I(\dot\gamma) &= \int_0^l \big(\theta'(\gamma_0(x))\,\dot\gamma(x) + w'(\varepsilon_0)(\dot\beta - \bar{\dot\gamma}) + \alpha\,\gamma_0'(x)\dot\gamma'(x)\big)\,dx\,,\\ J(\dot\gamma) &= \int_0^l \big(\theta''(\gamma_0(x))\,\dot\gamma^2(x) + w''(\varepsilon_0)(\dot\beta - \bar{\dot\gamma})^2 + \alpha\dot\gamma'^2(x)\big)\,dx\,. \end{aligned} \tag{20}$$

The solution of the incremental problem consists in minimizing the functional $E(\varepsilon_t,\gamma_t)$ for $t \to 0^+$, that is, neglecting the terms $o(t)$ in (19). In this way, the problem is reduced to the minimization of the linear functional $I(\dot\gamma)$. Using the definitions $(11)_1$ of σ and $(15)_2$ of f, after integration by parts this functional takes the form

$$I(\dot\gamma) = l\sigma_0\,\dot\beta + \int_0^l \big(f(\gamma_0,x) - \sigma_0\big)\dot\gamma(x)\,dx + \big[\alpha\gamma_0'(x)\,\dot\gamma(x)\big]_0^l\,. \tag{21}$$

The term $l\sigma_0\,\dot\beta$ is known, the integrand function is non-negative due to the yield condition (13) and to the dissipation condition $\dot\gamma \geq 0$, and the boundary terms are non-negative due to $\dot\gamma \geq 0$ and to the boundary conditions (14). Then the minimum is $l\sigma_0\,\dot\beta$, and is achieved when

$$(f(\gamma_0,x) - \sigma_0)\,\dot\gamma(x) = 0\,, \qquad \gamma_0'(0)\,\dot\gamma(0) = \gamma_0'(l)\,\dot\gamma(l) = 0\,. \tag{22}$$

Thus, the result of the minimization of $I(\dot\gamma)$ is that a non-zero plastic deformation rate $\dot\gamma(x)$ is allowed only at the interior points of $(0,l)$ at which the yield condition (15) is satisfied as an equality, and at the boundary points at which $\gamma_0'(x)$ is zero.

Clearly, these conditions are not sufficient to determine $\dot\gamma$. We then go further minimizing the second-order term $J(\dot\gamma)$, under the previous conditions $(17)_1$ and (18), plus the supplementary conditions (22) which force the first-order term $I(\dot\gamma)$ to keep its minimum value. It is convenient to make this minimization in two steps. Let us set

$$\dot\gamma(x) = \bar{\dot\gamma}\;y(x)\,, \tag{23}$$

and let us rewrite the functional in the form

$$J(\dot\gamma) = J(\bar{\dot\gamma}\;y) = lw''(\varepsilon_0)(\dot\beta - \bar{\dot\gamma})^2 + \bar{\dot\gamma}^{\,2} J_0(y)\,, \tag{24}$$

with

$$J_0(y) = \int_0^l \big(\theta''(\gamma_0(x))\,y^2(x) + \alpha y'^2(x)\big)\,dx\,. \tag{25}$$

Let y_{min} be a minimizer for J_0. Then, by (24), $J(\bar{\dot{\gamma}} y_{min}) \leq J(\bar{\dot{\gamma}} y)$, and the minimization of $J(\dot{\gamma})$ is reduced to the minimization of the quadratic function

$$J_1(\bar{\dot{\gamma}}) = lw''(\varepsilon_0)(\dot{\beta} - \bar{\dot{\gamma}})^2 + \bar{\dot{\gamma}}^{\,2} J_0(y_{min})\,. \tag{26}$$

A minimum for J_1 exists if and only if its second derivative is positive

$$lw''(\varepsilon_0) + J_0(y_{min}) > 0\,, \tag{27}$$

and in this case the minimum is attained at

$$\bar{\dot{\gamma}} \;=\; \frac{l\,w''(\varepsilon_0)}{lw''(\varepsilon_0) + J_0(y_{min})}\,\dot{\beta}\,. \tag{28}$$

In the presence of the dissipation constraint $\bar{\dot{\gamma}} \geq 0$, since $w''(\varepsilon_0)$ is positive by assumption (4), this is indeed a minimizer at *loading*, that is, for $\dot{\beta} \geq 0$. At *unloading*, $\dot{\beta} < 0$, the minimum of the constrained problem is attained at $\bar{\dot{\gamma}} = 0$. Then the solution of the constrained problem is

$$\bar{\dot{\gamma}}_{min} = \frac{l\,w''(\varepsilon_0)}{lw''(\varepsilon_0) + J_0(y_{min})}\,\dot{\beta}^{+}, \tag{29}$$

where $\dot{\beta}^{+} = \max\{0, \dot{\beta}\}$ is the positive part of $\dot{\beta}$. Thus, a second result of the incremental minimization is the property of *elastic response at unloading*: if $\dot{\beta} < 0$, then $\bar{\dot{\gamma}} = 0$, and therefore $\dot{\gamma}(x) = 0$, at all x.

5 Determination of the Stationarity Points

In the preceding section, the incremental problem has been reduced to the minimization of the functional $J_0(y)$ subject to the dissipation condition $(17)_1$ and to the normalization condition $\bar{y} = 1$ which follows from definition (23). To keep the first-order term $I(\dot{\gamma})$ to its minimum value, the conditions (22) must also be imposed. After the change of variable from $\dot{\gamma}$ to y, these conditions take the form

$$y(x) \geq 0\,, \quad \bar{y} = 1\,, \quad (f(\gamma_0, x) - \sigma_0)\, y(x) = 0\,, \quad \gamma_0'(0)\, y(0) = \gamma_0'(l)\, y(l) = 0\,. \tag{30}$$

The stationarity points of J_0 are obtained imposing the non-negativeness of the first variation

$$\begin{aligned} \delta J_0(y, \eta) &= 2 \int_0^l \big(\theta''(\gamma_0(x))\, y(x)\, \eta(x) + \alpha\, y'(x)\, \eta'(x)\big)\, dx \\ &= 2 \int_0^l \big(\theta''(\gamma_0(x))\, y(x) - \alpha\, y''(x)\big)\, \eta(x)\, dx + 2\Big[\alpha\, y'(x)\, \eta(x)\Big]_0^l \geq 0\,, \end{aligned} \tag{31}$$

for all perturbations η satisfying

$$y(x) + \eta(x) \geq 0\,, \quad \bar{\eta} = 0\,, \tag{32}$$

that is, such that the perturbed deformation $(y + \eta)$ satisfies the conditions $(30)_1$ and $(30)_2$.

We look for solutions $y \in C^1(0, l)$ whose support is an interval $(a, b) \subseteq (0, l)$. By $(30)_1$, $y(x)$ is positive in (a, b) and zero outside and, by $(30)_3$, $f(\gamma_0, x) = \sigma_0$ at all x in (a, b). Then there are perturbations η with support in (a, b) such as both η and $-\eta$ satisfy conditions (32). For such perturbations, from inequality (31) and condition $\bar{\eta} = 0$, it follows that

$$\theta''(\gamma_0(x))\, y(x) - \alpha\, y''(x) = c \qquad \forall\, x \in (a, b)\,, \tag{33}$$

with c a constant to be determined.

For the interval (a, b), we consider the three cases:

$$a = 0\,,\; b = l\,, \qquad a = 0\,,\; b < l\,, \qquad a > 0\,,\; b < l\,. \tag{34}$$

In the first case, $\eta(0)$ and $\eta(l)$ may have any sign. Then from the boundary term in (31) we get

$$y'(0) = y'(l) = 0\,. \tag{35}$$

In the second case, $y(x)$ is zero in (b, l). Then for $y \in C^1(0, l)$ both the function and its derivative must vanish at $x = b$

$$y'(0) = 0\,, \qquad y(b) = y'(b) = 0\,. \tag{36}$$

Similarly, in the third case we have

$$y(a) = y'(a) = 0\,, \qquad y(b) = y'(b) = 0\,. \tag{37}$$

Thus, all stationarity points are solutions of the differential Eq. (33) under the constraints

$$y(x) > 0 \quad \forall\, x \in (a, b)\,, \qquad \bar{y} = \frac{1}{l}\int_a^b y(x)\, dx = 1\,. \tag{38}$$

In the three cases (34), the boundary conditions are (35)–(37), respectively. Then we have *full-size solutions* in the first case and *localized solutions* in the two remaining cases.

6 The Homogeneous Case

In general, for the explicit determination of the stationarity points the use of numerical solution techniques is necessary. However, in the *homogeneous case*, that is, for $\theta''(\gamma_0(x))$ constant, a closed-form solution is possible. This case occurs if either θ is quadratic or the initial deformation γ_0 is homogeneous. Let us call θ''_0 the constant value of $\theta''(\gamma_0(x))$. The stability analysis will show that the full-size solution

$$y(x) = 1\,, \qquad x \in (0,l) \tag{39}$$

is stable for $\theta''_0 > 0$, but may be unstable for $\theta''_0 < 0$. Therefore, for $\theta''_0 < 0$ we consider all three cases (34). We set

$$k = \Big(\frac{-\theta''_0}{\alpha}\Big)^{1/2}. \tag{40}$$

and rewrite Eq. (33) in the form

$$y''(x) + k^2\,(y(x) - C) = 0\,, \qquad x \in (a,b)\,. \tag{41}$$

The solutions have the form

$$y(x) = A\ \sin\, k(x-a) + B\cos\, k(x-a) + C\,. \tag{42}$$

The boundary conditions $y'(a) = y'(b) = 0$, which are common to all solutions, imply $A = 0$ and the alternative between

$$B = 0\,, \qquad k\,(b-a) = n\pi\,, \quad n = 1, 2, \ldots \tag{43}$$

For $B = 0$ we have $y(x) = C$, and the homogeneous solution (39) is re-obtained. For $B \neq 0$ we get two types of localized conditions. For those of the first type, the condition $\bar{y} = 1$ and the remaining boundary condition $y(b) = 0$ yield

$$y(x) = \begin{cases} \dfrac{l}{b}(1 + \cos\, kx) & \text{if } x \in (0,b)\,, \\ \quad 0 & \text{if } x \in (b,l)\,, \end{cases} \qquad b = \frac{\pi}{k}\,, \tag{44}$$

and for those of the second type the further condition $y(a) = 0$ yields

$$y(x) = \begin{cases} \dfrac{l}{b-a}\big(1 - \cos k\,(x-a)\big) & \text{if } x \in (a,b)\,, \\ \quad 0 & \text{if } x \in (0,a) \cup (b,l)\,, \end{cases} \qquad b-a = \frac{2\pi}{k}\,, \tag{45}$$

Note that, since $(b-a) \leq l$, the solution (44) exists only for $kl \geq \pi$ and the solution (45) exists only for $kl \geq 2\,\pi$.

7 Stability Analysis

The quadratic functional J_0 admits the exact expansion

$$J_0(y+\eta) = J_0(y) + \delta J_0(y,\eta) + \tfrac{1}{2}\,\delta^2 J_0(y,\eta)\,, \tag{46}$$

with $\delta J_0(y,\eta)$ as in (31) and

$$\delta^2 J_0(y,\eta) = \int_0^l \big(\theta''(\gamma_0(x))\,\eta^2(x) + \alpha\,\eta'^2(x)\big)\,dx\,. \tag{47}$$

The first variation (31) is zero at all stationary points y, because at all such points Eq. (33) and all boundary conditions are satisfied. Therefore, a necessary and sufficient condition for a minimum at y, both local and global, is that the second variation (47) be non-negative for all η.

This condition is trivially satisfied if $\theta''(\gamma_0(x))$ is positive at all x. Then in this case all stationary points y are stable equilibrium configurations. If $\theta''(\gamma_0(x))$ is negative in some subinterval of $(0,l)$, for functions η with support in this subinterval the first term of the integrand function in (47) is negative. Then if α were zero all stationary points would be unstable. This shows the stabilizing role of the non-local term in the total energy (3).

In general, to test the positiveness of the functional (47) it is necessary to use some numerical technique. However, in the homogeneous case $\theta''(\gamma_0(x)) = \theta_0'' < 0$ an analytical procedure is possible. Indeed, in this case the functional has the form

$$J_0(\eta) = \alpha \int_0^l \big(\eta'^2(x) - k^2\eta^2(x)\big)\,dx\,, \tag{48}$$

with k as in (40). Its domain consists of functions η continuous over $(0,l)$, satisfying conditions (32), and with support contained in the support (a,b) of y. The last condition is necessary to keep the first-order term of the expansion (46) at its minimum value $\delta J_0(y,\eta) = 0$. Thanks to this condition, the non-negativeness condition $(32)_1$ on $(y+\eta)$ does not impose any restriction on the sign of $\eta(x)$, and therefore can be ignored.

By the Poincaré inequality, there is a positive constant k_p such that

$$\int_a^b \eta'^2(x)\,dx \;\geq\; k_p^2 \int_a^b \eta^2(x)\,dx\,, \tag{49}$$

for all η in the domain of definition of J_0. Then, a stationarity point y is stable if $k \leq k_p$ and unstable if $k > k_p$. The Poincaré constant k_p depends on the boundary conditions. For the full-size solution $y(x) = 1$, the domain of J_0 is restricted only by condition $(32)_2$, $\bar\eta = 0$. In this case the Poincaré constant is $k_p = \pi/l$, and therefore this solution is stable for $kl \leq \pi$.

For the localized solution (44), the domain of definition is further restricted by the continuity condition $\eta(b) = 0$. In this case the Poincaré constant is the solution of the equation $k_p\pi/b = \tan(k_p\pi/b)$, that is, $k_p \approx 1.43\,\pi/b$, and since $b = \pi/k$ by (44), we have $k_p \approx 1.43\,k$. Then this solution is stable for all kl for which it exists, that is, for all $kl \geq \pi$. Finally, for the localized solution (45), due to the further continuity condition $\eta(a) = 0$, the Poincaré constant is $k_p = 2\,\pi/(b-a)$, and since $(b-a) = 2\pi/k$ by (45), we have $k_p = k$. Therefore, this solution is stable for all kl for which it exists, that is, for all $kl \geq 2\pi$.

In conclusion, the stability analysis provides the following stable solutions:

- the full-size solution $y(x) = 1$ for $kl \leq \pi$,
- the localized solution (44) for $kl \geq \pi$,
- the localized solution (45) for $kl \geq 2\pi$.

Thus, whether a stable solution is full-size or localized depends on the non-dimensional product kl, which involves the geometrical and material data l, θ_0'', and α.

8 The Slope of the Response Curve

For the axial force, by differentiation of $(11)_1$ we have

$$\dot{\sigma} = w''(\varepsilon_0)(\dot{\beta} - \bar{\dot{\gamma}}_{min})\,. \tag{50}$$

Then by (29) the slope of the response curve $(\dot{\sigma}, \dot{\beta})$ is

$$\frac{\dot{\sigma}}{\dot{\beta}} = \begin{cases} w''(\varepsilon_0) & \text{if } \dot{\beta} < 0 \ (\textit{unloading})\,, \\ \dfrac{w''(\varepsilon_0)\, J_0(y_{min})}{l\, w''(\varepsilon_0) + J_0(y_{min})} & \text{if } \dot{\beta} \geq 0 \quad (\textit{loading})\,. \end{cases} \tag{51}$$

On the right-hand side, $w''(\varepsilon_0)$ is positive by assumption, and $(l\,w''(\varepsilon_0) + J_0(y_{min}))$ is positive by (27). Then the slope of the response curve is positive at unloading. At loading, it has the same sign of $J_0(y_{min})$. That is, we have a *hardening response* if $J_0(y_{min}) > 0$, and a *softening response* if $J_0(y_{min}) < 0$. The sign of $J_0(y_{min})$ depends on the initial deformation γ_0. If $\gamma_0(x)$ is positive for all x, from (25) we see that the minimum of J_0 is positive, and therefore the response is hardening. In any other case, there is no a priori estimate and the sign of $J_0(y_{min})$ is known only after the minimizer y_{min} has been determined.

In the homogeneous case $\theta''(\gamma_0(x)) = \theta_0''$, the minimizers are known. For $\theta_0'' < 0$, for the full-size minimizer $y(x) = 1$ and for the localized minimizers (44) and (45) we have

$$J_0(y_{min}) = l\,\theta_0''\,, \qquad J_0(y_{min}) = \frac{k\,l^2}{\pi}\,\theta_0''\,, \qquad J_0(y_{min}) = \frac{k\,l^2}{2\,\pi}\,\theta_0''\,, \tag{52}$$

respectively. Let us call l_i the length of the support of $\dot{\gamma}$, which in the three cases, is $l, \pi/k$, and $2\pi/k$. Then for the three minimizers we have the single expression

$$J_0(y_{min}) = \frac{l^2}{l_i}\,\theta_0'' \,. \tag{53}$$

The condition (27) for the existence of a minimum for J_1 then takes the form

$$w''(\varepsilon_0) + \frac{l}{l_i}\,\theta_0'' > 0\,, \tag{54}$$

the expression (29) of $\bar{\dot{\gamma}}_{min}$ becomes

$$\bar{\dot{\gamma}}_{min} = \frac{w''(\varepsilon_0)}{w''(\varepsilon_0)\,l_i/l + \theta_0''}\,\dot{\beta}^+ . \tag{55}$$

and the slope (51) of the response curve is

$$\frac{\dot{\sigma}}{\dot{\beta}} = \begin{cases} w''(\varepsilon_0) & \text{if } \dot{\beta} < 0 \ (\textit{unloading})\,, \\ \dfrac{w''(\varepsilon_0)\,\theta_0''}{w''(\varepsilon_0)\,l_i/l + \theta_0''} & \text{if } \dot{\beta} \geq 0 \quad (\textit{loading})\,. \end{cases} \tag{56}$$

Then we have a hardening response if $\theta_0'' > 0$, and a softening response if $\theta_0'' < 0$.

9 Quasi-Static Evolution Under Monotonic Loading

In the case of a softening response, $\theta''(0) < 0$, we wish to determine the quasi-static evolution from the initial configuration

$$\varepsilon_0 = 0\,, \quad \gamma_0(x) = 0\,, \tag{57}$$

under the monotonic load process

$$\beta_t = t\,\dot{\beta}\,, \qquad t \geq 0\,, \tag{58}$$

where $\dot{\beta}$ is a positive constant. The initial configuration is stress-free, $\sigma_0 = w'(0) = 0$. Then, by $(15)_2$, $f(0,x)$ is equal to $\theta'(0)$ for all x, and since $\theta'(0)$ is positive by assumption, σ_0 is strictly less than $f(0)$. By $(22)_1$ this implies $\dot{\gamma}(x) = 0$ for all x, and since $\gamma_0(x) = 0$ we have $\gamma_t(x) = 0$. Then we have an initial *elastic regime*, $\sigma_t = w'(\beta_t)$. This regime goes on as long as σ_t remains smaller than $\theta'(0)$. That is, until the load attains the critical value β_c at which

$$w'(\beta_c) = \theta'(0)\,. \tag{59}$$

The configuration $(\varepsilon, \gamma) = (\beta_c, 0)$ marks the *onset of the plastic regime*. To study this regime, we make the simplifying assumption that w is well approximated by a quadratic function. That is, we assume that

$$w''(\varepsilon_t) \approx K\,, \tag{60}$$

with K a positive constant. At the onset, according to condition (54) and to the stability conditions given in Sect. 7, a stable full-size solution exists only if

$$kl < \pi \qquad \text{and} \qquad K + \theta_0'' > 0\,, \tag{61}$$

and a localized solution exists only if

$$kl > \pi \qquad \text{and} \qquad K + \frac{kl}{\pi}\,\theta_0'' > 0\,. \tag{62}$$

Therefore, if

$$K + \theta_0'' < 0 \tag{63}$$

there are no solutions at the onset. This sudden loss of equilibrium at the very end of the elastic regime corresponds to brittle fracture. Therefore, *inequality* (63) *is a sufficient condition for brittle fracture*.

If $K + \theta_0'' > 0$, a plastic regime begins. The solution is full-size if $kl < \pi$ and localized if $kl > \pi$. For $kl > 2\pi$, both localized solutions (44) and (45) are possible. By (52), the latter has a higher energy. Therefore, it can be ignored in the present analysis. In the case of a localized solution, the results of the previous sections do not allow us to follow the subsequent evolution. On the contrary, this is possible for a full-size solution. In this case, we set

$$\gamma_t(x) = \gamma_t\,, \qquad \theta''(\gamma_t) = \theta_t''\,, \qquad k_t = \sqrt{\frac{-\theta_t''}{\alpha}}\,, \tag{64}$$

and we note that the function $t \mapsto \gamma_t$ is non-decreasing because of the dissipation condition $(17)_1$.

In the full-size plastic regime the minimizer of J_0 is still $y(x) = 1$, its minimum is $l\theta_t''$, and the minimizer of J_1 is given by (29) with $J_0(y_{min}) = l\theta_t''$. This regime goes on until either of the two following events occurs:

- $k_t l$ reaches the value π,
- $-\theta_t''$ reaches the value K.

In the first case we have localization. In the second case, inequality (54) becomes an equality, and the slope (56) of the response curve becomes $-\infty$. This corresponds to *ductile–brittle fracture*. By $(64)_3$, this occurs when k_t reaches the value $\sqrt{K/\alpha}$. Therefore, the regime of full-size plastic deformation ends with

- localization, if $K\,l^2/\alpha\ >\ \pi^2$,
- brittle fracture, if $K\,l^2/\alpha\ <\ \pi^2$.

Thus, whether or not brittle fracture occurs is decided by the non-dimensional number $K\,l^2/\alpha$, independently of the shape of the function θ.

To deal with evolutions in the regime of localized plastic deformation, in addition to the assumptions made in Sect. 2, for the plastic energy density θ we now assume that

$$\lim_{\gamma\to+\infty}\theta'(\gamma)=0\,,\qquad \lim_{\gamma\to+\infty}\theta''(\gamma)=0\,, \tag{65}$$

and that there is a $\gamma_{lim}>0$ such that

$$\theta''(\gamma)<0\qquad\forall\,\gamma>\gamma_{lim}. \tag{66}$$

In view of the assumed continuity of θ' and θ'', these assumptions imply that $\theta'(x)$ has a maximum θ'_{max} and that $\theta''(x)$ has a minimum θ''_{min} for x in $[0,+\infty)$. Then we set

$$k_{max}=\sqrt{\frac{-\theta''_{min}}{\alpha}}\;, \tag{67}$$

and we consider first the case

$$k_{max}\,l<\pi\;. \tag{68}$$

This inequality guarantees that after the onset there is a full-size plastic regime, and that in this regime localization cannot occur. Then we have ductile–brittle fracture if $Kl^2/\alpha<\pi^2$, and an endless full-size plastic regime otherwise.

Let us see what happens in this second case. By the yield condition (15), σ_t is bounded from above by θ'_{max}, and from the assumed quadratic form of the elastic energy density w we have

$$\theta'_{max}\ \geq\ \sigma_t\ =\ w'(\varepsilon_t)\ =\ K\,\varepsilon_t\ =\ K\,(\beta_t-\gamma_t)\,. \tag{69}$$

Then $\gamma_t\ \to\ +\infty$ when $\beta_t\ \to\ +\infty$. By consequence, $\theta'(\gamma_t)$ tends to zero by $(65)_1$, and therefore σ_t tends to zero by the yield condition (15). Thus, there is an unlimited growth of plastic deformation, accompanied by a gradual decay to zero of the axial force. This is the *ductile fracture mode* observed in tension tests of bars made of non-reinforced concrete and other amorphous materials. In conclusion, if $k_{max}\,l<\pi$ there is an initial full-size plastic regime which ends with ductile–brittle fracture if $Kl^2/\alpha<\pi^2$, and with ductile fracture otherwise.

If $k_{max}l\ >\ \pi$, at some time a localized plastic regime is established. During this regime the localization may become extreme, producing ductile–brittle fracture. The occurrence of this event cannot be predicted, but only verified by numerical simulation. If ductile–brittle fracture does not occur, from (69) we still have $\theta'(\gamma_t) \to 0$ and therefore $k_t \to 0$. Then $k_t l$ is smaller than π for sufficiently large t, that is, the plastic regime becomes full-size. Thus, depending on the shape of the function θ, there may be a sequence of alternating full-size and localized regimes, but eventually ductile fracture occurs, preceded by a regime of full-size plastic deformation.

10 Numerical Simulations

In this section the predictions of the theory are compared with some experimental results. The traction tests on steel bars reported below in Sects. 10.1 and 10.3 were made expressly for the papers [3] and [7], while the test on a concrete bar reported in Sect. 10.2 was taken from the literature. In the comparison, some difficulties arise in the representation of the boundary conditions. Indeed, in the solution (44) the plastic deformation is concentrated near the boundary, while in the experiments the plastic deformation is intentionally kept away from the boundary by increasing the cross-sectional area at the bar's ends. Therefore, it makes no sense to compare the experiments with the theoretical solution (44), which is the only stable solution for $\pi < kl < 2\pi$ and the solution of least energy for $kl > 2\pi$.

In the paper [3], solutions closer to the experimental situation were obtained replacing the natural boundary conditions $y'(0) = y'(l) = 0$ with the "Dirichlet" conditions

$$y(0) = y(l) = 0\,, \tag{70}$$

which prevent the formation of plastic zones near the bar's ends. In the theoretical solutions obtained in [3] imposing these boundary conditions, for $kl < \pi$ the plastic regime starts with a full-size solution which slightly differs from the homogeneous solution $y(x) = 1$ only very close to the boundary. The same solution extends to kl in $(\pi, 2\pi)$. For $kl > 2\,\pi$, the plastic regime starts with the localized solution (45).

The numerical simulations reported in Sects. 10.1 and 10.2 were obtained with the boundary conditions (70). A better agreement with the experimental conditions is obtained taking into account the variability of the cross section. This has been done in the paper [7], some results of which are discussed in Sect. 10.3.

10.1 Simulation of a Tensile Test on a Steel Bar

The dotted curve in Fig. 1 is the response curve of a tensile test on a steel bar of length $l = 200\,\text{mm}$, with a circular cross section of diameter 16 mm. In [3], this curve has been compared with the result of a numerical simulation, whose input data have been chosen in the following way.

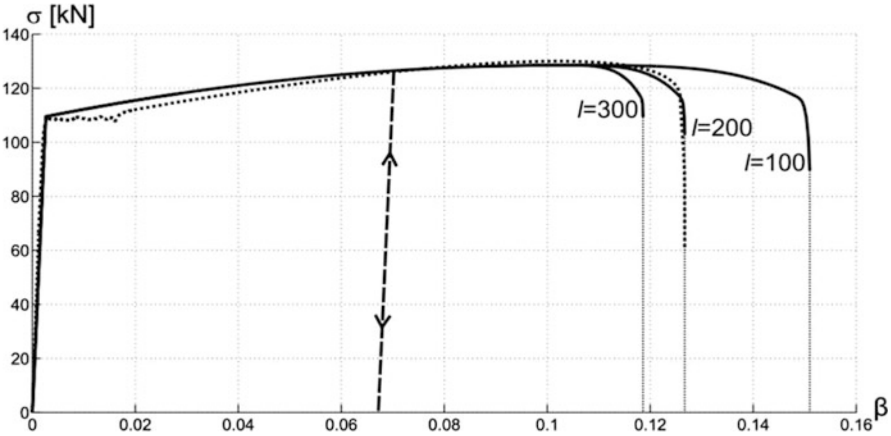


Fig. 1 The experimental response curve of a steel bar (*dotted line*), compared with the curves obtained by numerical simulation with different values of l. The *dashed line* shows the model's response to unloading and reloading

The elastic energy density w has been taken quadratic, and the axial stiffness $w''(\varepsilon) = K = 42 \times 10^3\,\mathrm{kN}$ is the product of the Young modulus of steel, $E = 210\,\mathrm{kN/mm}^2$, by the area $A = 201\,\mathrm{mm}^2$ of the cross section. For the nonlocality factor, the value $\alpha = 10\,\mathrm{kN\,mm}^2$ has been chosen on the basis of some rather informal microscopic considerations.

For the plastic energy density θ, a $C^1[0, +\infty)$ function with finite support $(0, \gamma_c)$ has been chosen. Inside the support, θ has been taken C^2 and piecewise cubic. Subdividing the support into four intervals and imposing the end conditions $\theta(0) = \theta'(\gamma_c) = 0$ and the continuity conditions for the function and for its first and second derivatives at the three internal nodal points, the function is described by five constants. One of them, $\theta'(0)$, which by (59) is the value of σ at the onset of the plastic regime, has been measured directly on the experimental curve. The remaining four have been chosen as follows.

The second derivative θ'' has been taken decreasing from a maximum of 380 kN at $\gamma = 0$ to a minimum $\theta''_{min} = -350\,\mathrm{kN}$ at $\gamma_c = 2.10$. In the first interval θ'' has been taken positive to reproduce the hardening branch of the experimental curve, and in the three following intervals it has been taken negative to reproduce the softening branch. As shown in Fig. 1, in spite of the rough way in which the data were selected, a quite good approximation of the experimental curve for the whole load process, from the unstressed state to final rupture, has been obtained. Indeed, the result of the numerical simulation, which is the solid line $l = 200$ in the figure, is quite close to the experimental curve. It is worth noting that this has been obtained using a very simple numerical program and a very small number of parameters: K, l, α, and the five constants which determine the function θ.

The numerical simulation does not reproduce the horizontal plateau exhibited by the experimental curve just after the onset of the plastic regime. A closer approximation can be obtained with a more accurate definition of the function θ, for example, subdividing its support into a larger number of intervals. This has been done in the simulation discussed in Sect. 10.3.

The two remaining solid lines in Fig. 1 show the results of numerical simulations on beams with lengths $l = 100$ and $l = 300$ mm. While the hardening branch of the curve is the same for all beams, the different softening branches show the presence of a *size effect*, by which longer bars break at lower values of β. In the same figure, the dashed line represents a loading–unloading cycle inserted in the simulation. The response perfectly reproduces the elastic unloading–reloading phenomenon.

The profiles of the plastic deformation obtained in the numerical simulation are shown in Fig. 2. They are in excellent agreement both with the experiments, see, e.g., [3, Fig. 5], and with other theoretical models, see, e.g., [6, Fig. 9]. Concerning the theoretical solutions of Sect. 6, the profiles in the hardening regime, Fig. 2a, are quite close to the theoretical solution $\gamma(x) = \bar{\gamma}$ except near the bar's ends, where a very local effect is due to the boundary conditions (70). The profiles in the softening regime, Fig. 2b, correspond to the solution (45). They show a progressive localization, which ends with ductile–brittle fracture. This type of fracture has been obtained choosing a sufficiently large value of $-\theta''_{min}$. The theoretical predictions made in the previous sections are confirmed. Indeed, at the onset of plastic deformation, $Kl^2/\alpha = 1.68 \times 10^8 \gg \pi^2$ excludes brittle fracture, and $k^2_{max}l^2 = 1.4 \times 10^6 \gg \pi^2$ implies that at some time a localization of the plastic deformation must occur.

10.2 Simulation of a Tensile Test on a Concrete Bar

The second simulation considered in [3] had the purpose of reproducing the response curve of a bar made of non-reinforced concrete. The dotted curve shown in Fig. 3, taken from [5, Fig. 38], is the response curve of a tensile test on a bar

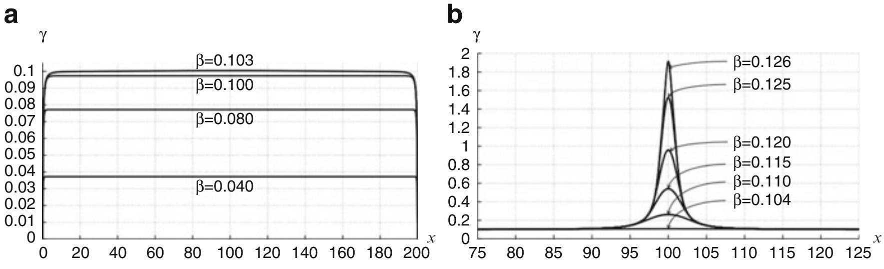


Fig. 2 Simulation of a tensile test on a steel bar. The plastic deformation profile in the hardening regime (**a**) and in the subsequent softening regime (**b**). Note the different scales in the two pictures

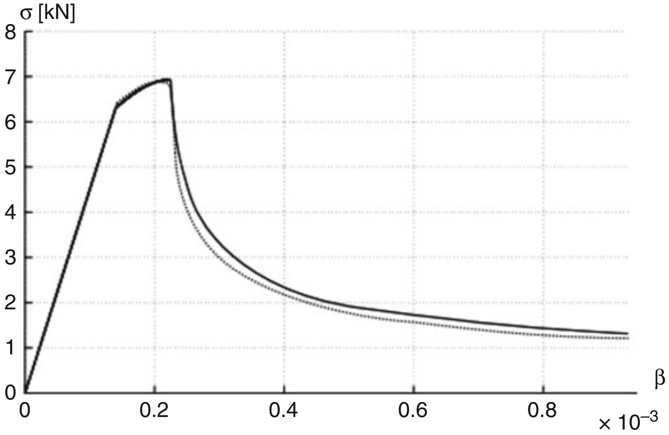


Fig. 3 The experimental response curve of a concrete bar (*dotted line*) compared with the curves obtained by numerical simulation

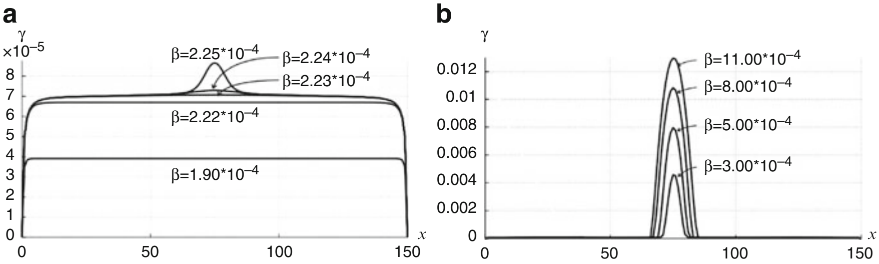


Fig. 4 Simulation of a tensile test on a concrete bar. The plastic deformation profile in the regime of progressive (**a**) and regressive (**b**) localization. Note the different scales in the two pictures

of length 150 mm and with square cross section of size 50 mm. The elastic energy density w is supposed to be quadratic, with $K = 45 \times 10^3$ kN, and the non-locality constant $\alpha = 3.5 \times 10^3$ kN mm^2 has been chosen with considerations related to the maximum estimated dimension of the grains.

Like in the case of the steel bar, the plastic energy density θ is supposed to be $C^1[0, +\infty)$, and C^2 and piecewise cubic in the support $(0, \gamma_c)$. The support has again been subdivided into four intervals, and θ'' has been supposed positive in the first interval and negative in the remaining ones, with a maximum of 18×10^3 kN at $\gamma = 0$ and a minimum of -2×10^3 kN at $\gamma = 1.2 \times 10^{-4}$.

At the onset of the plastic regime, the inequality $Kl^2/\alpha = 2.9 \times 10^5 \gg \pi^2$ guarantees that there is no brittle fracture, and $k^2_{max} l^2 = 1.29 \times 10^4 \gg \pi^2$ guarantees that at some time a localized plastic regime is established. However, in the present case the minimum of θ'' is not sufficiently large to produce ductile–brittle fracture. As shown in Fig. 4, the progressive localization started at $\beta = 2.22 \times 10^{-4}$ stops at $\beta = 3 \times 10^{-4}$, and for larger β the plastic zone enlarges. A further simulation, not

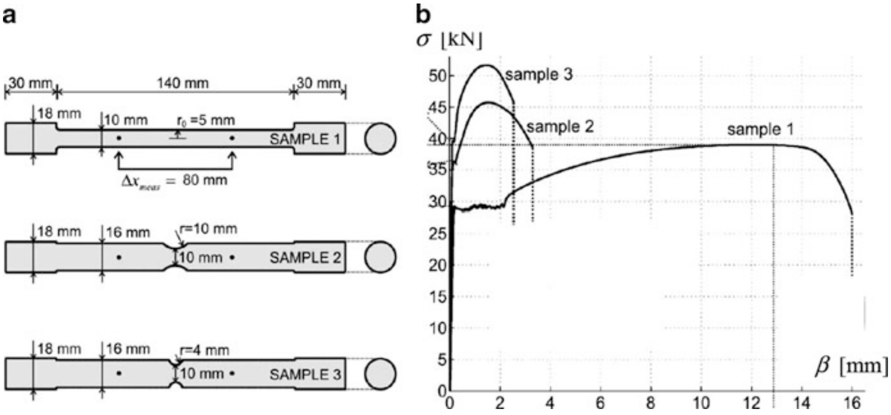


Fig. 5 Simulations of tensile tests on steel bars with variable cross section. The three samples (**a**), and the corresponding response curves (**b**)

reported here, shows that for even larger β the plastic zone spreads over the whole interval $(0, l)$, in accordance with the ductile fracture mode. This is the behavior shown in Fig. 6 of [6], where it is considered "spurious" and "nonphysical."

10.3 Simulations on Steel Bars with Variable Cross Section

An investigation on the effects of the variability of the cross section has been made in [7]. The three samples shown in Fig. 5a have been subjected to a tensile test, and the results have been compared with those of corresponding numerical simulations. In the latter, an energy of the form

$$E(\varepsilon,\gamma) = \int_0^l A(x)\big(w(\varepsilon(x)) + \theta(\gamma(x)) + \tfrac{1}{2}\,\alpha(\gamma(x))\,\gamma'^2(x)\big)\,dx\,, \tag{71}$$

has been assumed, with $A(x)$ the area of the cross section. A dependence of the nonlocality factor α on the plastic deformation has also been taken into account, but it will not be discussed here.

As shown in the figure, the first sample has a reinforced cross section at the ends, and the remaining two have circular notches of different radii at the mid section. All samples have the same length and the same cross-sectional area at the mid section. The response curves in Fig. 5b show that the last two samples have a larger strength. This is due to the larger cross-sectional area away from the mid section. By contrast, the presence of the notches considerably reduces the ductility, that is, the gap between the values of β at the onset of the plastic regime and at fracture. Also, the gap is larger for the notch with a smaller radius.

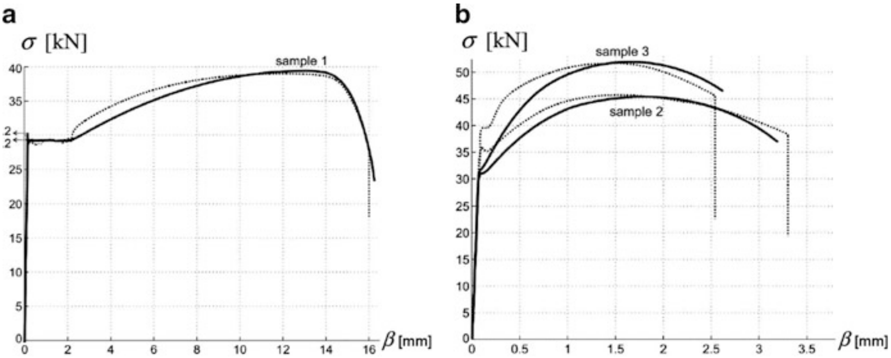


Fig. 6 Comparison between the response curves obtained in the experiments (*dotted lines*) and in the numerical simulations (*full lines*)

A comparison with the numerical simulations is shown in Fig. 6. The plateau observed at the beginning of the inelastic regime has been reproduced taking an energy density θ concave for small γ, then convex in correspondence to the hardening regime, and again concave for large γ, to reproduce the subsequent softening regime up to the final rupture. Though the overall reproduction of the evolution process is satisfactory, there are still quantitative discrepancies, whose elimination requires some further work.

References

1. Aifantis, E.C.: On the microstructural origin of certain inelastic models. ASME J. Eng. Mater. Technol. **106**, 326–330 (1984)
2. Del Piero, G.: A variational approach to fracture and other inelastic phenomena. J. Elast. **112**, 3–77 (2013)
3. Del Piero, G., Lancioni, G., March, R.: A diffuse energy approach for fracture and plasticity: the one-dimensional case. J. Mech. Mater. Struct. **8**, 109–151 (2013)
4. Hillerborg, A.: Application of the fictitious crack model to different types of materials. Int. J. Fract. **51**, 95–102 (1991)
5. Hordijk, D.A.: Tensile and tensile fatigue behaviour of concrete; experiments, modelling and analyses. Heron **37**, 1–79 (1992)
6. Jirásek, M., Rolshoven, S.: Localization properties of strain-softening gradient plasticity models. Part I: strain-gradient theories. Int. J. Solids Struct. **46**, 2225–2238 (2009)
7. Lancioni, G.: Modeling the response of tensile steel bars by means of incremental energy minimization. J. Elast. **121**, 25–54 (2015)
8. Pham, K., Amor, H., Marigo, J.-J., Maurini, C.: Gradient damage models and their use to approximate brittle fracture. Int. J. Damage Mech. **20**, 618–652 (2011)

Minimum Induced Drag Theorems for Nonplanar Systems and Closed Wings

Luciano Demasi, Giovanni Monegato, and Rauno Cavallaro

Abstract An analytical formulation for the induced drag minimization of generic single-wing non-planar systems, biwings, and closed systems is presented. The method is based on a variational approach, which leads to the Euler–Lagrange integral equations in the unknown circulation distributions. The relationship between quasi-closed C-wings, biwings, and closed systems is discussed and several induced drag theorems/properties are introduced. It is shown that under optimal conditions these systems present the same minimum induced drag and the circulation can be obtained from a fundamental one by just adding a constant. The shape of the optimal aerodynamic load on the Box Wing is shown to change with the distance between the wings; differently that what assumed in previous works, it is not the superposition of a constant and an elliptical function.

1 Introduction

Minimizing the airplane drag can result in large savings in fuel consumption. Moreover, the overall pollution and emissions would be reduced favoring a more sustainable air transportation system. Researchers have been working on the drag minimization problem since the first decades of aviation history [1, 2]. They initially focused on induced drag, one of the main drag components [3–7]. On that respect, it was recognized the superior induced drag performance of the Box Wing [2, 8–10], which can be described (an example is shown in Fig. 1) as a biplane with the tips joined so that in a frontal view the system reminds of a box shape.

L. Demasi
Department of Aerospace Engineering, San Diego State University, San Diego, CA, USA
e-mail: ldemasi@mail.sdsu.edu

G. Monegato
Dipartimento di Scienze Mathematiche, Politecnico di Torino, Turin, Italy
e-mail: giovanni.monegato@polito.it

R. Cavallaro (✉)
Departamento de Bioingeniería e Ingeniería Aeroespacial, Universidad Carlos III de Madrid, Madrid, Spain
e-mail: rauno.cavallaro@gmail.com

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_8

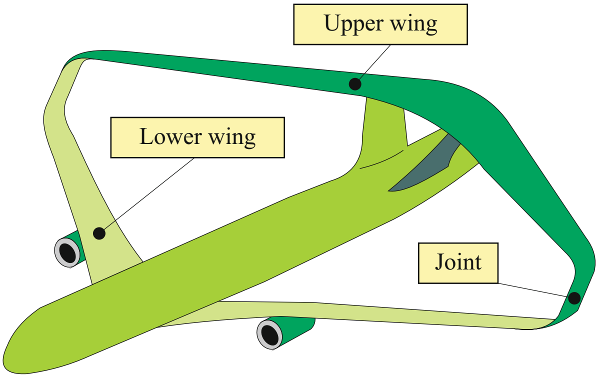


Fig. 1 Artistic example of a Box Wing

Many other works investigated Joined Wings/closed systems [11–15] because of their promising advantages [16] in aerodynamics, engine integration, flight mechanics, etc. Some studies were quite detailed and involved multidisciplinary design optimization [17], others focused on the theoretical aspects of induced drag performance, such as the optimal aerodynamic load achieving the optimum (see [8, 18, 19]) or the theoretical asymptotic behavior for large distance between the upper and lower wings [9, 20, 21] of a Box Wing.

Additional theoretical works on induced drag performance of various wing systems [3–5, 22, 23] confirmed the positive effects [10] of nonplanar geometries. Recently [24], using variational formulations, the minimization of induced drag of various systems (including generic closed wings, biwings, and multiwings) has been investigated with the introduction of new theorems/properties and efficient numerical procedures tailored for the early design phases of new configurations. Several known results, such as Munk's Minimum Induced Drag Theorem, have also been augmented with new findings and insights.

The present work summarizes all the major findings presented in publications [5, 20, 21] and specifically addresses theorems which relate closed systems, quasi-closed C-wings, biwings whose lifting lines essentially identify a closed path, and other relevant induced drag properties.

2 Problem Formulation and Minimum Induced Drag Conditions

The induced drag minimization procedure presented in this work [5, 20, 21] is invariant and can be applied to generic non-planar wing systems with relatively small chord compared to the wingspan (see Fig. 2). In particular, the following steps are taken to set up the aerodynamic model:

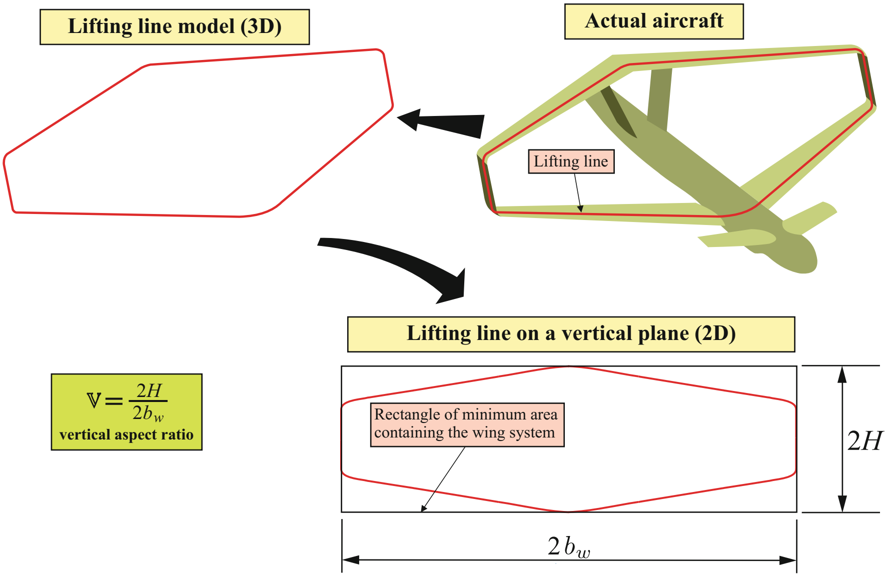

Lifting line model (3D)
Actual aircraft
Lifting line
Lifting line on a vertical plane (2D)
$\mathbb{V} = \frac{2H}{2b_w}$
vertical aspect ratio
Rectangle of minimum area containing the wing system
$2H$
$2\,b_w$

Fig. 2 Example of actual aircraft, 3D and 2D lifting line models. Definition of vertical aspect ratio

- The lifting lines are identified (usually they connect the first-quarter-chord points). A three-dimensional model is then built, as shown in Fig. 2, reporting an example of closed system.
- Since the goal is to minimize the induced drag, according to *Munk's stagger theorem* (applicable under the assumption of rigid wake aligned with the frestream velocity) the longitudinal position of the vortices is not relevant as far as the induced drag is concerned. The problem is then simplified and a two-dimensional lifting line model can be adopted: the entire wing system is actually analyzed in a vertical plane, as shown in Fig. 2. Figure 2 also shows the definition of *vertical aspect ratio*, a wing parameter which has relevant importance in the induced drag performance of the system, as will be discussed later. The two-dimensional lifting line is assumed to be given by a smooth parametric representation, symmetric with respect to the vertical axis and defined in the interval $[-a, a]$. Furthermore, the circulation Γ vanishes at the endpoints $\pm a$ for open systems.

The proposed formulation can be applied not only to single-wing systems but also biwings, closed systems, and multiwings (addressed in [24]). The purpose of this paper is to outline the main ideas, so only the key expressions for the *open* single-wing case (see Fig. 3 and [5]) are reported.

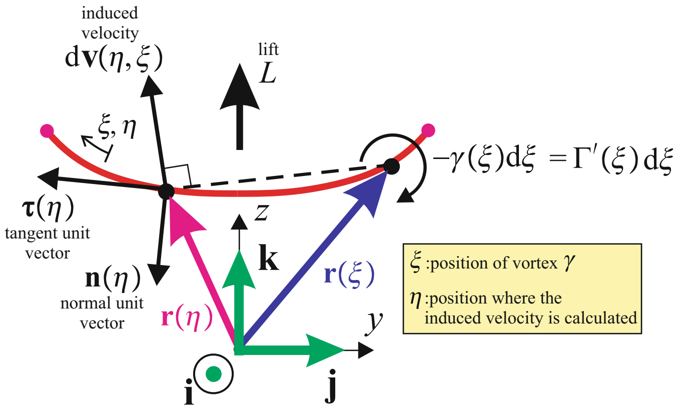


Fig. 3 Single-wing case: contribution to the induced velocity calculated at η

After the wing has been translated into a vertical plane (as shown in Fig. 2), and the lifting line parametric representation is used, the expressions of the *lift* L and *induced drag* D_{ind} are written in terms of the *circulation* Γ as follows:

$$L = -\int_{-a}^{a} \rho_\infty V_\infty \tau_y(\eta)\ \Gamma(\eta)\, d\eta \tag{1}$$

$$D_{\text{ind}} = -\int_{-a}^{a} \rho_\infty v_n(\eta)\ \Gamma(\eta)\, d\eta \tag{2}$$

The quantities ρ_∞ and V_∞ indicate the density and freestream velocity, respectively. v_n is the so-called *normalwash*: it is the component of the induced velocity in the direction perpendicular to the lifting line (see Fig. 3).

Using the quantities reported in (1) and (2), the *functional* J which needs to be minimized, under the constraint of prescribed[1] lift $L = L_{\text{pres}}$, and in a proper functional space V, is written as

$$J(\Gamma, \lambda) = D_{\text{ind}} - \lambda\left(L - L_{\text{pres}}\right) \tag{3}$$

The above space V is the weighted Sobolev type space

$$V = \{w = \mu\overline{w}, \overline{w} \in L^2_{\mu,1}(-a,a)\}, \quad \mu(\xi) = \sqrt{a^2-\xi^2},$$

$$L^2_{\mu,1}(-a,a) = \{u : \int_{-a}^{a} (a^2-\xi^2)^{1/2+k}|u^{(k)}(\xi)|^2 d\xi < \infty,\ k = 0, 1\}$$

[1] At this stage in the discussion there is a need to distinguish between the lift L and the prescribed lift L_{pres}. Later in the text L will be used to indicate the prescribed lift since there will be no ambiguity.

Recalling that (see [25]) $\overline{w} \in L^2_{\mu,1}(-a,a)$ implies $\overline{w} \in C^{0,\gamma}(-a,a) \cap L^p(-a,a)$ for any $0 < \gamma < \frac{1}{2}$ and $1 < p < 4$, we remark that we necessarily have $w \in C[-a,a], w(-a) = w(a) = 0$, a property that the optimal circulation must satisfy.

Thus, the minimization problem is formulated as follows: *find $\Gamma \in V$ and $\lambda \in R$ such that* $J(\Gamma, \lambda) = \min$. In [20] it has been proven that the functional J is convex when acting on $V \times R$, and, furthermore, that there is a unique $\Gamma \in V$ and a unique real λ minimizing J. Finally, it has been shown that the extremum condition for the functional J is obtained when the following integral expression holds:

$$\int_{-a}^{a} \rho_\infty \left[-2v_n^{\text{opt}}(\eta) + \lambda V_\infty \tau_y(\eta)\right] \delta(\eta)\, \mathrm{d}\eta = 0, \quad \forall \delta \in V \tag{4}$$

After application of the fundamental lemma of calculus of variations, it is possible to demonstrate that the optimal conditions corresponding to the minimum induced drag for a given lift and wingspan can be achieved if the *Euler–Lagrange Equation (ELE)* is satisfied:

$$v_n^{\text{opt}}(\eta) = \frac{V_\infty}{E^{\text{opt}}} \cos(\vartheta(\eta)) \tag{5}$$

In other words, the Augmented Munk's Minimum Induced Drag Theorem (AMMIDT) has been obtained:

> When the lifting system has been translated into a single plane (Munk's stagger theorem), the induced drag will be minimum when the component of the induced velocity normal to the lifting element at each point is proportional to the cosine of the angle of inclination of the lifting element at that point. The constant of proportionality is the ratio between the freestream velocity and the optimal aerodynamic efficiency.

Munk discussed this theorem in his original work [1]. However, the constant of proportionality ($\frac{V_\infty}{E^{\text{opt}}}$) relating the normalwash and the cosine of the angle ϑ in (5) was not provided.

Regarding the Lagrange multiplier λ previously introduced, it can be shown that

$$\lambda = \frac{2D_{\text{ind}}^{\text{opt}}}{L} = \frac{2}{E^{\text{opt}}} \tag{6}$$

where E^{opt} is the *aerodynamic efficiency* under optimal conditions.

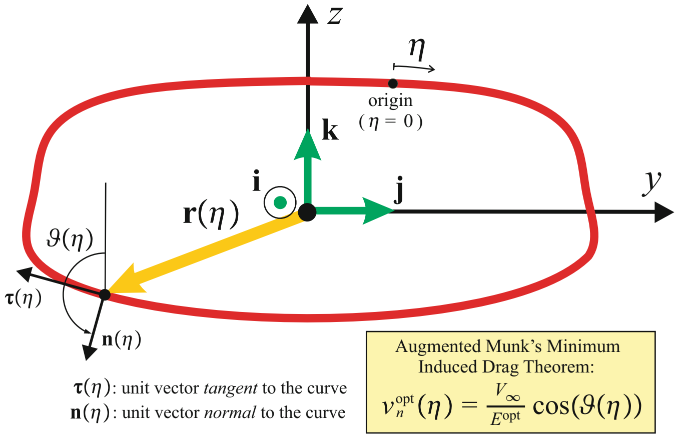

z
η
origin
(η = 0)
k
i
r(η)
j
y
ϑ(η)
τ(η)
n(η)
τ(η): unit vector tangent to the curve
n(η): unit vector normal to the curve
Augmented Munk's Minimum Induced Drag Theorem:
$v_n^{\rm opt}(\eta) = \frac{V_\infty}{E^{\rm opt}} \cos(\vartheta(\eta))$

Fig. 4 Augmented Munk's minimum induced drag theorem

AMMIDT is valid not only for open non-planar wings, but also for closed systems, biwings, and multiwings. The case of closed systems is graphically presented in Fig. 4.

3 Explicit Forms for the Euler–Lagrange Equation

To solve the ELE and find the unknown optimal circulation $\Gamma^{\rm opt}$, it is convenient first to express the normalwash as a function of the circulation (both evaluated under the assumption of optimal conditions). This is accomplished by considering its representation given by

$$v_n^{\rm opt}(\eta) = \frac{1}{4\pi} \fint_{-a}^{a} \frac{{\rm d}\Gamma^{\rm opt}(\xi)}{{\rm d}\xi} Y(\eta, \xi)\, {\rm d}\xi, \quad -a \le \eta \le a \tag{7}$$

where the symbol $\fint$ means that the integral must be interpreted in the Cauchy principal value sense. This because the kernel

$$Y(\eta, \xi) = -\frac{\rm d}{{\rm d}\eta} \ln |\mathbf{r}(\xi) - \mathbf{r}(\eta)| \tag{8}$$

presents a singularity of order 1 when sending and receiving points coincide (i.e., when $\xi = \eta$).

Using the latter expression, the ELE can be written in its explicit *integro-differential form*:

$$\frac{1}{4\pi}\fint_{-a}^{a} \frac{\mathrm{d}\varGamma^{\mathrm{opt}}(\xi)}{\mathrm{d}\xi}\, Y(\eta,\xi)\mathrm{d}\xi = \frac{V_\infty}{E^{\mathrm{opt}}}\, y'(\eta)\,, \quad -a<\eta<a \tag{9}$$

where the (known) function $y(\eta)$ denotes the abscissa of a generic point of the lifting line (see Fig. 4). Furthermore, we can transform (9) into an integral equation where the circulation is not differentiated:

$$-\frac{1}{4\pi}\fint_{-a}^{a} \varGamma^{\mathrm{opt}}(\xi)\; Y(\xi,\eta)\,\mathrm{d}\xi = \frac{V_\infty}{E^{\mathrm{opt}}}\, y(\eta) \tag{10}$$

The constraint of prescribed total lift is still that given by $L = L_{\mathrm{pres}}$.

We remark that expression (9) is useful for examining the endpoint behavior of the first derivative of the optimal circulation (see Sect. 4.2), while (10) is needed for the explicit computation of $\varGamma^{\mathrm{opt}}$.

In particular, after writing the physical quantities in dimensionless form and introducing the parametric representation of the lifting curve, it is possible to rewrite the ELE, as well as the associated constraint for the prescribed lift, in the following dimensionless form:

$$\begin{cases} -\dfrac{1}{\pi}\displaystyle\fint_{-1}^{1} \tilde{\varGamma}^{\,\mathrm{opt}}(u_v)\,\widetilde{Y}(u_v,u)\;\mathrm{d}u_v = \widetilde{y}\,(u)\,, \quad -1<u<1 \\[2ex] -\dfrac{2}{\pi b_w^2}\displaystyle\int_{-1}^{1} \tilde{\varGamma}^{\,\mathrm{opt}}(u)\widetilde{y}'\,(u)\;\mathrm{d}u = \varepsilon \end{cases} \tag{11}$$

The quantity $2b_w$ indicates the wingspan, $\tilde{\varGamma}^{\,\mathrm{opt}}(u_v)$ is the dimensionless optimal circulation, and $\widetilde{y}\,(u)$ is the non-dimensional parameterized y coordinate. The new kernel still presents a singularity of order 1 when sending and receiving points coincide (i.e., when $u = u_v$).

Note that the two equations in (11) are decoupled. Thus, one first determines the unknown circulation $\widetilde{\varGamma}^{\mathrm{opt}}$ by solving the singular integral equation, and then compute the corresponding parameter ε, which is the *optimal aerodynamic efficiency ratio*:

The optimal aerodynamic efficiency ratio ε for a given wing represents the ratio between its aerodynamic efficiency and the corresponding efficiency of a reference classical cantilevered wing with the same wing span and total lift. Both efficiencies are evaluated under their respective optimal conditions.

For the solution of “system” (11), an efficient numerical method has been proposed in [20].

4 Single-Wing Non-planar Systems: Induced Drag Theorems

4.1 *Minimum Induced Drag Curvature-Invariance Theorem*

The problem of finding the minimum induced drag for a given lift and wingspan presents an important property of invariance. To discuss this property, consider a generic single-wing system symmetric with respect to the $x - z$ plane. Suppose that the minimum induced drag problem has been solved and that $\tilde{\Gamma}^{\text{opt}}$ is the corresponding solution.

Consider now a lifting line perfectly identical to the one previously considered. Assume that this second lifting line is mirrored with respect to the first one (i.e., the two lifting lines are the symmetric image of each other with respect to the $\widetilde{y}$ axis, as Fig. 5 shows). The *Minimum Induced Drag Curvature-Invariance Theorem (MIDCIT)* states that

Changing the sign of the curvature of the lifting line (i.e., the arc from convex is changed to concave or viceversa) does not modify the optimal induced drag and circulation distribution: the optimal solution is then invariant if the sign of the curvature is modified.

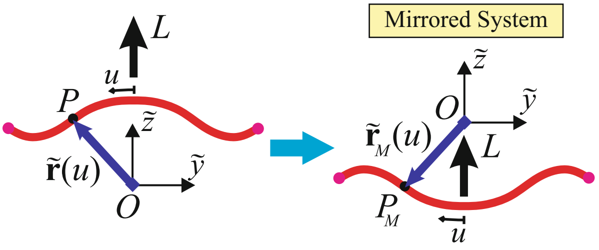


Fig. 5 Identical lifting lines which are the symmetric image of each other with respect to the $\tilde{y}$ axis

4.2 Quasi-Closed C-Wing Zero Gradient Optimal Circulation Theorem

When the tips of a C-wing are (symmetrically and smoothly) brought close to each other, an interesting property of the optimal circulation arises. This is expressed by the *Quasi-closed C-Wing Zero-gradient Optimal Circulation Theorem (QCWZOCT)*:

> If the two tips of a C wing are brought indefinitely close to each other, then the values that the optimal circulation first derivative take at the two tips tend to zero.

We recall that for any given value of a in (10), we have $\Gamma^{\mathrm{opt}}(\pm a) = 0$.

5 Quasi-Closed C-Wing Minimum Induced Drag Conditions

A C-wing presents excellent aerodynamic performance in terms of induced drag. This is confirmed in Fig. 6. The parameter χ is used to define the equation of the lifting line. When χ is close to 1 the curve is almost closed and the wingtips are very near to each other (but the wing is still an open system). The case analyzed in Fig. 6 presents a vertical aspect ratio of aeronautical interest ($\mathbb{V} = 0.2$). It can be seen that Box Wing and C-wing (both under their respective optimal conditions) practically have the same induced drag performance. However, it is conceptually

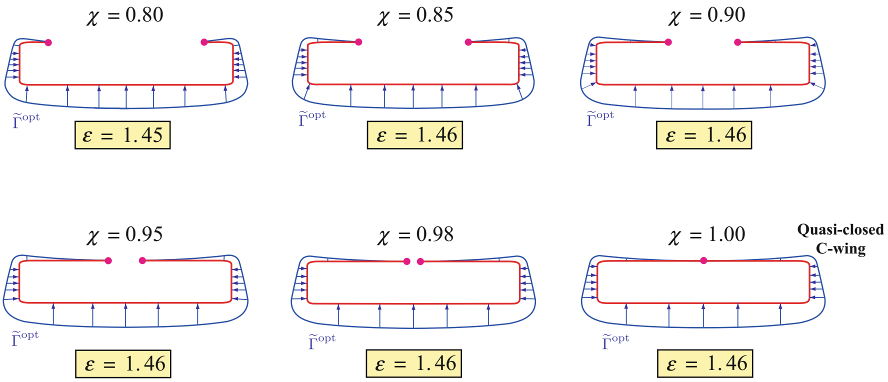


Fig. 6 Optimal aerodynamic loads and efficiency ratios of various C-wings

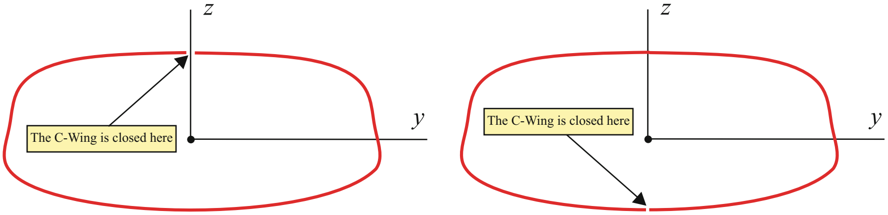


Fig. 7 Two alternative possibilities to close a C-wing at a point on the z axis: north pole and south pole options

not sound to base the design only on aerodynamic efficiency: other disciplines need to be included in a multidisciplinary design optimization to actually reach a true optimal configuration. In other words, Fig. 6 alone cannot be used to claim that C-wings are better than Box Wings.

The relationship between C-wing and closed system (see Fig. 7) may be formally stated with the *Closed System's C-Wing Limit Theorem (CSCWLT)* (for which no formal proof is still available, but this property has been numerically verified):

> The optimal induced drag of a C-wing is equal to the one relative to the corresponding closed wing system, when the tips of the C-wing are smoothly brought indefinitely close to each other. The resulting optimal circulation is the fundamental optimal circulation for the closed wing system for the case of origin of the curve on the z *axis*.

6 Effects of Vertical Aspect Ratio and Dihedral

Figures 8, 9, 10, and 11 report several highly non-planar systems. The following can be observed:

- Increasing the vertical aspect ratio is very beneficial: the optimal aerodynamic efficiency ratio ε is positively affected.
- Adding vertical portions to the wings, even if they aerodynamically do not contribute to the lift (which is in the vertical direction) has a beneficial effect. This is evident if Figs. 8 and 10 are compared.

In all the cases the quasi-closed C-wing achieved the same optimal aerodynamic efficiency ratio of the corresponding closed system.

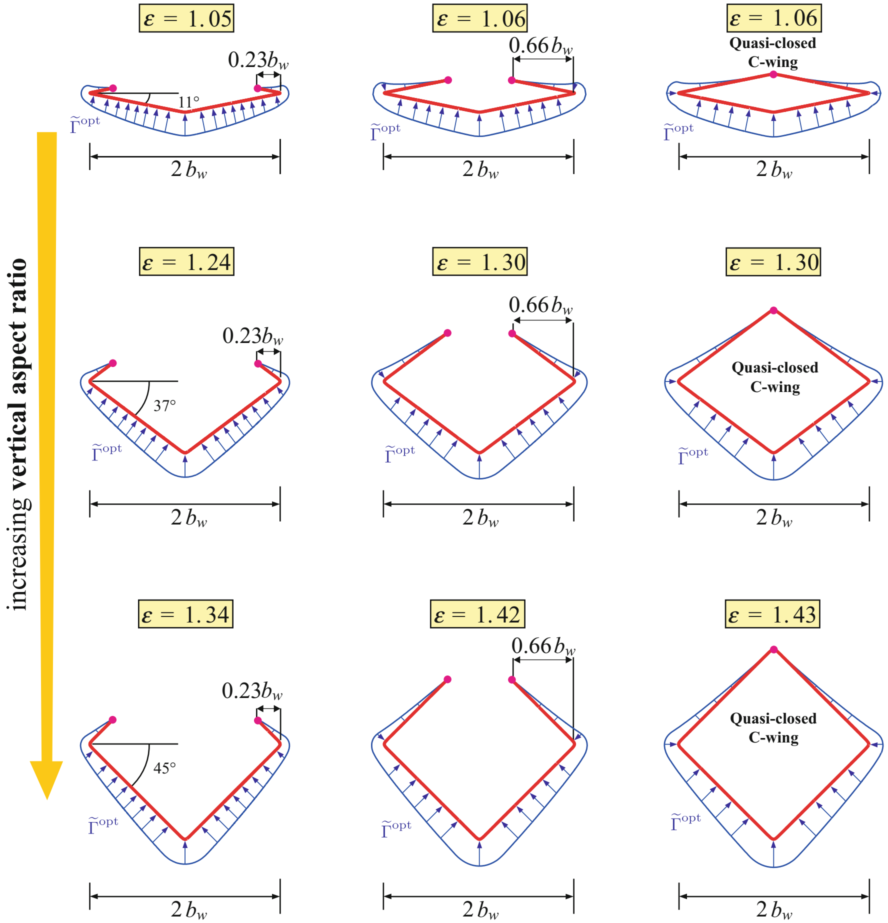


Fig. 8 Optimal conditions for highly non-planar wings (I)

7 Biwings

The above discussion showed the key equation for open single-wing non-planar systems. Similar procedure can be applied when more wings are present. This is the case, for example, of bi-wing configurations. The term *biwing* indicates a system made of two (smooth) non-planar and disjoint wings (a biplane is a particular case of biwing). The study of this new configuration is a straightforward generalization of the previous (open) single-wing case; thus, we simply summarize the main steps in Fig. 12.

Note that in a biwing there are two optimal circulations: one for each wing, and a system of ELEs needs to be solved. The quantity l_{wj} represents the reference length used to write the equations in dimensionless form when wing j is considered.

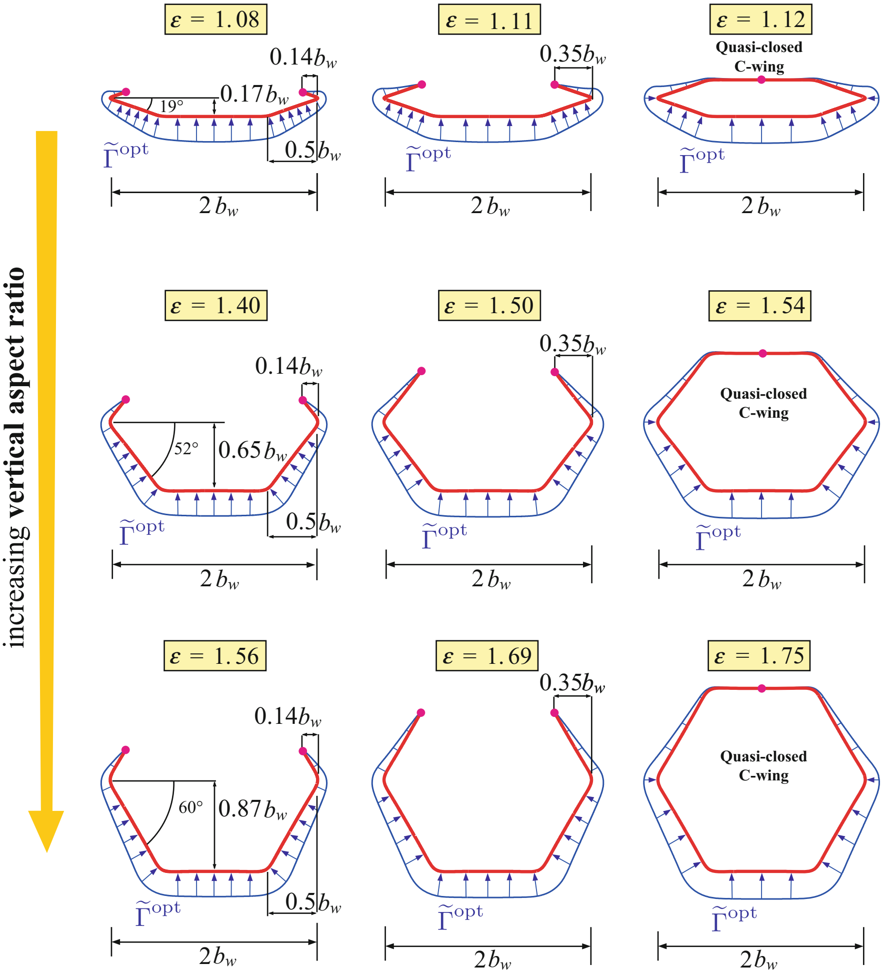


Fig. 9 Optimal conditions for highly non-planar wings (II)

For example, l_{wj} can be selected to be the semi-wingspan b_{wj} relative to wing j. The other quantity appearing in Fig. 12 is l_{kj} which is the ratio between the reference lengths: $l_{kj} = l_{wk}/l_{wj}$.

Several induced drag theorems have been demonstrated [20]. They are briefly reported below.

In a general biwing under optimal conditions, the aerodynamic efficiency of each wing is equal to the aerodynamic efficiency of the entire wing system. This theorem holds even if the two wings are not the same and present different shapes and wingspans.

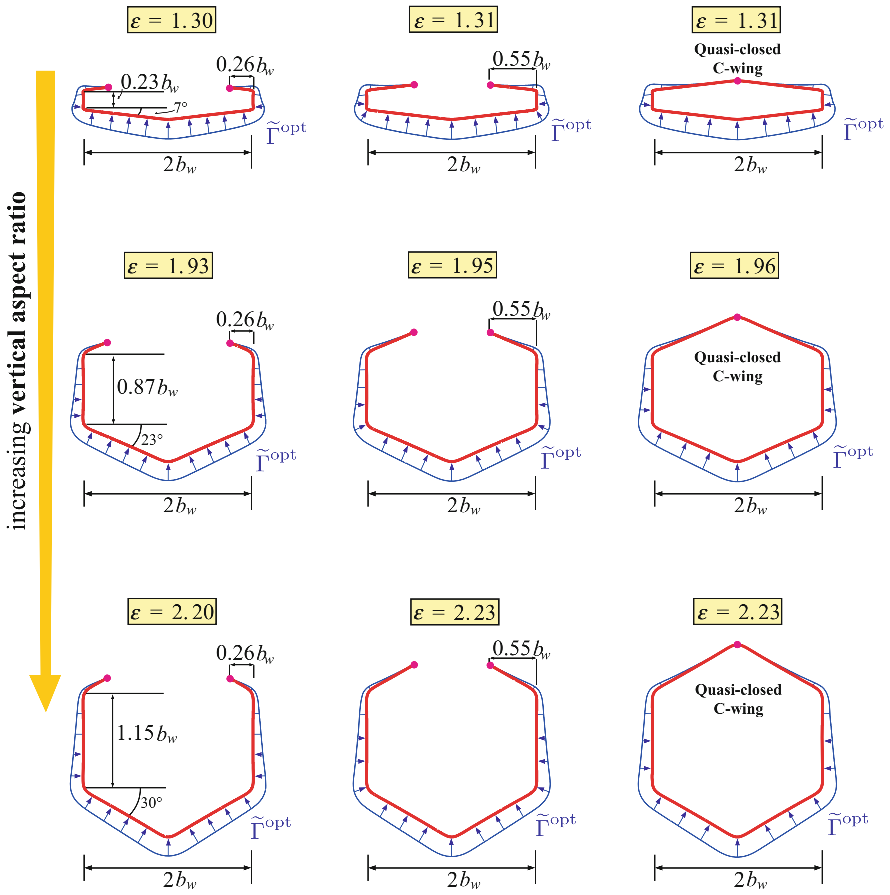


Fig. 10 Optimal conditions for highly non-planar wings (III)

It is immediate to recognize a direct practical consequence:

> In a general biwing under optimal conditions and subjected to a positive lifting force (constraint), the aerodynamic lift on each wing cannot be negative.

From Figs. 13, 14, 15, 16, 17, 18, 19, 20, 21, and 22 the following can be observed:

- The load repartition (under optimal conditions) changes when the distance between the wings is changed.

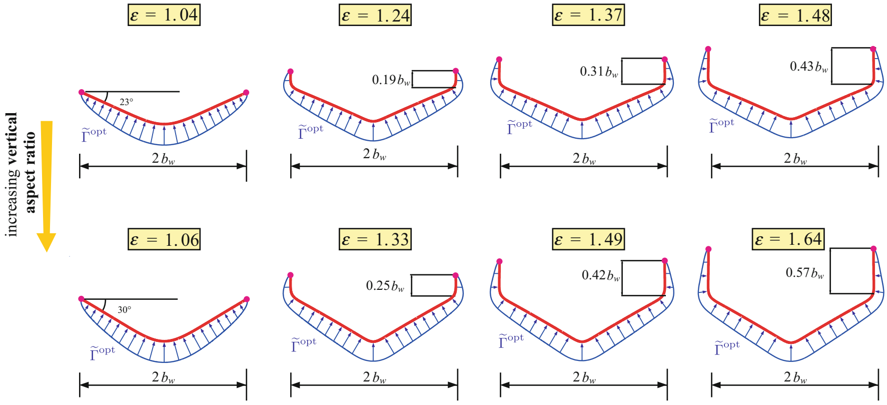
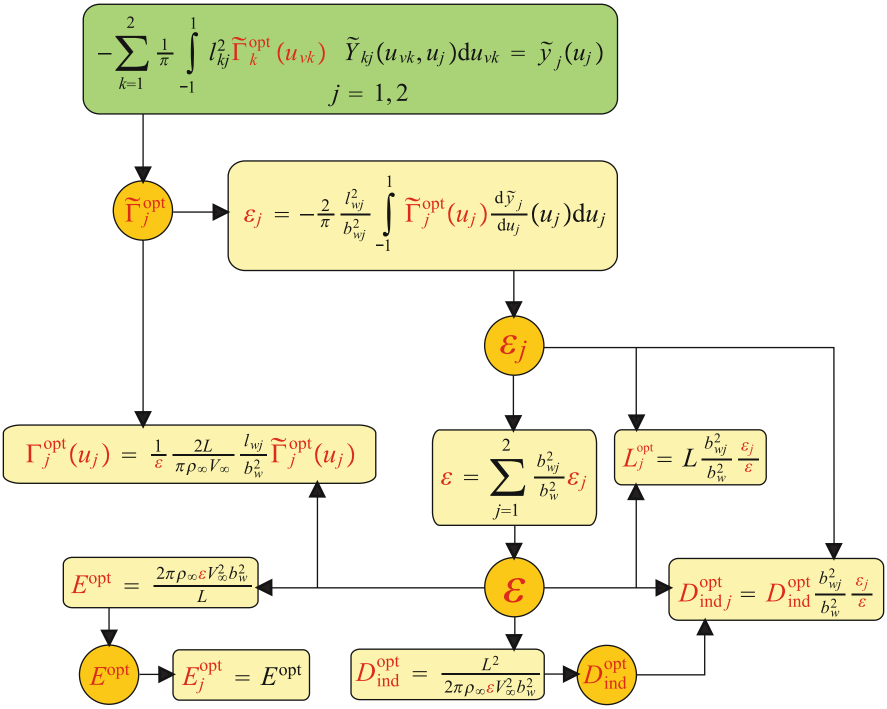


Fig. 11 Optimal conditions for highly non-planar wings (IV)



$-\sum_{k=1}^{2}\frac{1}{\pi}\int_{-1}^{1} l_{kj}^2\widetilde{\Gamma}_k^{\text{opt}}(u_{vk})\ \widetilde{Y}_{kj}(u_{vk},u_j)\,\mathrm{d}u_{vk} = \widetilde{y}_j(u_j)$
$j = 1,2$
$\widetilde{\Gamma}_j^{\text{opt}}$
$\varepsilon_j = -\frac{2}{\pi}\frac{l_{wj}^2}{b_{wj}^2}\int_{-1}^{1}\widetilde{\Gamma}_j^{\text{opt}}(u_j)\frac{\mathrm{d}\widetilde{y}_j}{\mathrm{d}u_j}(u_j)\,\mathrm{d}u_j$
ε_j
$\Gamma_j^{\text{opt}}(u_j) = \frac{1}{\varepsilon}\frac{2L}{\pi\rho_\infty V_\infty}\frac{l_{wj}}{b_w^2}\widetilde{\Gamma}_j^{\text{opt}}(u_j)$
$\varepsilon = \sum_{j=1}^{2}\frac{b_{wj}^2}{b_w^2}\varepsilon_j$
$L_j^{\text{opt}} = L\frac{b_{wj}^2}{b_w^2}\frac{\varepsilon_j}{\varepsilon}$
$E^{\text{opt}} = \frac{2\pi\rho_\infty\varepsilon V_\infty^2 b_w^2}{L}$
ε
$D_{\text{ind}\,j}^{\text{opt}} = D_{\text{ind}}^{\text{opt}}\frac{b_{wj}^2}{b_w^2}\frac{\varepsilon_j}{\varepsilon}$
E^{opt}
$E_j^{\text{opt}} = E^{\text{opt}}$
$D_{\text{ind}}^{\text{opt}} = \frac{L^2}{2\pi\rho_\infty\varepsilon V_\infty^2 b_w^2}$
$D_{\text{ind}}^{\text{opt}}$

Fig. 12 Minimization of induced drag for a biwing: numerical procedure

- Larger wings take an important share of the aerodynamic load. This is intuitively explained by observing that the reverse would imply much higher gradients of circulation on the smaller wing (thus determining a penalty in induced drag).

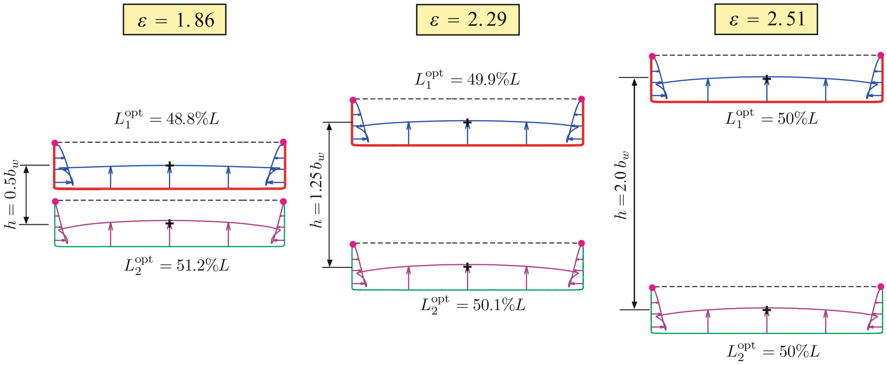

$\varepsilon = 1.86$
$L_1^{\text{opt}} = 48.8\%L$
$h = 0.5b_w$
$L_2^{\text{opt}} = 51.2\%L$
$\varepsilon = 2.29$
$L_1^{\text{opt}} = 49.9\%L$
$h = 1.25\,b_w$
$L_2^{\text{opt}} = 50.1\%L$
$\varepsilon = 2.51$
$L_1^{\text{opt}} = 50\%L$
$h = 2.0\,b_w$
$L_2^{\text{opt}} = 50\%L$

Fig. 13 Biwing made of straight wings with winglets. Both wings have the same wingspan

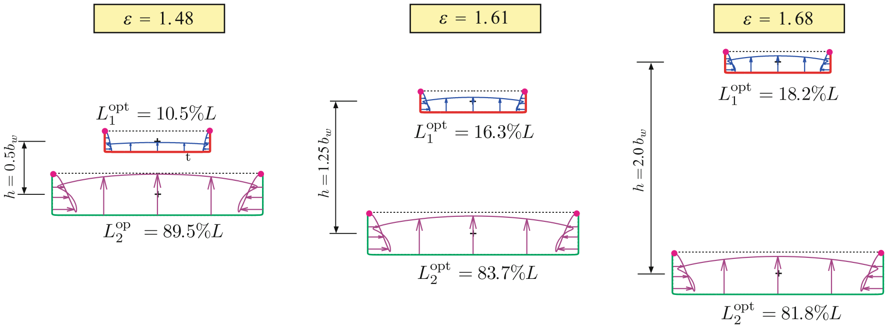

$\varepsilon = 1.48$
$L_1^{\text{opt}} = 10.5\%L$
t
$h = 0.5b_w$
$L_2^{\text{op}} = 89.5\%L$
$\varepsilon = 1.61$
$L_1^{\text{opt}} = 16.3\%L$
$h = 1.25\,b_w$
$L_2^{\text{opt}} = 83.7\%L$
$\varepsilon = 1.68$
$L_1^{\text{opt}} = 18.2\%L$
$h = 2.0\,b_w$
$L_2^{\text{opt}} = 81.8\%L$

Fig. 14 Biwing made of straight wings with winglets. The wings have different wingspans

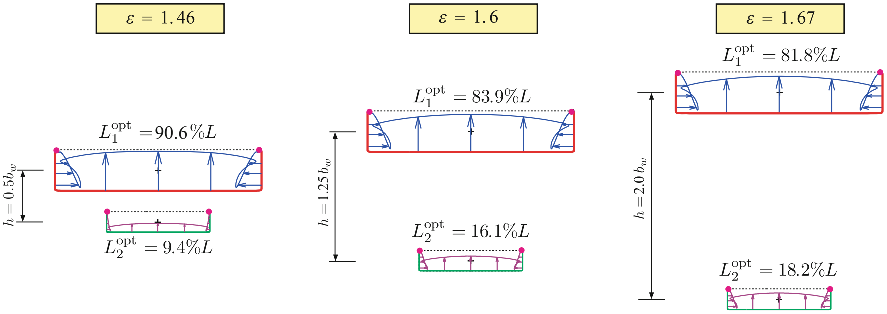

$\varepsilon = 1.46$
$L_1^{\text{opt}} = 90.6\%L$
$h = 0.5b_w$
$L_2^{\text{opt}} = 9.4\%L$
$\varepsilon = 1.6$
$L_1^{\text{opt}} = 83.9\%L$
$h = 1.25\,b_w$
$L_2^{\text{opt}} = 16.1\%L$
$\varepsilon = 1.67$
$L_1^{\text{opt}} = 81.8\%L$
$h = 2.0\,b_w$
$L_2^{\text{opt}} = 18.2\%L$

Fig. 15 Biwing made of straight wings with winglets. The wings have different wingspans

- Non-planar wings have an excellent induced drag performance. However, it is more convenient to have horizontal wings augmented with winglets rather than having dihedral.

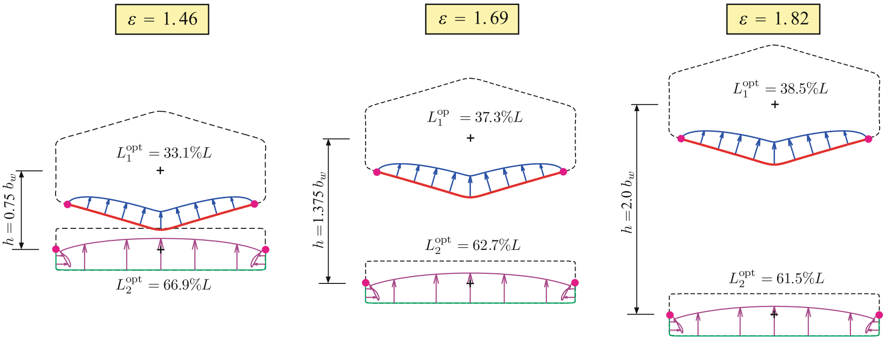

ε = 1.46
ε = 1.69
ε = 1.82
$L_1^{\text{opt}} = 33.1\%L$
$L_2^{\text{opt}} = 66.9\%L$
$h = 0.75\,b_w$
$L_1^{\text{op}} = 37.3\%L$
$L_2^{\text{opt}} = 62.7\%L$
$h = 1.375\,b_w$
$L_1^{\text{opt}} = 38.5\%L$
$L_2^{\text{opt}} = 61.5\%L$
$h = 2.0\,b_w$

Fig. 16 Biwing made of a lower straight wing with winglets and an upper wing presenting dihedral

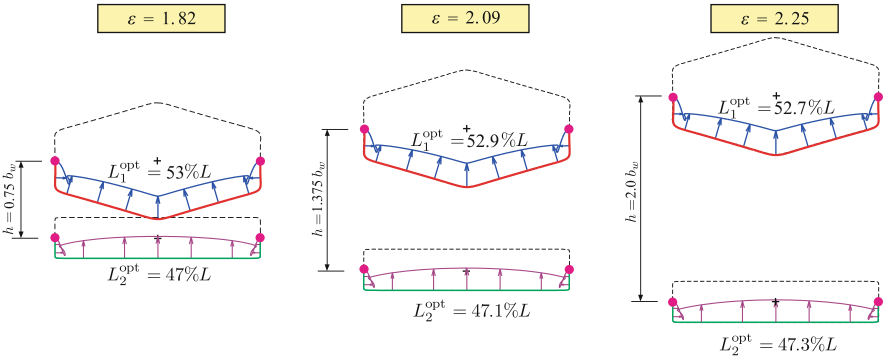

ε = 1.82
ε = 2.09
ε = 2.25
$L_1^{\text{opt}} = 53\%L$
$L_2^{\text{opt}} = 47\%L$
$h = 0.75\,b_w$
$L_1^{\text{opt}} = 52.9\%L$
$L_2^{\text{opt}} = 47.1\%L$
$h = 1.375\,b_w$
$L_1^{\text{opt}} = 52.7\%L$
$L_2^{\text{opt}} = 47.3\%L$
$h = 2.0\,b_w$

Fig. 17 Biwing made of a lower straight wing and an upper wing presenting dihedral. Both wings present winglets

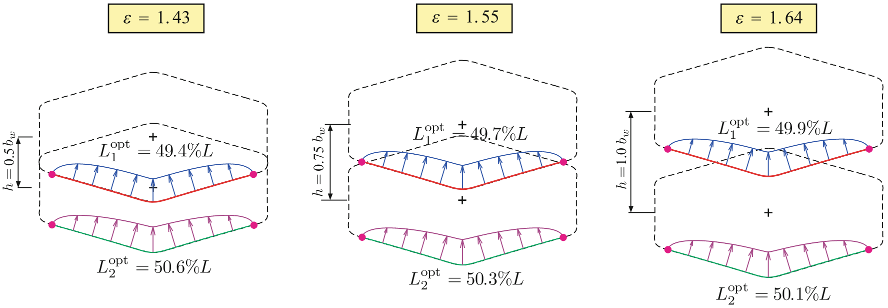

ε = 1.43
ε = 1.55
ε = 1.64
$L_1^{\text{opt}} = 49.4\%L$
$L_2^{\text{opt}} = 50.6\%L$
$h = 0.5\,b_w$
$L_1^{\text{opt}} = 49.7\%L$
$L_2^{\text{opt}} = 50.3\%L$
$h = 0.75\,b_w$
$L_1^{\text{opt}} = 49.9\%L$
$L_2^{\text{opt}} = 50.1\%L$
$h = 1.0\,b_w$

Fig. 18 Biwing made of wings presenting dihedral. Both wings have the same wingspan

- Increasing the vertical aspect ratio (which in the specific case means that the distance between the wings is larger) improves the wing system's performance: the optimal aerodynamic efficiency ratio is higher.

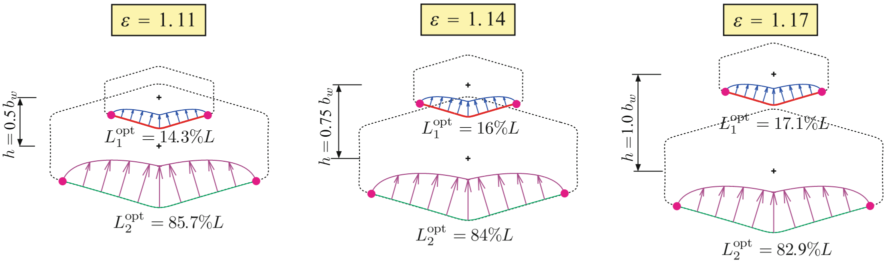


Fig. 19 Biwing made of wings presenting dihedral. The wings have the different wingspans

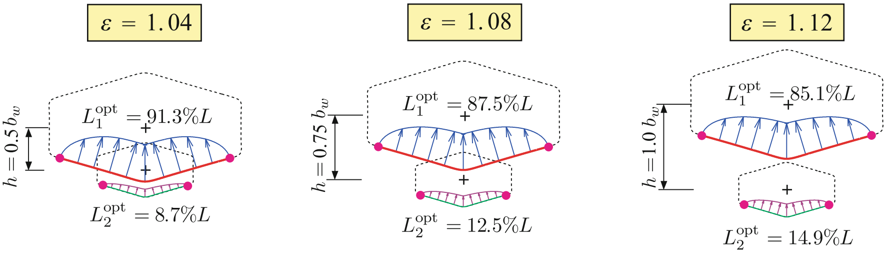


Fig. 20 Biwing made of wings presenting dihedral. The wings have the different wingspans

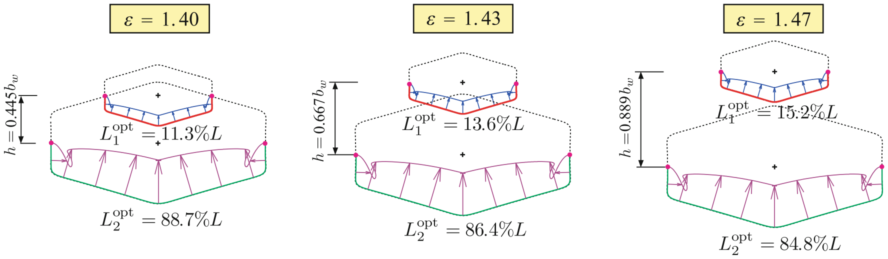

$\varepsilon = 1.40$
$h = 0.445 b_w$
$L_1^{\rm opt} = 11.3\%L$
$L_2^{\rm opt} = 88.7\%L$
$\varepsilon = 1.43$
$h = 0.667 b_w$
$L_1^{\rm opt} = 13.6\%L$
$L_2^{\rm opt} = 86.4\%L$
$\varepsilon = 1.47$
$h = 0.889 b_w$
$L_1^{\rm opt} = 15.2\%L$
$L_2^{\rm opt} = 84.8\%L$

Fig. 21 Biwing made of wings presenting dihedral and winglets. The wings have the different wingspans

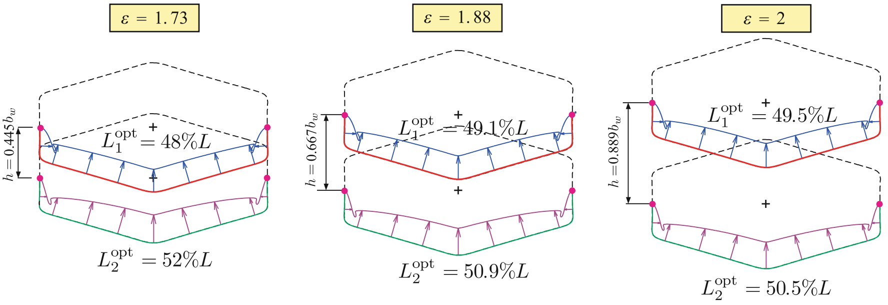


Fig. 22 Biwing made of wings presenting dihedral and winglets. Both wings have the same wingspan

8 Closed Systems

For Box Wings, and in general closed systems, it can be shown that the Augmented Munk's Minimum Induced Drag Theorem holds. Additional properties are now presented.

The *Closed System's Indeterminate Optimal Circulation Theorem (CSIOCT)* states that

> Given a closed wing system, the optimal circulation is indeterminate: there exists an infinite number of equivalent solutions obtained by adding an arbitrary constant to a reference optimal circulation. However, they all have the same optimal induced drag and the same optimal circulation lift.

For practical applications, an important closed system is the Box Wing (see Fig. 1). The optimal aerodynamic load is not uniquely defined due to the above discussed CSIOCT. According to an early work published by Prandtl [2], the optimal induced drag of the Box Wing was related to the induced drag $D_{\text{ind}}^{\text{ref}}$ of an elliptically loaded classical wing with the following formula:

$$\frac{D_{\text{ind}}^{\text{opt}}}{D_{\text{ind}}^{\text{ref}}} \approx \frac{1 + 0.45\mathbb{V}}{1.04 + 2.81\mathbb{V}} \tag{12}$$

Unfortunately, Prandtl did not show how the formula was obtained. From (12) it can be immediately observed that for very small vertical aspect ratios (i.e., $\mathbb{V} \to 0$) Prandtl's equation provides

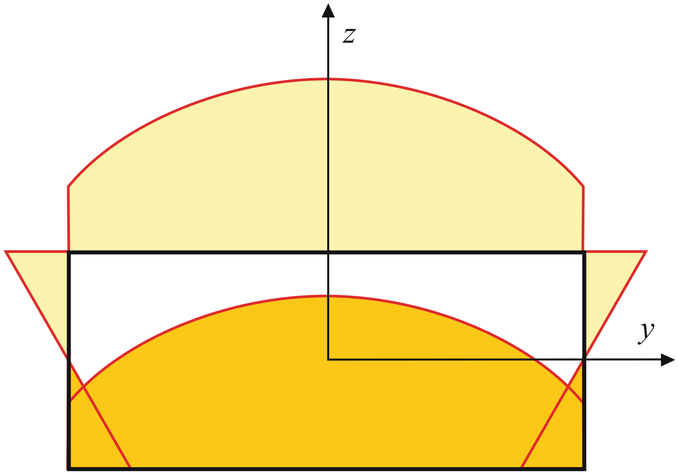
$$\lim_{\mathbb{V}\to 0} \frac{D_{\text{ind}}^{\text{opt}}}{D_{\text{ind}}^{\text{ref}}} \approx \frac{1}{1.04} = 0.96 < 1 \tag{13}$$

In the authors' experience gained with a large number of numerical simulations, the optimal induced drag of the Box Wing actually equates the value of the corresponding elliptically loaded classical monoplane wing. Thus, Prandtl's finding presented a 4 % error.

It is interesting to investigate what (12) predicts when large vertical aspect ratios are considered (i.e., $\mathbb{V} \to \infty$). From (12) it is deduced that

$$\lim_{\mathbb{V}\to \infty} \frac{D_{\text{ind}}^{\text{opt}}}{D_{\text{ind}}^{\text{ref}}} \approx \frac{0.45}{2.81} = 0.16 \tag{14}$$

Other studies predicted the values 0 [9] and 0.5 [18], respectively. It has been numerically verified [21] that the optimal induced drag of a Box Wing asymptotically reaches zero when $\mathbb{V} \to \infty$.



A "*rough sketch*" of the load distribution: see Figure 81 of
T., von Kármán and J. M., Burgers
"*General Aerodynamic Theory - Perfect Fluids*",
Vol II of Aerodynamic Theory, pp. 201-222, 1935

Fig. 23 Representation of the optimal aerodynamic load of a Box Wing as deduced from an early work [8]

Another important aspect in the conceptual understanding of the aerodynamic properties of a Box Wing is the optimal aerodynamic load which minimizes the induced drag. Implicitly assuming an equal load repartition among the wings, the first discussion on this matter [8] showed a circulation distribution which *resembled* the superposition of a constant and an elliptical function (see Fig. 23). The assumption of elliptical plus constant function for the description of the optimal aerodynamic load over the horizontal wings was adopted later in other research [17–19]. Recently, it has been shown that it is an acceptable approximation for relatively small vertical aspect ratios (the ones of aeronautical interest). The actual optimal distribution is shown in Fig. 24. It has numerically been verified that when the vertical aspect ratio is increased, the optimal aerodynamic load on the horizontal wings becomes increasingly similar to a constant function (see Fig. 25). The gradients of circulation are progressively reduced (also in the joints), explaining the asymptotically reached zero optimal induced drag for very large vertical aspect ratio, a behavior similar to a classical monoplane wing when the "horizontal" aspect ratio is indefinitely increased.

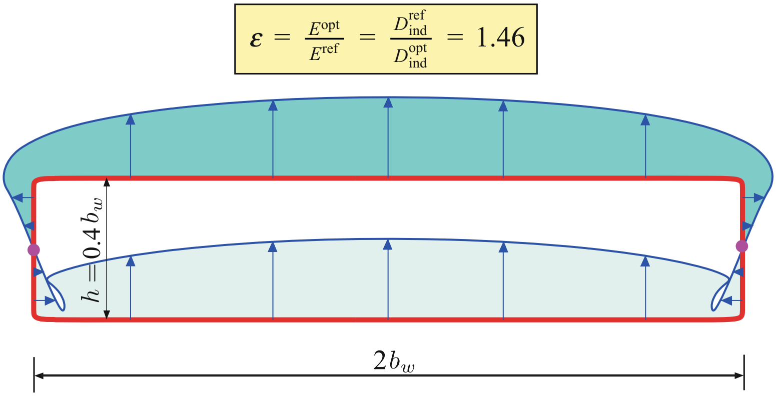

$\varepsilon = \frac{E^{\text{opt}}}{E^{\text{ref}}} = \frac{D_{\text{ind}}^{\text{ref}}}{D_{\text{ind}}^{\text{opt}}} = 1.46$
$h = 0.4\, b_w$
$2b_w$

Fig. 24 Optimal aerodynamic load distribution in a Box Wing with equally loaded wings and with $\mathbb{V} = 0.2$

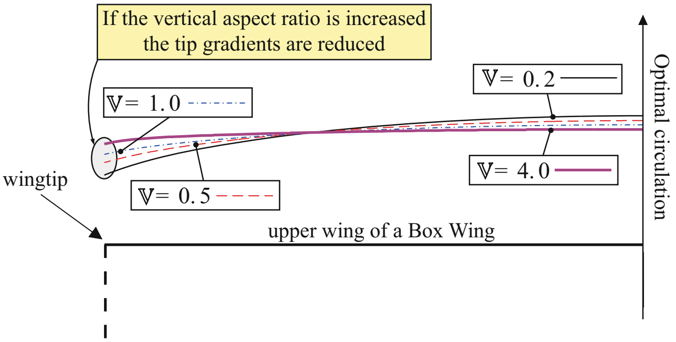
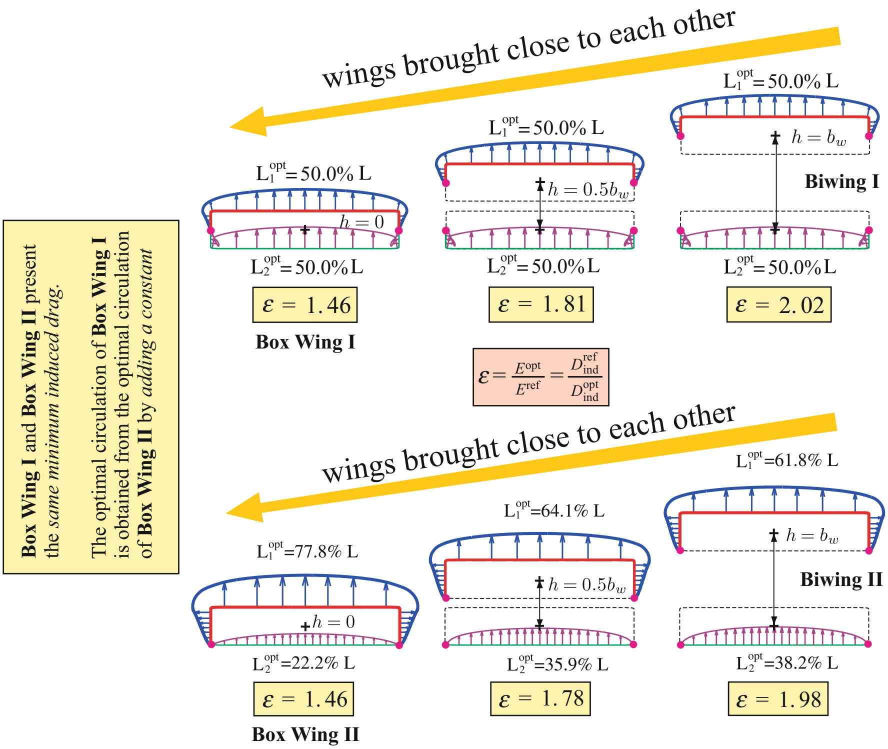


Fig. 25 Aerodynamic load of minimum induced drag on the upper wing of a Box Wing with equally loaded wings. Effect of the vertical aspect ratio

9 Relationship Between Quasi-Closed Systems, Biwings, and C-Wings Under Optimal Conditions

It has been observed that when the tips of C-wings are brought close to each other, then the *open* lifting line essentially identifies the corresponding closed system (see Fig. 6). The optimal induced drags of both the quasi-closed C-wing and closed system are the same. This is also the case if the closed system is "obtained" by bringing the upper and lower wings of a biwing close to each other, so that they essentially identify the lifting line of the closed system: in that case the resulting optimal induced drag coincides with the one of the corresponding closed wing.



Box Wing I and **Box Wing II** present the *same minimum induced drag*.

The optimal circulation of **Box Wing I** is obtained from the optimal circulation of **Box Wing II** by *adding a constant*

Fig. 26 Equivalence between biwing and Box Wing under optimal conditions (minimization of induced drag)

The conceptual equivalence between biwings and closed systems (under their respective optimal conditions) is shown in Fig. 26 for the Box Wing and in Fig. 27 for a more generic closed system presenting dihedral. This equivalence has been numerically verified for a large number of configurations, but a formal mathematical proof is not available at the moment. Actually, the optimal circulations of quasi-closed C-wings, biwings identifying the closed path, and closed systems are all obtained from the same fundamental curve by just adding a constant. This is explained in Figs. 28 and 29. Note that the optimal circulations shown in Figs. 28 and 29 appear to have discontinuity of their slopes. This is not a physical fact, but only a convenient graphical representation (postprocessing) of the circulations.

The equivalence between biwing and closed systems under optimal conditions can be stated by formulating the *Closed Systems's Biwing Limit Theorem (CSBLT)*:

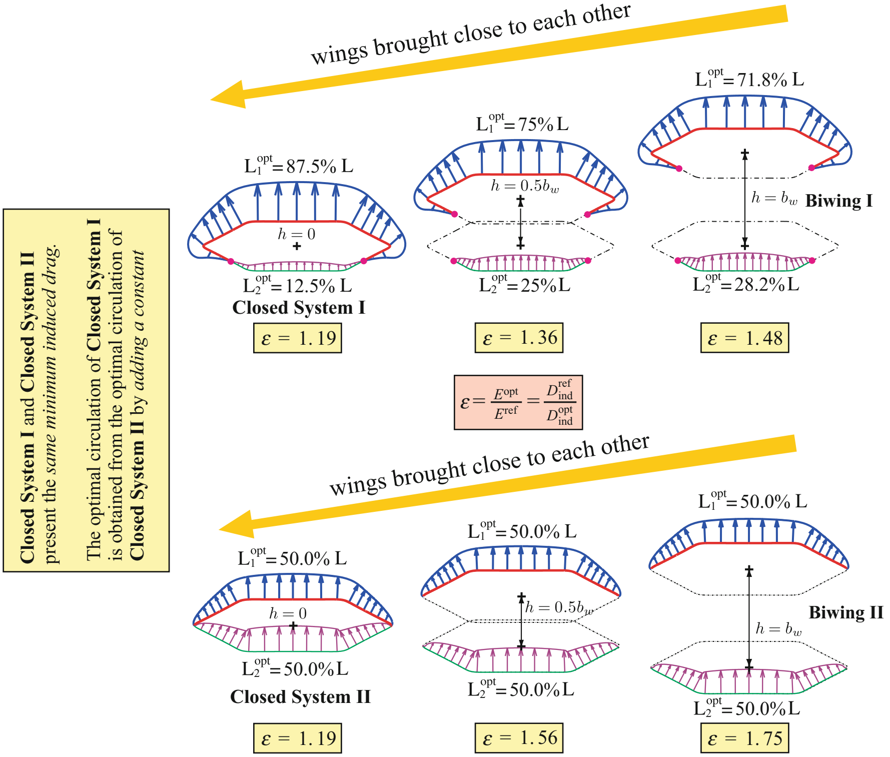


Fig. 27 Equivalence between biwing and a closed system under optimal conditions (minimization of induced drag)

> A wing system defined by a closed and smooth curve can always be considered as a biwing formed by two disjoint wings, whose tips are smoothly brought indefinitely close to each other, so that the limit biwing coincides with the original smooth curve. This biwing provides the same optimal induced drag of the corresponding closed wing system. In particular, the biwing provides the closed wing's fundamental circulation distribution. The origin of the closed curve is on any of the biwing system's tips.

All the findings shown in Figs. 26, 27, 28, and 29 can be summarized with the *Closed System's Minimum Induced Drag Theorem (CSMIDT)*:

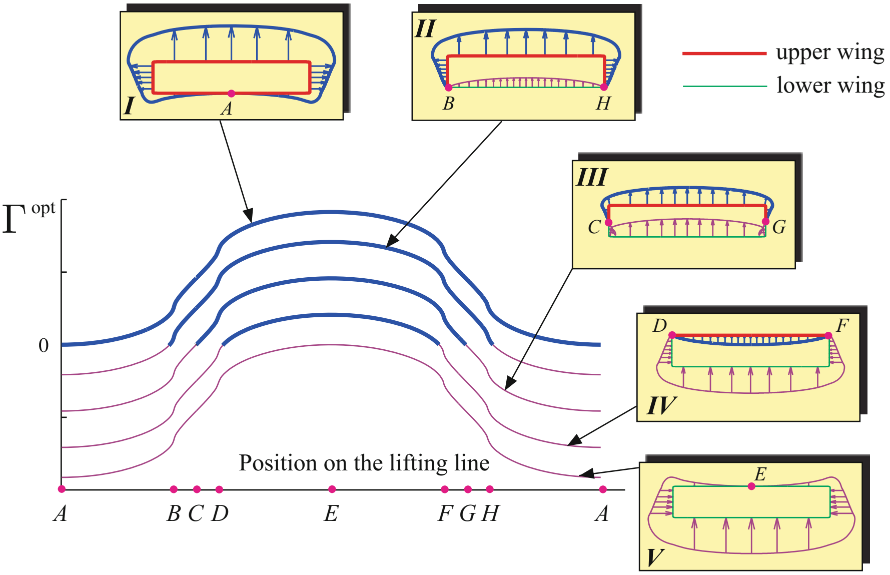


Fig. 28 Relationship between C-wings, biwings, and Box Wings under optimal conditions

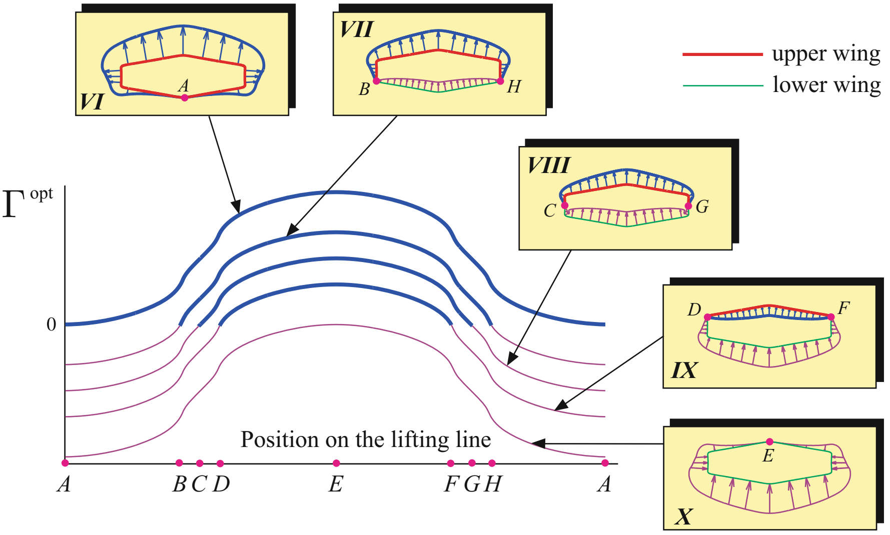


Fig. 29 Relationship between C-wings, biwings and closed systems under optimal conditions

A closed system presents an infinite number of optimal load distributions that minimize the induced drag. These optimal distributions correspond to different load repartitions among the wings, but they all present the same value for the minimum induced drag.

CSMIDT is a direct consequence of CSCWLT, CSIOCT, and CSBLT. Formal mathematical proof of the equivalence, under optimal conditions, of quasi-closed C-wings, biwing, and closed systems is *not available yet.*

10 Numerical Methods

The numerical methods used to solve the ELEs are discussed for the particular case of biwings.

After having rewritten the optimal circulations in the following form:

$$\widetilde{\Gamma}_1^{\text{opt}}(u_{v1}) = \sqrt{1-u_{v1}^2}\,\Upsilon_1(u_{v1}), \qquad \widetilde{\Gamma}_2^{\text{opt}}(u_{v2}) = \sqrt{1-u_{v2}^2}\,\Upsilon_2(u_{v2}), \tag{15}$$

we apply to the ELE system a very simple and effective quadrature method. This is defined by collocating the equation on the zeros of the first kind Chebyshev polynomial of degree $2n+1$ and approximating the integrals by the $2n$-point Gauss-Chebyshev quadrature associated with the weight function $\sqrt{1-u_v^2}$. The resulting system is

$$\begin{cases} -\sum_{s=1}^{2n} w_s^{2n}\left[l_{11}^2\,\widetilde{Y}_{11}\left(u_{v1\,s}^{2n}, u_{1q}^{2n+1}\right)a_{1s} + l_{21}^2\,\widetilde{Y}_{21}\left(u_{v2\,s}^{2n}, u_{1q}^{2n+1}\right)a_{2s}\right] = \widetilde{y}_1(u_{1q}^{2n+1}), \\ -\sum_{s=1}^{2n} w_s^{2n}\left[l_{12}^2\,\widetilde{Y}_{12}\left(u_{v1\,s}^{2n}, u_{2q}^{2n+1}\right)a_{1s} + l_{22}^2\,\widetilde{Y}_{22}\left(u_{v2\,s}^{2n}, u_{2q}^{2n+1}\right)a_{2s}\right] = \widetilde{y}_2(u_{2q}^{2n+1}), \end{cases} \quad q = 1:2n+1 \tag{16}$$

where

$$w_s^{2n} = \frac{1}{2n+1}\sin^2\frac{s\pi}{2n+1}, \quad s = 1:2n,$$

$$u_{v1\,s}^{2n} = u_{v2\,s}^{2n} = \cos\frac{s\pi}{2n+1}, \; s = 1:2n;$$

$$u_{1q}^{2n+1} = u_{2q}^{2n+1} = \cos\frac{(2q-1)\pi}{4n+2}, \; q = 1:2n+1.$$

It is an over-determined set of equations, which has, however, a unique symmetric solution. By taking into account this property, which implies

$$a_{1\;2n+1-s} = a_{1s}, \qquad a_{2\;2n+1-s} = a_{2s}, \quad s = 1:n,$$

as well as the following symmetries

$$\begin{aligned}
&\widetilde{y}_1(-u_1) = -\widetilde{y}_1(u_1), \qquad \widetilde{y}_2(-u_2) = -\widetilde{y}_2(u_2),\\
&\widetilde{Y}_{11}\left(-u_{v1}, u_1\right) = -\widetilde{Y}_{11}\left(u_{v1}, -u_1\right), \quad \widetilde{Y}_{12}\left(-u_{v1}, u_2\right) = -\widetilde{Y}_{12}\left(u_{v1}, -u_2\right),\\
&\widetilde{Y}_{21}\left(-u_{v2}, u_1\right) = -\widetilde{Y}_{21}\left(u_{v2}, -u_1\right), \quad \widetilde{Y}_{22}\left(-u_{v2}, u_2\right) = -\widetilde{Y}_{22}\left(u_{v2}, -u_2\right),\\
&w_s^{2n} = w_{2n+1-s}^{2n}, \quad s = 1:n,\\
&u_{v1\;s}^{2n} = -u_{v1\;2n+1-s}^{2n}, \quad u_{v2\;s}^{2n} = -u_{v2\;2n+1-s}^{2n}, \qquad s = 1:n,\\
&u_{1q}^{2n+1} = -u_{1\;2n+2-q}^{2n+1}, \quad u_{2q}^{2n+1} = -u_{2\;2n+2-q}^{2n+1}, \qquad q = 1:n,
\end{aligned}$$

system (16) is reduced to the following one, which is of order $2n$:

$$\begin{cases}
\displaystyle\sum_{s=1}^{n} w_s^{2n}\Bigg[l_{11}^2\left[\widetilde{Y}_{11}\left(u_{v1\;s}^{2n}, u_{1q}^{2n+1}\right) + \widetilde{Y}_{11}\left(-u_{v1\;s}^{2n}, u_{1q}^{2n+1}\right)\right] a_{1s},\\
\qquad +l_{21}^2\left[\widetilde{Y}_{21}\left(u_{v2\;s}^{2n}, u_{1q}^{2n+1}\right) + \widetilde{Y}_{21}\left(-u_{v2\;s}^{2n}, u_{1q}^{2n+1}\right)\right] a_{2s}\Bigg] = \widetilde{y}_1(-u_{1q}^{2n+1}), \qquad q = 1:n,\\
\displaystyle\sum_{s=1}^{n} w_s^{2n}\Bigg[l_{12}^2\left[\widetilde{Y}_{12}\left(u_{v1\;s}^{2n}, u_{2q}^{2n+1}\right) + \widetilde{Y}_{12}\left(-u_{v1\;s}^{2n}, u_{2q}^{2n+1}\right)\right] a_{1s},\\
\qquad +l_{22}^2\left[\widetilde{Y}_{22}\left(u_{v2\;s}^{2n}, u_{2q}^{2n+1}\right) + \widetilde{Y}_{22}\left(-u_{v2\;s}^{2n}, u_{2q}^{2n+1}\right)\right] a_{2s}\Bigg] = \widetilde{y}_2(-u_{2q}^{2n+1}), \qquad q = 1:n.
\end{cases} \tag{17}$$

We remark that the equations in (16), corresponding to the collocation points $u_{1\;n+1}^{2n+1} = 0$ and $u_{2\;n+1}^{2n+1} = 0$ (i.e., for $q = n + 1$), are trivially satisfied, since both of their members are equal to zero for any given values of the coefficients a_{1s} and a_{2s}. Thus they must be deleted from the system.

Once we have solved this system, the approximate solution is given by the following expressions:

$$\Upsilon_1(u_{v1}) \approx \Upsilon_{1\,n}(u_{v1}) = \sum_{s=1}^{2n} a_{1s}\mathscr{L}_s\left(u_{v1}\right), \qquad \Upsilon_2(u_{v2}) \approx \Upsilon_{2\,n}(u_{v2}) = \sum_{s=1}^{2n} a_{2s}\mathscr{L}_s\left(u_{v2}\right), \tag{18}$$

where $\{\mathscr{L}_s\left(u_{v1}\right), s = 1:2n\}$ and $\{\mathscr{L}_s\left(u_{v2}\right), s = 1:2n\}$ are the $2n - 1$-degree fundamental Lagrange polynomials associated with the $2n$ zeros $\{u_{v1\;d}^{2n}\}$ and $\{u_{v2\;d}^{2n}\}$ of the $2n$-degree Chebyshev polynomial of the second kind ($U_{2n}(u_{v1})$ and $U_{2n}(u_{v2})$, respectively) and defined by the interpolation conditions $\mathscr{L}_s(u_{v1\;w}^{2n}) = \delta_{sd}$ and $\mathscr{L}_s(u_{v2\;d}^{2n}) = \delta_{sd}$, respectively, where δ_{sw} represents the Kronecker symbol. Note that these latter imply

$$a_{1s} = \Upsilon_{1\,n}(u_{v1\;s}^{2n}), \qquad a_{2s} = \Upsilon_{2\,n}(u_{v2\;s}^{2n}).$$

The optimal aerodynamic efficiency ratios ε_1 and ε_2 are then computed by applying the same Gaussian rule mentioned above to their integral representations, that is,

$$\begin{aligned}\varepsilon_i = -\frac{2}{\pi}\frac{l_{wi}^2}{b_{wi}^2}\int\limits_{-1}^{1}\sqrt{1-u_i^2}\,\Upsilon_i\left(u_i\right)\tilde{y}_i'\left(u_i\right)\mathrm{d}u_i \approx -2\frac{l_{wi}^2}{b_{wi}^2}\sum_{s=1}^{2n} w_s^{2n}\,a_{is}\tilde{y}'(u_{2s}^{2n}) = \\ -4\frac{l_{wi}^2}{b_{wi}^2}\sum_{s=1}^{n} w_s^{2n}\,a_{2s}\tilde{y}'(u_{is}^{2n}), \qquad i = 1, 2.\end{aligned} \tag{19}$$

11 Conclusions

A wing-invariant formulation for the induced drag minimization of various open and closed systems is proposed. Under the assumption of potential flow with the wake aligned with the freestream velocity, a variational approach is used and the resulting Euler–Lagrange equations are numerically solved to find the optimal aerodynamic load and induced drag. Quasi-closed C-wings and biwings whose lifting lines essentially identify the closed path of the corresponding closed system have been shown to have the same optimal induced drag. It is also shown that the optimal aerodynamic load of a Box Wing is not given by the superposition of a constant and an elliptical function. This can be an acceptable approximation only for small gap between the upper and lower wings. If the horizontal lifting surfaces of a Box Wing are placed at a large distance, the optimal circulation resembles more a constant function and asymptotically the optimal induced drag approaches zero. Prandtl's formula relating the aerodynamic performance of a Box Wing with its vertical aspect ratio has been shown to have a 4 % error for upper wing very close to the lower wing and is penalizing for large vertical aspect ratios (although Prandtl never suggested to use his formula in that case).

References

1. Munk, M.: The minimum induced drag in airfoils. NACA, Report 121 (1921)
2. Prandtl, L.: Induced drag of multiplanes. National Advisory Committee for Aeronautics, Technical Note No. 182, From Technische Berichte, Volume III, No. 7 pp. 309–315 (1924)
3. Kroo, I.: Nonplanar wing concepts for increased aircraft efficiency. In: VKI Lecture Series on Innovative Configurations and Advanced Concepts for Future Civil Aircraft (2005)
4. Kroo, I.: Drag due to lift: concepts for prediction and reduction. Annu. Rev. Fluid Mech. **33**, 587–617 (2001)

5. Demasi, L., Dipace, A., Monegato, G., Cavallaro, R.: Invariant formulation for the minimum induced drag conditions of non-planar wing systems. AIAA J. **52**(10), 2223–2240 (2014). doi:10.2514/1.J052837
6. Céron-Muñoz, H.D., Cosin, R., Coimbra, R.F.F., Correa, L.G.N., Catalano, F.M.: Experimental investigation of wing-tip devices on the reduction of induced drag. J. Aircr. **50**(2), 441–449 (2013)
7. Henderson, W.P., Holmes, B.J.: Induced drag: historical perspective. In: Society of Automotive Engineers Paper 892341 (1989)
8. Von Karman, T., Burgers, J.M.: General aerodynamic theory - perfect fluids. In: Durand, W.F. (ed.) Aerodynamic Theory, vol. II, pp. 201–222. Dover, New York (1935).
9. De Young, J.: Induced drag ideal efficiency factor of arbitrary lateral-vertical wing forms. NASA Contractor Report 3357 (1980)
10. Cone, C.D.: The theory of induced lift and minimum induced drag of nonplanar lifting systems. Tech. rep., NASA (1962)
11. Wolkovitch, J.: The joined wing aircraft: an overview. J. Aircr. **23** (1986). doi:10.2514/3.45285
12. Lange, R.H., Cahill, J.F., Bradley, E.S., Eudaily, R.R., Jenness, C.M., Macwilkinson, D.G.: Feasibility Study of the Transonic Biplane Concept for Transport Aircraft Applications (1974). NASA CR–132462, Lockheed–Georgia Company
13. Miranda, L.R.: Boxplane Wing and Aircraft US Patent 3,834,654, September 1974
14. Frediani, A.: Large Dimension Aircraft US Patent 5,899,409, 1999
15. Frediani, A., Cipolla, V., Rizzo, E.: The PrandtlPlane configuration: overview on possible applications to civil aviation. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering: Mathematical Challenges for Aerospace Design. Springer Optimization and Its Applications, vol. 66, pp. 179–210. Springer, New York (2012). http://dx.doi.org/10.1007/978-1-4614-2435-2_8
16. Cavallaro, R., Demasi, L.: Challenges, ideas, and innovations of joined-wing configurations: A concept from the past, an opportunity for the future. Progress in Aerospace Sciences (2016, in press). http://dx.doi.org/10.1016/j.paerosci.2016.07.002
17. Rizzo, E.: Optimization Methods Applied to the Preliminary Design of Innovative, Non Conventional Aircraft Configurations. Edizioni ETS, Pisa (2009). ISBN 978-884672458-8
18. Frediani, A., Montanari, G.: Best wing system: an exact solution of the Prandtl's problem. In: Variational Analysis and Aerospace Engineering. Springer Optimization and Its Applications, vol. 33, pp. 183–211. Springer, New York (2009). doi:10.1007/978-0-387-95857-6-11
19. Frediani, A., Montanari, G., Pappalardo, M.: Sul problema di prandtl della minima resistenza indotta di un sistema portante. In: AIDAA, pp. 267–278 (November 1999)
20. Demasi, L., Monegato, G., Dipace, A., Cavallaro, R.: Minimum induced drag theorems for joined wings, closed systems, and generic biwings: theory. J. Optim. Theory Appl. **169**(1), 200–235 (2016). doi:10.1007/s10957-015-0849-y
21. Demasi, L., Monegato, G., Rizzo, E., Cavallaro, R., Dipace, A.: Minimum induced drag theorems for joined wings, closed systems, and generic biwings: applications. J. Optim. Theory Appl. **169**(1), 236–261 (2016)
22. Demasi, L.: Induced drag minimization: a variational approach using the acceleration potential. J. Aircr. **43**, 669–680 (2006)
23. Demasi, L.: A theoretical investigation on the conditions of minimum induced drag of closed wing systems and C-wings. J. Aircr. **44**, 81–99 (2007)
24. Demasi, L., Monegato, G., Cavallaro, R.: Minimum induced drag theorems for multi-wing systems. In: 57th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, AIAA SciTech, (AIAA 2016-0236), San Diego (2016). doi:http://dx.doi.org/10.2514/6.2016-0236
25. Berthold, D., Hoppe, W., Silbermann, B.: A fast algorithm for solving the generalized airfoil equation. J. Integr. Equ. Appl. **43**(1–2), 185–219 (1992)

On Aerodynamic Design with a POD Surrogate Model

Valentina Dolci and Renzo Arina

Abstract Three surrogate models, or reduced-order models (ROMs), are constructed using the proper orthogonal decomposition (POD) applied in the parameter space. We use a reduced snapshot set adopting full and fractional factorial planes together with quadtree distribution for the initial positioning of the snapshots. To compute the POD coefficients, response surface methodology is employed. In the first application a ROM is constructed in order to analyze the subsonic flow past a two-dimensional airfoil. The second example regards a transonic two-dimensional flow in a five-dimensional shape parameter space. In the last case a surrogate model for database generation considering a three-dimensional aircraft configuration is constructed. In all the three cases a posteriori error estimates were performed and the surrogate models showed good agreement with the CFD reference solution.

1 Introduction

Since the last few years, surrogate models are becoming a promising research field for engineering applications. A surrogate, or reduced-order model (ROM), is a mathematical tool able to extract the main features of a more computational demanding high-order model, starting from a reduced set of information. Once the surrogate model is built, it can be used to perform faster analyses of the problem. The fields of optimization and database generation in aerodynamics can be best candidates for the application of ROMs. In the case of aerodynamic optimization, a cost function should be evaluated several times requiring many CFD simulations in order to achieve sufficient information to identify an optimum target. This operation requires a great amount of computational time and effort and the adoption of a surrogate model replacing the CFD high-order model can be attractive.

However the construction of an accurate and robust surrogate model is a delicate process. Particular attention should be paid to the type of surrogate model chosen for a specific problem and to the number and position of the initial set of high-order simulations. In this construction phase there is a constant trade-off between required

V. Dolci (✉) • R. Arina
DIMEAS, Politecnico di Torino, Corso duca degli Abruzzi 24, 10129, Torino (TO), Italy
e-mail: valentina.dolci@polito.it; renzo.arina@polito.it

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_9

accuracy and computational effort reduction: the building of the surrogate model must be fast and not computational demanding preserving in the meantime the primary characteristics of the problem. Several applications can be found in recent literature regarding surrogate models in aerodynamics. For example, in [10] a surrogate model is built to perform aerodynamic database predictions for flight simulations and in [7] a Kriging method is used to estimate the drag polar of a supercritical airfoil. In this work we present the construction of a surrogate model based on a parametrized proper orthogonal decomposition (POD), that is a POD applied in the parameter space. We follow the method of snapshots proposed by Sirovich [11] and we use a reduced initial snapshot set. The POD coefficients are computed using different interpolation techniques and several a posteriori error calculations are performed. The paper is structured as follows: in the first section a theoretical description can be found. After that we present three applications in which we build and test three different surrogate models. In the first application the steady subsonic two-dimensional flow field past a NACA 0012 airfoil is analyzed and a ROM is built in a parameter space composed by the Mach number of the undisturbed flow and the angle of attack of the airfoil. The second surrogate model regards the two-dimensional transonic case of the RAE 2822 airfoil. This model is built using five shape parameters extracted from a Bézier curve. Finally a three-dimensional aircraft configuration is analyzed in a two-dimensional parameter space considering a subsonic steady flow. The parameter space is composed by angle of attack and sideslip angle of the aircraft and a database generation is performed. In the end, some conclusions are given.

2 ROM Through POD

The proper orthogonal decomposition technique can be seen as a mathematical procedure able to build a basis $\{\boldsymbol{\varphi}_j(x)\}_{j=1}^{\infty}$ in order to represent the function u($\boldsymbol{x}$) using the linear approximation $\tilde{u}(\boldsymbol{x})$

$$\tilde{u}(\boldsymbol{x}) = \sum_{j=1}^{M} a_j \boldsymbol{\varphi}_j(\boldsymbol{x}) \sim u(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^p. \tag{1}$$

In this framework the basis $\{\boldsymbol{\varphi}_j(x)\}_{j=1}^{\infty}$ is indicated as POD basis and the functions $\boldsymbol{\varphi}_j(x)$ are called POD functions or modes. p is the cardinality of the parameter space. Many procedures are available in literature to compute the POD basis. In this work we followed the method of snapshots [11]. In this case a set of M representative samples $\{\boldsymbol{u}^k(\boldsymbol{x})\}_{k=1}^{M}$ named snapshots is considered.

The functions composing the basis $\{\boldsymbol{\varphi}_j(x)\}_{j=1}^{\infty}$ are chosen to describe the M snapshots $\{\boldsymbol{u}^k(\boldsymbol{x})\}_{k=1}^{M}$ in an optimal way in the energetic sense, in fact they maximize the squared mean projection of the set $\{\boldsymbol{u}^k(\boldsymbol{x})\}_{k=1}^{M}$ on the basis itself. This can be represented by the optimization problem

$$\max_{\varphi} \frac{\langle |(\boldsymbol{u}(\boldsymbol{x}), \varphi(\boldsymbol{x}))|^2 \rangle}{\|\varphi(\boldsymbol{x})\|^2}, \tag{2}$$

that the POD basis has to satisfy. In Eq. (2) $\langle \cdot \rangle$ is the average operator, $(\cdot, \cdot)$ indicates the inner product between two functions, $|\cdot|$ is the absolute value, and $\|\cdot\|$ is the L^2 norm.

For practical applications the technique should be used in finite dimensions. In this case the set of functions $\{\boldsymbol{u}^k\}_{k=1}^{M}$ becomes a set of M vectors and the basis $\{\varphi_j\}_{j=1}^{M}$ is no longer composed by functions but by vectors. The kernel necessary to solve the maximum problem (2) in finite dimensions is the correlation matrix R of dimensions $[M \times M]$. In this work the samples $\{\boldsymbol{u}^k\}_{k=1}^{M}$ consist of numerical solutions of a PDEs system and the superscript k indicates the varying boundary conditions [1]. In order to compute the basis vectors, the eigenvalue problem

$$R\boldsymbol{b} = \lambda \boldsymbol{b}, \tag{3}$$

has to be solved, where with R is indicated the correlation matrix

$$R_{ij} = \frac{1}{M}(\boldsymbol{u}^i, \boldsymbol{u}^j) \qquad i = 1, ..., M \quad j = 1, ..., M. \tag{4}$$

Once the eigenvalue problem is solved, M eigenvalues $\{\lambda_i\}_{i=1}^{M}$ and M eigenvectors, each one with M components $\{\boldsymbol{b}_i^k\}_{i=1}^{M}, k = 1, ..., M$ are obtained. The optimal POD basis can be calculated recombining the initial snapshots with the eigenvectors and each basis vector will be constructed as

$$\boldsymbol{\varphi}^k = \sum_{i=1}^{M} \boldsymbol{b}_i^k \boldsymbol{u}^i \quad k = 1, ..., M. \tag{5}$$

The physical meaning of the eigenvalues differs with the content of the vectors $\{\boldsymbol{u}^k\}_{k=1}^{M}$. Typically they represent a field variable. In the case of an incompressible flow, for example, these vectors can consist of the discretization of the velocity field obtained from numerical solutions of the Navier–Stokes equations with different boundary conditions and each eigenvalue will be equal to the double of the mean kinetic energy captured by the corresponding POD mode [1].

The ratio

$$\frac{\sum_{1=1}^{t} \lambda_i}{\sum_{1=1}^{M} \lambda_i}, \tag{6}$$

where t corresponds to the number of modes considered, can be used as an indicator of the energy contained in the different modes. The general approach of the POD technique considers this eigenvalue estimate. Typically an a priori threshold is chosen and the modes corresponding to the eigenvalues with an amount of energy under the threshold are neglected. With this procedure, a truncation error is introduced and only q modes are considered, with $q \ll M$.

In the present work, a highly reduced number of snapshots (and therefore of POD modes) is used and no truncation is generated, avoiding the introduction of the corresponding error. All the POD modes are retained in the building of the POD surrogate model because the saving of computational cost performing the truncation was negligible.

Once the POD basis is constructed, it is possible to obtain exactly the initial snapshots $\{\boldsymbol{u}^k\}_{k=1}^M$ as the linear combination

$$\boldsymbol{u}^k = \sum_{i=1}^{M} a_i^k \boldsymbol{\varphi}^i \quad k = 1, ..., M. \tag{7}$$

The coefficients a_i^k can be calculated as angles between the snapshots and the POD basis vectors:

$$a_i^k = (\boldsymbol{u}^k, \boldsymbol{\varphi}^i) \quad k = 1, ..., M. \tag{8}$$

In fluid dynamics usually the truncated POD basis is used to project the Navier–Stokes equations along the optimal POD vectors and the mean operator is applied in time. In this way the reduced-order model (ROM) is formed by a set of ordinary differential equations in the unknowns $a_i^k(t)$.

In the present work a different approach is used: the POD is parametrized, no projection onto a reduced-order dimension space is performed, the eigenvectors are all used to construct the POD basis and weighted with the coefficients a_i^k. In literature this particular application is called POD with interpolation or PODI [12].

With the PODI approach we follow the method of snapshots. Let's call $\{\boldsymbol{u}^{\delta_i}\}_{i=1}^M$ the snapshot set corresponding to the parameter combination set $\{\boldsymbol{\delta}_i\}_{i=1}^M$. The parameter vector δ has cardinality p and the M snapshots are related to M different parameter combinations. It is possible to obtain the POD basis $\{\boldsymbol{\varphi}_j\}_{j=1}^M$ following the method of snapshots described previously: a correlation matrix R should be calculated and an eigenvalue problem of order M has to be solved. The exact reconstruction of each snapshot is given by the linear combination

$$\boldsymbol{u}^{\delta_i} = \sum_{j=1}^{M} a_j^{\delta_i} \varphi^j, \quad i = 1, ..., M. \tag{9}$$

If we can assume that the coefficients $a_j^{\delta_i}$ are the discretization of the function $a_j(\delta)$ and if the function $a_j(\delta)$ is smooth enough with respect to δ, interpolation methods can be used to calculate the PODI coefficients $\widetilde{a_j}^{\delta_l}$ for values of the parameters not belonging to the initial snapshot set.

Once the PODI coefficients $\widetilde{a_j}^{\delta_l}$ are computed, it is possible to evaluate the field variable $\boldsymbol{u}_{un}^{\delta_l}$ corresponding to unknown parameter combinations $\boldsymbol{\delta_l}$ as

$$u_{un}^{\delta_l} \sim \sum_{i=1}^{M} \widetilde{a_i}^{\delta_l} \varphi^i. \tag{10}$$

Equation (10) is the representation of the surrogate model built in the present work and used for the following applications. The Navier–Stokes equations can be considered as the high-order model of the problem, reference or “true” solution. Starting from M numerical solutions of the Navier–Stokes equations, or snapshots, the surrogate model can be used to evaluate the field of a generic variable of interest (as density, temperature, or the turbulent quantities) associated with specific boundary conditions corresponding to different parameter combinations $\boldsymbol{\delta_l}$ not belonging to the initial snapshot set. No extrapolation procedure is allowed at the moment therefore $\boldsymbol{\delta_l}$ has to stay in the parameter space defined by the snapshots. The use of the surrogate model instead of solving the system of PDEs allows to save computational time and effort maintaining acceptable accuracy with respect to the true solution. The accuracy is guaranteed by the fact that basis truncation is avoided and that the POD basis vectors are optimal in the energetic sense.

The method can be outlined in three main steps:

1. Generation of the initial snapshot set: the snapshots are obtained with high-fidelity CFD calculations. The position of the snapshots in the parameter space as well as their number strongly influence the ROM accuracy. This is the most computational expensive step but can be performed off-line only once.
2. POD decomposition: the correlation matrix R is computed and an eigenvalue problem of order M, equal to the snapshot number, is solved. As previously explained, in our approach M is small and all the POD modes are used to construct the surrogate model. There is no particular effort in the solution of the eigenvalue problem. Even this part can be performed one time off-line.
3. PODI reconstruction: this last step should be performed on-line for each reconstruction. The PODI coefficients $\widetilde{a_i}(\delta_l)$ are computed using different interpolation techniques. For parameter spaces with dimension greater than one, the response surface method is applied, with first and second order least square regression or radial basis functions with Gaussian or multiquadric bases. Once the coefficients are calculated, the linear combination of the POD basis can be performed in order to obtain the desired $u_{un}(\delta_l)$.

3 Response Surface Methodology

In the case of parameter spaces with dimension greater than one, the response surface methodology is applied in order to compute the PODI coefficients $\widetilde{a_i}^{\delta_l}$ presented in Sect. 2. This methodology allows interpolation in more than one dimension through the generation of an analytical surface starting from a certain number of output evaluations.

The expression

$$\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon \tag{11}$$

can be used [9] to approximate the input–output relation between $\boldsymbol{x}$ and $\boldsymbol{y}$. In Eq. (11) $\boldsymbol{y}$ is the dependent variable, called response of the system that we are analyzing and from that term arises the definition *response surface*. In our problem the vector $\boldsymbol{y}$ will contain the coefficients $\{a_j^{\delta_i}\}_{i=1}^{M}$ corresponding to the angles between the POD modes and the snapshots. For each a_j a different response surface has to be built, with $j = 1, ..., M$; that is, we are building M different response surfaces to approximate the PODI coefficients corresponding to each POD mode. The vector $\boldsymbol{x}$ contains the independent variables or parameters of the problem. The error ϵ with respect to the true solution is modeled as a random error with 0 mean.

The function $f(\boldsymbol{x})$ can be composed by a certain combination of the components of the parameter vector $\boldsymbol{x}$. The complexity of this combination varies with the accuracy needed in the specific problem and in this work least square regression of the first and second order is used, together with radial basis functions considering Gaussian and multiquadric basis.

3.1 Least Square Regression

In the framework of response surface methodology, least square regression can be a simple technique to estimate a best fit approximation of the PODI coefficients $\widetilde{a_i}^{\delta_l}$.

The general form of a first order least square surface can be written as:

$$y_i = \beta_0 + \sum_{j=1}^{r} \beta_j x_{ij} + \epsilon_i, \quad i = 1, .., M \tag{12}$$

The quantities indicated with β_j are called regressors and shall be estimate solving the system in the least square sense. The quantity r represents the number of the regressors and therefore the number of unknowns of the system.

In the case of a second order regression we have

$$y_i = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i=1}^{p} \beta_{ii} x_i^2 + \sum_{i=1}^{p-1} \sum_{j=2}^{p} \beta_{ij} x_i x_j + \epsilon_i, \quad i = 1, .., M. \tag{13}$$

where p is the cardinality of the parameter space. The second order model can be useful in the case of a strong curvature in the true input–output relation.

In order to solve the problem, (12) or (13) can be written in matrix form as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{14}$$

with $\boldsymbol{y} \in \mathbb{R}^M$ and $\boldsymbol{X} \in \mathbb{R}^{M \times r}$. The vector $\boldsymbol{\beta}$ contains the r regressors of the model that are the unknowns of the system. M is the number of available input–output relations, or the number of already performed evaluations of the system state. In our case M is coincident with the number of snapshots.

System (14) can be solved through least square minimization

$$\boldsymbol{\beta} = (X'X)^{-1} X' \boldsymbol{y}. \tag{15}$$

Once the vector $\boldsymbol{\beta}$ that contains the regressors of the model is found, the response surface is defined and the PODI reconstruction coefficients $\widetilde{a_i}^{\delta_l}$ can be calculated directly from (12) or (13).

3.2 Radial Basis Functions

With the radial basis function method we can avoid regression and perform a classical interpolation, in order to build an analytical surface that is coincident with the data in the starting points, corresponding to evaluations of the high-order model. Radial basis functions can be considered a generalization of splines to the multivariate setting and they are able to avoid the curse of dimensionality treating all space dimensions in the same way.

The general form of a response surface built with radial basis functions is

$$y(x) = \sum_{j=1}^{M} w_j \phi(x - x_{j_2}), \quad x \in \mathbb{R}^p. \tag{16}$$

As we can see from Eq. (16), different positions of the radial function $\Phi = \phi(| \cdot |)$, with $\phi[0, \infty) \rightarrow \mathbb{R}$, produce the linear combination describing the response surface. The term radial is referring to the Euclidean norm $\cdot_2$. In our approach the vector $y(x)$ contains the known PODI coefficients $\widetilde{a_i}^{\delta_l}$ and M is the number of snapshots.

Starting from this general form many particularizations can be and have been made [5]. In this work we tested only two different basis forms: the Gaussian basis $\phi(x) = e^{-\frac{1}{2\sigma_i^2} x - x_{i2}^{2}}$ and the multiquadric basis $\phi(x) = \sqrt{1 + \frac{x - x_{i2}^{2}}{2\sigma_i^2}}$. In these expressions σ is a shape parameter that determines the aspect of the radial basis function.

The solution of the interpolation (16) leads to the system

$$A\boldsymbol{w} = \boldsymbol{y}. \tag{17}$$

The interpolation matrix A has components $A_{ji} = \phi(x_i - x_{j2}), \quad j, i = 1, \ldots, M$. To avoid the problem of ill-conditioning a relaxation of the interpolation condition can be foreseen [4]. In this case data points and centers of the RBF functions are no longer coincident. The exact problem (16) becomes a problem of linear optimization that can be solved in the least square sense introducing the Moore–Penrose pseudoinverse. A more recent method to improve the conditioning of the interpolation problem is the addition of a multivariate polynomial $p(x)$ leading to

$$y(x) = \sum_{j=1}^{M} w_j \phi(x - x_{j2}) + p(x), \quad x \in \mathbb{R}^p. \tag{18}$$

Problem (18) is underdetermined and we have to impose the orthogonality condition [8]

$$\sum_{j=1}^{M} w_j p(x_j). \tag{19}$$

4 Applications

In this section applications of the surrogate model are presented for two- and three-dimensional problems in the case of transonic and subsonic flows. The parameter spaces are composed by shape variables, flow properties as the Mach number, or airfoil characteristics as the angle of attack. The initial collocation of the snapshots is performed using full and factorial planes or quadtree distribution.

4.1 NACA 0012 Surrogate Model

A surrogate model is constructed considering the flow around a NACA 0012 airfoil. The model is used to evaluate unknown pressure and velocity fields corresponding to desired parameter configurations in the two-dimensional parameter space composed

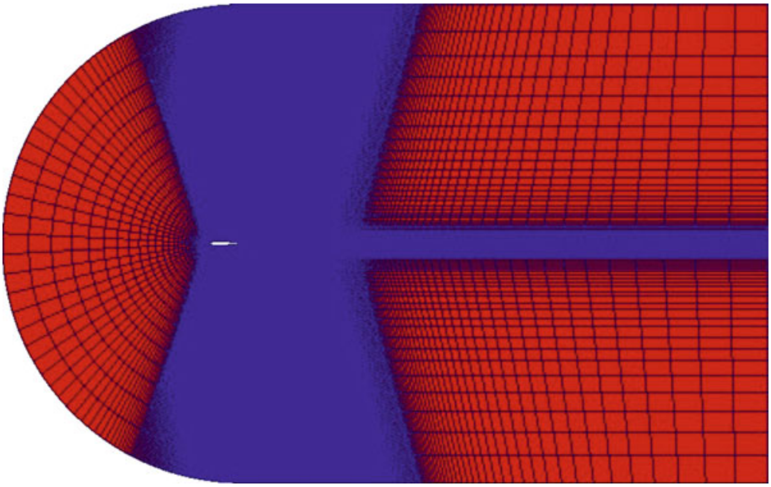


Fig. 1 Problem discretization for the NACA 0012 surrogate model

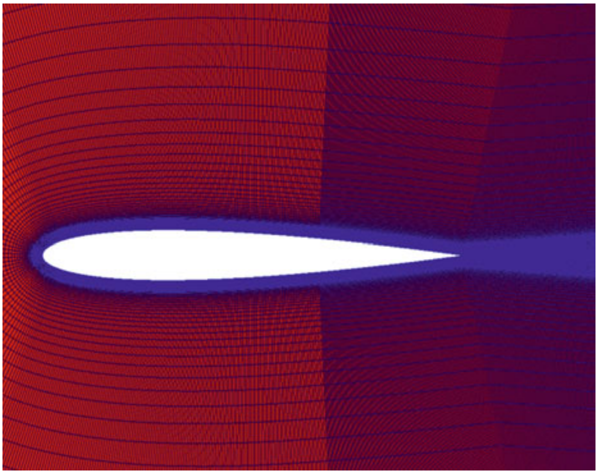


Fig. 2 Problem discretization for the NACA 0012 surrogate model—airfoil zoom

by the Mach angle of the undisturbed flow M and the angle of attack α of the airfoil. As previously explained the high-order model of the problem are the Navier–Stokes equations. A steady subsonic flow is considered and a grid characterized by a number of points N equal to 260,000 is used to discretize the geometry. To compute the snapshots the SIMPLE algorithm implemented in the open source software OpenFOAM is used. To model the turbulence a Spalart–Allmaras equation is solved and wall functions are used to estimate the boundary layer quantities near the wall.

In Fig. 1 a representation of the problem discretization can be seen and in Fig. 2 a zoom near the airfoil is shown.

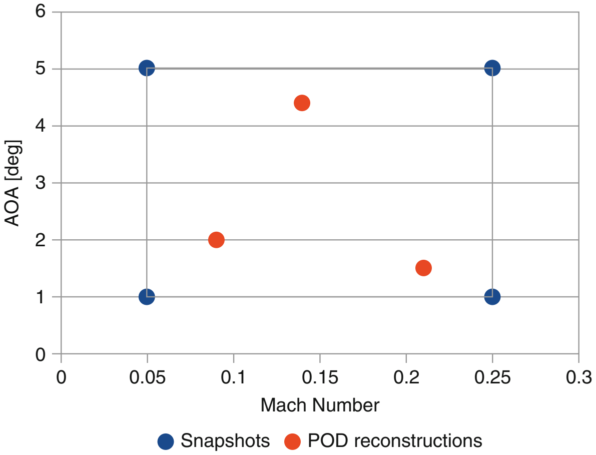


Fig. 3 Snapshot and PODI reconstruction positions in the parameter space

4.1.1 Interpolation Technique: Preliminary Phase

A first section is dedicated to the choice of the right interpolation technique for the computation of the PODI coefficients $\widetilde{a_i}^{\delta_l}$. For this purpose a snapshot set $\{\boldsymbol{u}^k\}_{k=1}^M$ composed by only four snapshots is used. The snapshots are placed at the four vertices of the parameter space and this configuration corresponds to a two-level full factorial plane.

In Fig. 3 a representation of the snapshot position in the parameter space can be seen. Three points are randomly chosen to test the surrogate model and decide which interpolation technique has to be preferred for the model. The NACA 0012 surrogate model is only a preliminary application and this number can be increased in practical cases. It is necessary to consider that higher is the reconstruction number in this first section, higher will be the total computational cost to build the surrogate model. A trade-off between accuracy and computational resources, in terms of hardware and time, shall be taken into account. If an a priori knowledge of the problem is available, it is possible to start testing the model in portions of the parameter space that will be more interesting or critical from the design point of view.

In Fig. 3 the three test points can be seen in red. In Table 1 snapshots and test points are reported.

Four techniques are tested to compute the PODI reconstruction coefficients $\widetilde{a_i}^{\delta_l}$: a least square regression of the first order (LS), radial basis functions using Gaussian basis with and without polynomial term and using multiquadric basis considering the polynomial term. A comparison of the results can be seen in Fig. 4.

In this figure, only the error on the maximum and minimum values of pressure and velocity fields associated with the three reconstruction points 1, 2, and 3 are considered.

Table 1 Snapshot and POD reconstruction positions in the parameter space

Snapshot	*Mach number*	*Alpha* (°)
1	0.05	1
2	0.25	1
3	0.05	5
4	0.25	5
PODI reconstruction	*Mach number*	*Alpha* (°)
1	0.09	2
2	0.21	1.5
3	0.14	4.4

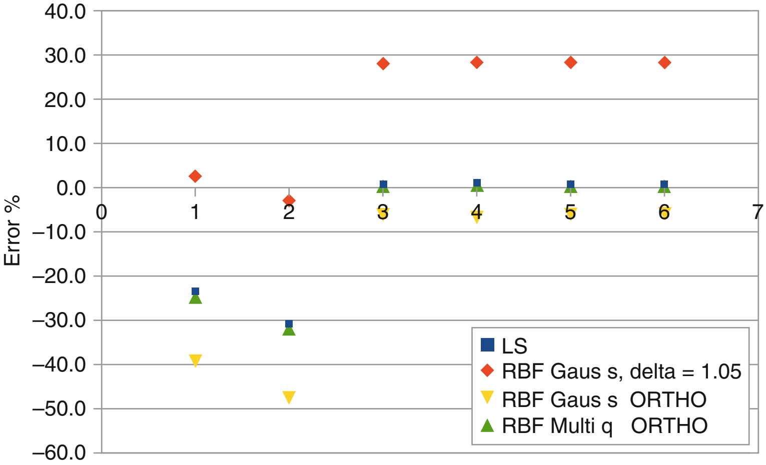


Fig. 4 NACA0012 surrogate model. Errors for different response surfaces (x = 1: error on the maximum value of the pressure field, x = 2: error on the minimum value of the pressure field, x = 3 and 4: errors on maximum and minimum values of the x component of the velocity field, x = 5 and 6: errors on maximum and minimum values of the y component of the velocity field)

The error $e_{\%}$ is calculated with respect to the CFD computation, taken as reference or "true" solution and is computed as

$$e_{\%} = \frac{x - x_{POD}}{x} \cdot 100. \tag{20}$$

The best results for the reconstruction of the pressure field are obtained using a radial basis function with Gaussian basis, no relaxation of the interpolation condition and a value of the shape parameter σ of 1.05. In this application a constant parameter σ is chosen for all the different radial basis functions therefore $\sigma_i \equiv \sigma, i = 1, \ldots, M$. The behavior of the response surfaces is different in the PODI reconstruction of the velocity field. In this case the lower error is obtained using a multiquadric basis, with the orthogonality condition. These interpolation methods will be used for the airfoil surrogate models built in the following section to reconstruct pressure and velocity fields.

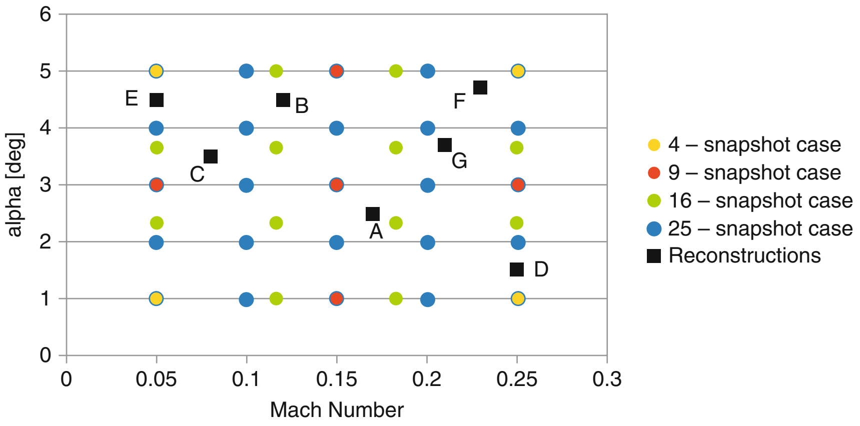


Fig. 5 NACA0012 surrogate model. Snapshot sets and reconstruction positions in the (Mach-α) plane

Table 2 NACA0012 surrogate model

Reconstruction point	Mach number	α (°)
A	0.17	2.5
B	0.12	4.5
C	0.08	3.5
D	0.25	1.5
E	0.05	4.5
F	0.23	4.7
G	0.21	3.7

List of the reconstruction combinations

4.1.2 Influence of the Number of Snapshots

Once the best interpolation technique has been selected for the PODI coefficients, an analysis has been performed to understand the influence of the number of snapshots on the accuracy of the surrogate model. In this section four PODI-surrogate models are tested using 4, 9, 16, and 25 snapshots corresponding to a 2, 3, 4, and 5 level full factorial design for the two parameters Mach and α.

With respect to the surrogate model of the previous section, the parameter ranges are extended: the Mach number now is varying from 0.05 to 0.25 and α is between 1° and 5°. In Fig. 5 a visualization of the snapshot positions is shown. The surrogate models are used to reconstruct pressure and velocity fields in seven random points in the parameter space. The positions of the reconstructed cases in the (Mach-α) plane are summarized in Table 2.

In this case the error $E\%$ with respect to the CFD reference solution is computed using the L_2 norm

$$E\% = \|x - x_{POD}\|_2 \cdot 100 = \frac{\sqrt{\sum_{i=1}^{N}(x_i - x_i^{POD})^2}}{N} \cdot 100, \qquad (21)$$

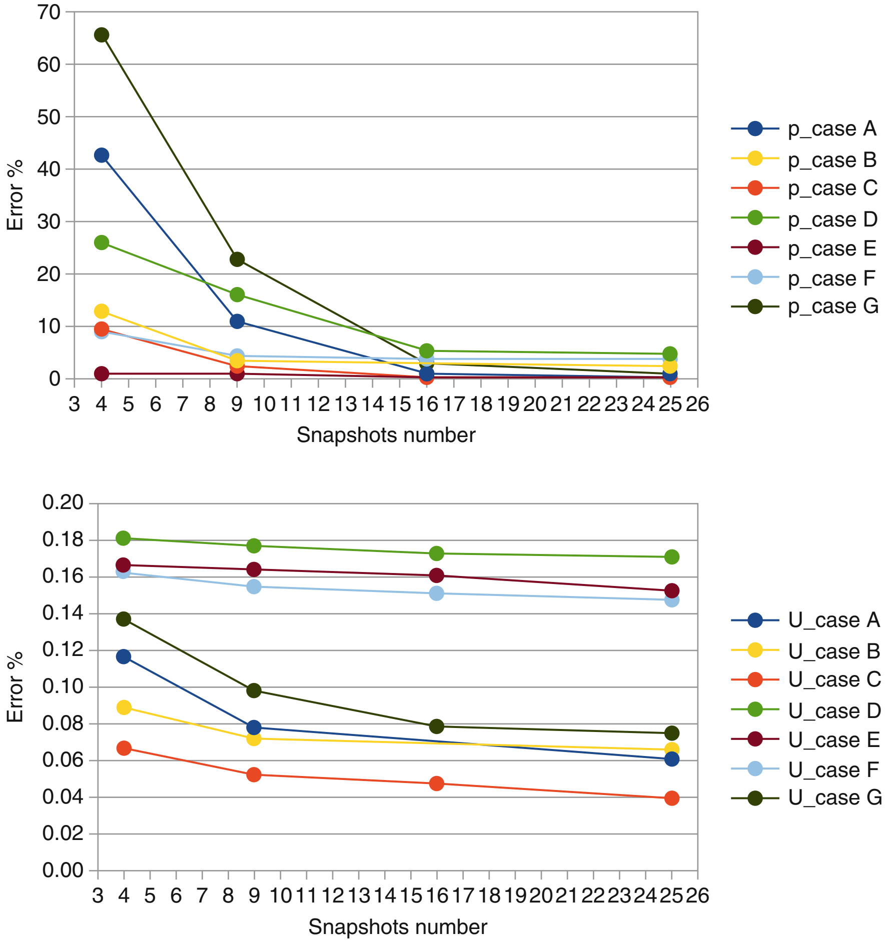


Fig. 6 NACA0012 surrogate model. Errors generated by the PODI-surrogate model for the reconstruction of the pressure (*top*) and velocity (*bottom*) fields

where with x is indicated the CFD value of the field of interest in a single cell and with x^{POD} the corresponding value obtained with PODI. In Fig. 6, the error trends obtained with Eq. (21) are shown for the pressure and velocity fields.

As expected the errors are decreasing with the increasing of the snapshot number. Considering the velocity field, the error is already under 2 % using four snapshots and is slowly decreasing. Therefore, depending on the a priori threshold of the surrogate model error, the use of 25 snapshots can be avoided and 9 or 16 snapshot sets can be used. For the pressure field higher errors are generated, but again the use of the 16-snapshot set can fulfill accuracy requirements. In Fig. 7 a visualization of the velocity field is shown for the full model and for the 25, 16, and 9 snapshot cases for point C.

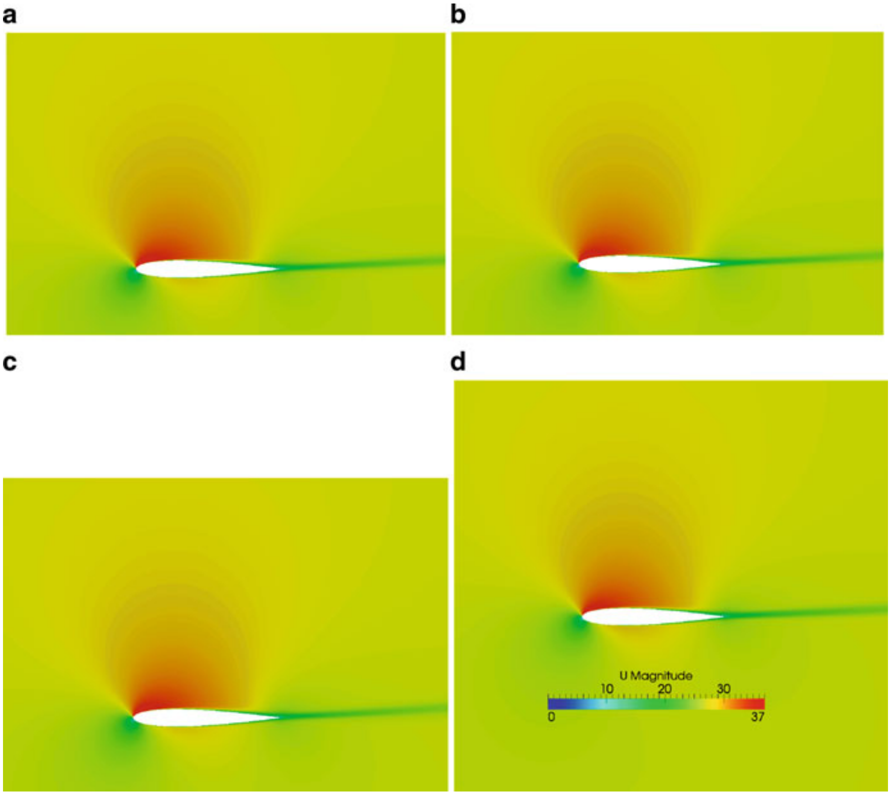


Fig. 7 NACA0012 surrogate model. Comparison between velocity fields. $M = 0.08$, $\alpha = 3.5°$. (**a**) PODI-surrogate model using 25 snapshots. (**b**) PODI-surrogate model using 16 snapshots. (**c**) PODI-surrogate model using nine snapshots. (**d**) CFD solution

4.1.3 Influence of the Snapshot Position: Quadtree Initial Distribution

Quadtree is a specific subdivision of two-dimensional spaces used first in digital imaging. It can be considered as a hierarchical data structure [3]. The subdivision is described by the so-called tree data structure that is an iterative splitting of the area of interest. At the beginning we can consider the whole space as a cell. This cell has four vertices at its corners. In the first iteration the single cell is divided into four sub-cells and in the following iterations each cell can be divided again in other four sub-cells.

In Fig. 8 the quadtree approach is visually explained. Examples of a quadtree distribution applied to the initial snapshot sampling for the building of a surrogate model can be found in [2]. In the present work no leave-one-out procedure is present because we are using a reduced number of snapshots. In Figs. 9 and 10 a comparison between five-level full factorial and quadtree distribution can be made and in Fig. 11 the results are reported.

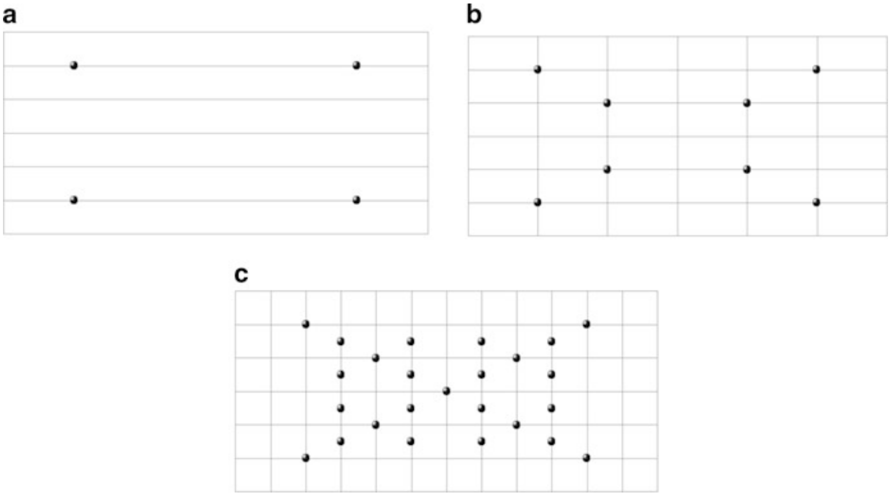
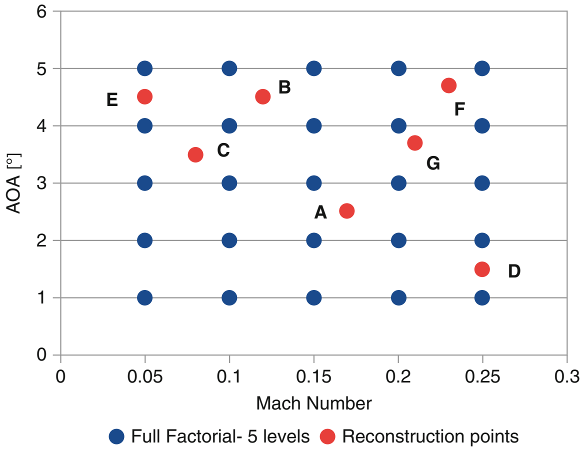


Fig. 8 Iterations of a quadtree distribution. (**a**) Start of the quadtree distribution. (**b**) First iteration of the quadtree distribution. (**c**) Second iteration of the quadtree distribution

Fig. 9 NACA0012 surrogate model. Snapshot and reconstruction positions in the five-level full factorial case



The quadtree distribution is able to reduce the surrogate model errors in all the seven test points. This can be obvious for the internal points B, C, A, and F, since in the quadtree distribution the snapshots are now nearer to the reconstruction point but is not trivial for the other points D, E, and G. In this three points, with respect to the five-level case, the snapshots in the quadtree distribution are farther but still the errors are decreasing.

Finally, in Fig. 12, visualizations of the reconstructed fields compared with the CFD high-order solution can be seen.

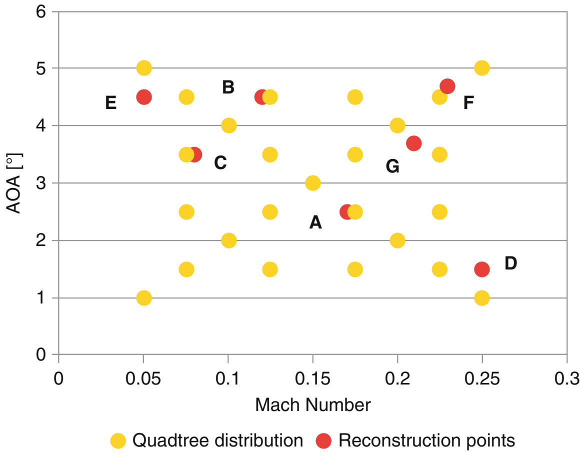


Fig. 10 NACA0012 surrogate model. Snapshot and reconstruction positions in the quadtree case

4.2 *RAE 2822 Surrogate Model*

In this section the generation of a surrogate model for a supercritical airfoil is performed. The flow past a transonic wing profile is complex and highly nonlinear and particular care must be taken in the shock wave treatment.

4.2.1 Problem Setting

The two-dimensional transonic flow field past the supercritical airfoil RAE 2822 is analyzed. A surrogate model is built in the shape parameter space. The airfoil is modeled using two Bézier curves [6], one for the upper surface and one for the lower surface of the airfoil. Each Bézier curve is computed using nine control points as can be seen in Fig. 13.

The high-order simulations are performed considering a flow field characterized by a Mach number of 0.729 and a Reynolds number of $6.5 \cdot 10^6$. The airfoil has an angle of attack equal to 2.31°.

To construct the parameter space of the surrogate model, only five control points are considered. The points P0, P1, P7, and P8 of Fig. 13 are fixed to maintain a constant positions of leading and trailing edges. Points P2, P3, P4, P5, and P6 instead are moved along the y-axis, normal to the flow field. The parameters of the surrogate model are the positive or negative variations of the point positions along the y-axis with respect to the base configuration, within a range of ±2 % of the chord. The parameter space is five-dimensional and we are affected by the curse of dimensionality: if a two-level full factorial plane is required for a preliminary screening of the space, 2^5 simulations should be performed. In order to reduce the number of initial simulations required for the construction of the surrogate model, a two-level fractional factorial plane 2^{5-1} is adopted to compute the snapshots. In this

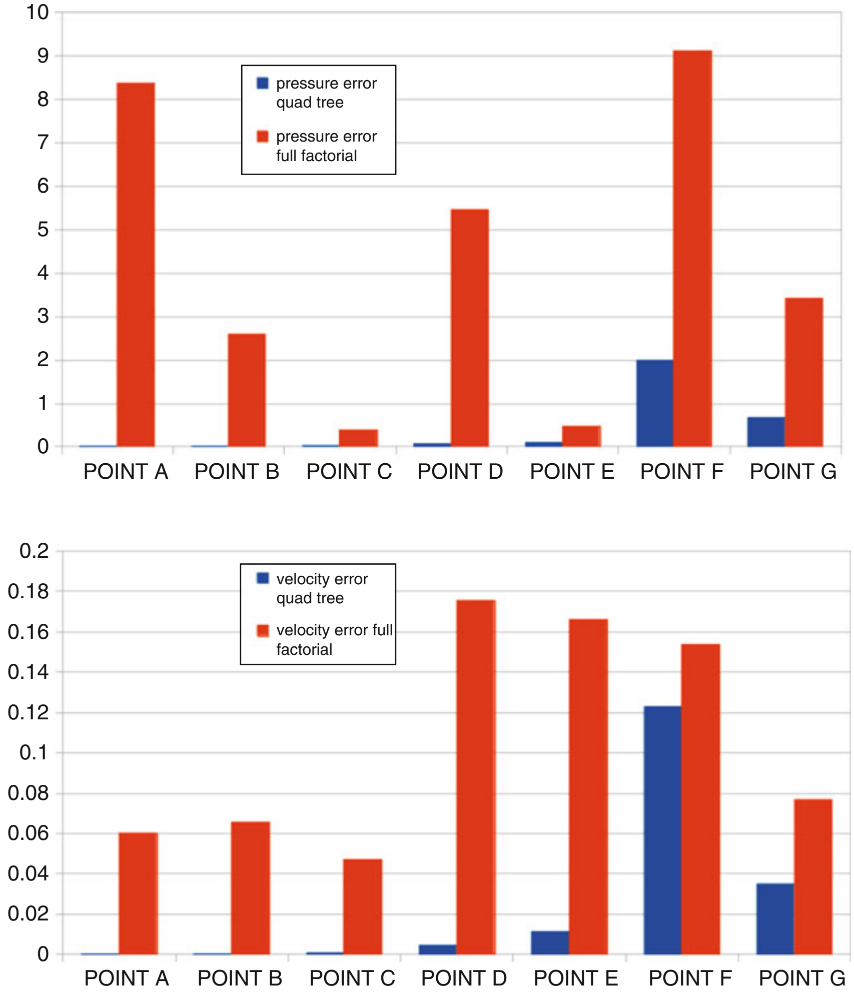


Fig. 11 NACA0012 surrogate model. Comparison of the surrogate model errors on the field reconstruction using quadtree or five-level full factorial distribution. Pressure (*top*) and velocity (*bottom*) fields

way only 16 high-order model evaluations are required instead of 32. Fractional factorial planes are a useful instrument typically applied in design of experiment (DOE) methodology. They allow to investigate the response with respect to multiple parameters with a reduced number of samples [9]. In Table 3 the 16 parameter combinations associated with each snapshot are reported.

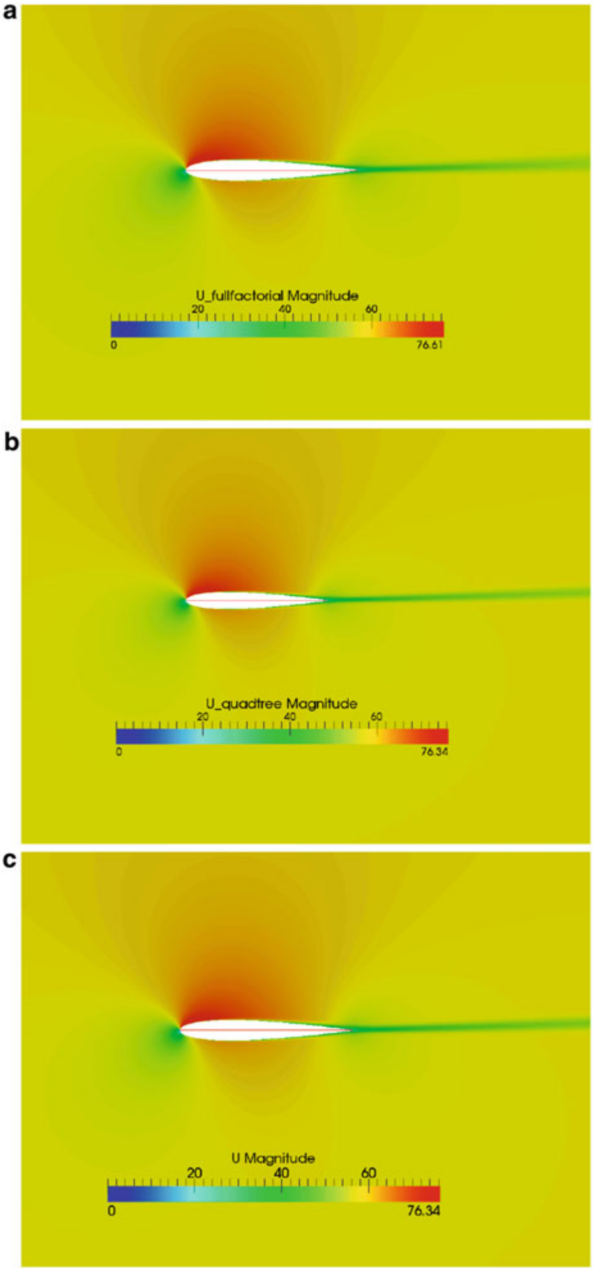


Fig. 12 Velocity fields, m/s. (**a**) Velocity field obtained with the reduced-order model using a five-level full factorial distribution. (**b**) Velocity field obtained with the reduced-order model using a quadtree distribution. (**c**) Velocity field obtained with the CFD full model

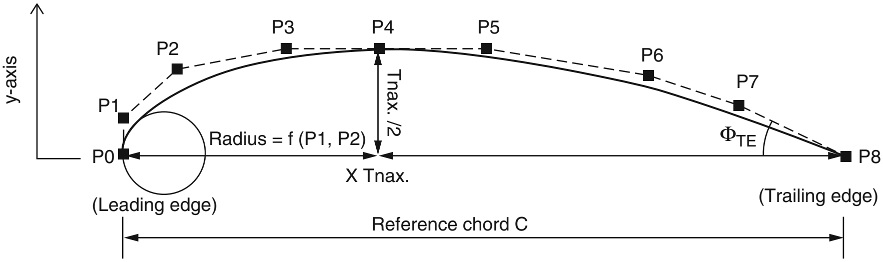


Fig. 13 RAE 2822 surrogate model. Bézier curve for the upper surface of the airfoil. Visualization of the nine control points

Table 3 RAE 2822 surrogate model

Snapshot number	$\Delta P2$	$\Delta P3$	$\Delta P4$	$\Delta P5$	$\Delta P6$
1	−0.2	−0.2	−0.2	−0.2	−0.2
2	0.2	−0.2	−0.2	−0.2	−0.2
3	−0.2	0.2	−0.2	−0.2	−0.2
4	0.2	0.2	−0.2	−0.2	0.2
5	−0.2	−0.2	0.2	−0.2	−0.2
6	0.2	−0.2	0.2	−0.2	0.2
7	−0.2	0.2	0.2	−0.2	0.2
8	0.2	0.2	0.2	−0.2	−0.2
9	−0.2	−0.2	−0.2	0.2	−0.2
10	0.2	−0.2	−0.2	0.2	0.2
11	−0.2	0.2	−0.2	0.2	0.2
12	0.2	0.2	−0.2	0.2	−0.2
13	0.2	0.2	−0.2	−0.2	−0.2
14	0.2	−0.2	0.2	0.2	−0.2
15	−0.2	0.2	0.2	0.2	−0.2
16	0.2	0.2	0.2	0.2	0.2

List of the parameter combinations associated with the snapshots

4.2.2 Results

In this application the POD surrogate model is used to reconstruct the conservation variables (density ρ, momentum $\rho\boldsymbol{u}$, and total energy ρE) and the coordinates of the grid points. A mesh morphing is used to build only the initial snapshots, for the remaining reconstruction points the computational mesh is computed with the POD/ROM. The reconstructed fields are associated with shape parameter combinations not belonging to the initial set of snapshots. Four random points, listed in Table 4, are chosen to test the ROM. The reference solution is obtained through CFD simulation using the Alenia solver UNS3D. With respect to this solution, the normalized root mean square error E_n is computed for each reconstructed field as

Table 4 RAE 2822 surrogate model

Point number	$\Delta P2$	$\Delta P3$	$\Delta P4$	$\Delta P5$	$\Delta P6$
1	0.2	0	0	0	0
2	0.1	0.1	0.1	0.1	0.1
3	0.2	0.1	0.2	0.1	0.2
4	0.15	−0.1	0.15	−0.1	0.15

List of the four reconstruction points

Table 5 RAE 2822 surrogate model

	Point 1	Point 2	Point 3	Point 4
Field variable	E_n %	E_n %	E_n %	E_n %
ρ	0.4	1.2	0.5	0.6
ρu	1.0	3.6	1.3	1.1
ρv	5.4	23.9	8.8	6.9
ρE	0.5	1.2	0.5	0.5

Errors on the reconstruction of the four test points

$$E_n = \frac{\sqrt{N \sum_{i=1}^{N} (x_i - x_i^{POD})^2}}{\sum_{i=1}^{N} x_i} \cdot 100, \tag{22}$$

where N is the number of grid points, x the CFD value of a cell, and x^{POD} the corresponding value obtained with the surrogate model. In Table 5 this error is listed for the four points.

The values are high considering the transverse momentum especially for point 2. In Figs. 14, 15, 16, and 17, qualitative comparisons are reported considering the derived fields of Mach and pressure.

A partial agreement with the reference solution is achieved. A relative error e is calculated considering the computation of lift and drag coefficients c_L and c_D:

$$e = \frac{|x_{CFD} - x_{POD}|}{x_{CFD}} \cdot 100 \tag{23}$$

and the results are listed in Table 6. If we take into account only drag and lift coefficient reconstructions, considering, for example, a link of the surrogate model to an optimization procedure, the errors e are lower than the root mean square computation E_n and a good agreement with the CFD result is obtained.

4.3 3D Aircraft Surrogate Model

The last application deals with the construction of a POD surrogate model for a three-dimensional aircraft configuration. The two-dimensional parameter space is composed by the angle of attack α and the sideslip angle β of the aircraft. In Fig. 18 a geometry visualization is reported.

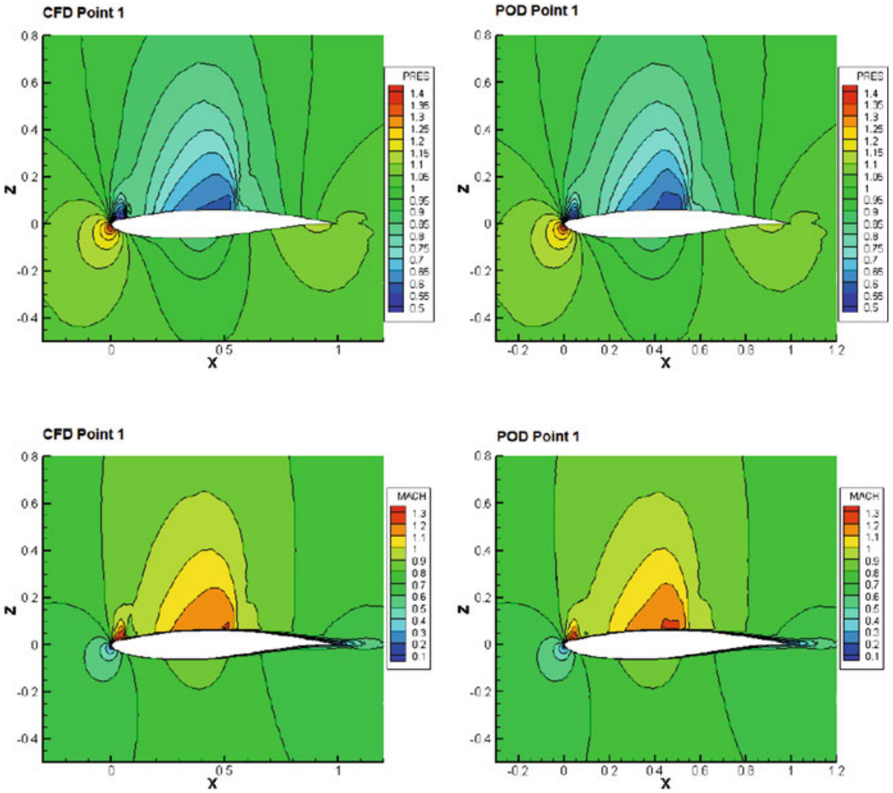


Fig. 14 RAE 2822 surrogate model. CFD and POD/ROM comparison for reconstruction point 1: pressure (*top*) and Mach (*bottom*) fields

4.3.1 Problem Setting

A subsonic flow characterized by a Mach number of 0.25 and a Reynolds number of $4 \cdot 10^6$ is considered. The computational domain is composed by 10^6 points and can be seen in Fig. 19. The Alenia UNS3D solver is used for the CFD simulations.

The POD/ROM is used to build a database for all the aircraft operative configurations therefore the parameter space is large: the angle of attack α varies between 0° and 14°, the sideslip angle β is between 0° and 6°. Taking into account the good performances of the previous subsonic application, a quadtree distribution is exploited to position the snapshots composing the initial high-order simulations. A visualization of the snapshot distribution is presented in Fig. 20.

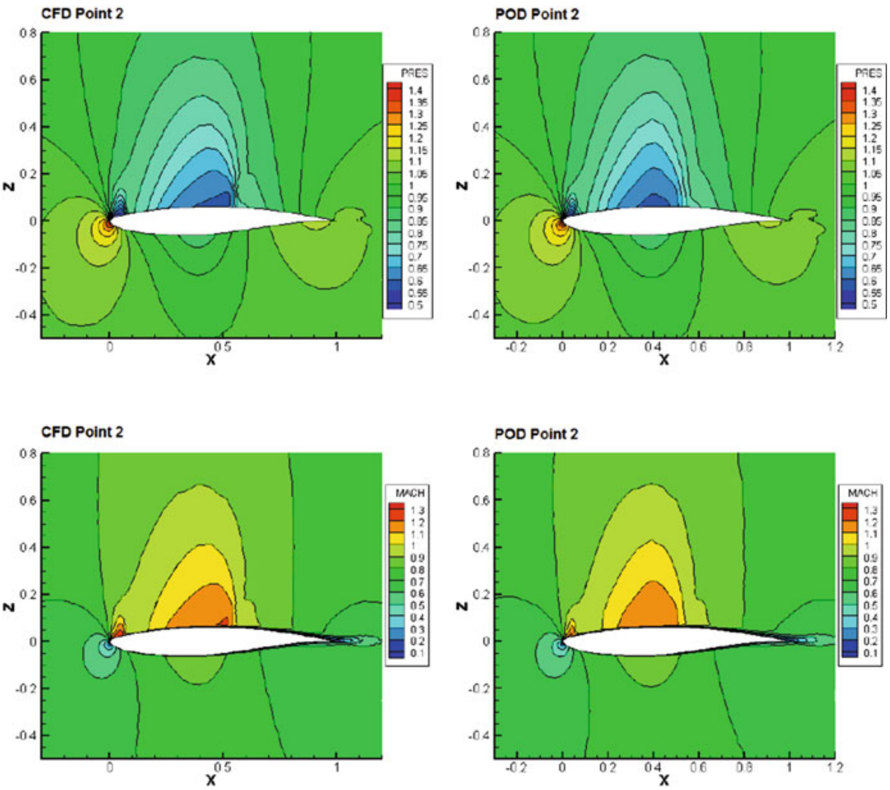


Fig. 15 RAE 2822 surrogate model. CFD and POD/ROM for reconstruction point 2: pressure (*top*) and Mach (*bottom*) fields

4.3.2 Results

20 random points are chosen to test the surrogate model. Lift and drag coefficients are computed starting from the fields of the conservation variables generated using the surrogate model. The results are listed in Table 7. The relative error *e* % is computed following Eq. (23).

A good agreement with the CFD solution is obtained for the lift coefficient calculation. On the other hand, the drag coefficient comparison is characterized by larger errors. In Fig. 21, the points corresponding to higher errors are visualized and it can be remarked that they belong to an "outer area" of the parameter space.

In the inner area underlined in Fig. 22 a good agreement with the reference solution is obtained considering both lift and drag coefficient computations. For a further application of the POD/ROM to database generation therefore, this distinction between outer and inner area of the parameter space can be taken into account and the use of the surrogate model for the internal area is recommended.

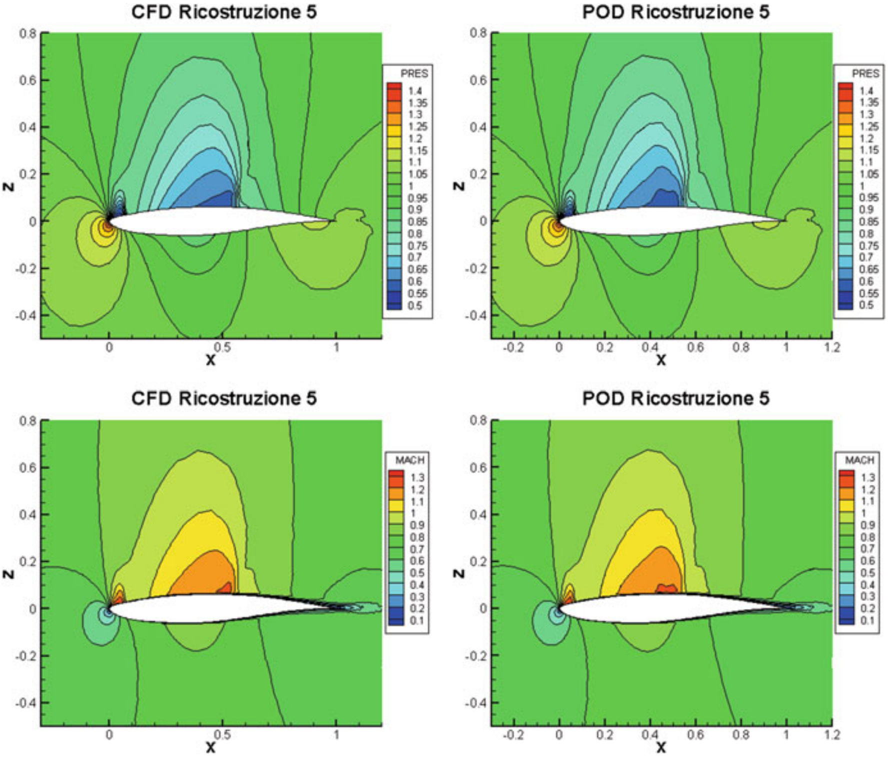


Fig. 16 RAE 2822 surrogate model. CFD and POD/ROM comparison for reconstruction point 3: pressure (*top*) and Mach (*bottom*) fields

5 Conclusions

In this work a surrogate model using the proper orthogonal decomposition has been described. The method of snapshots was adopted and POD was applied in the parameter space. Interpolation techniques were exploited to compute the expansion coefficients. For the initial position of the snapshots full and fractional factorial planes together with quadtree distribution were tested.

Three applications were presented. In the first one a surrogate model has been constructed to analyze the subsonic two-dimensional flow past a NACA 0012 airfoil. The Mach number of the undisturbed flow and the angle of attack of the airfoil were chosen as parameters. The second application was in the transonic field. Five shape parameters composed the parameter space used to build the surrogate model. A fractional factorial plane was adopted to generate the initial snapshot set. Finally in the last test case a three-dimensional aircraft was considered in an extended parameter space and a quadtree snapshot distribution was adopted. In all the three applications the surrogate model with a reduced number of snapshots

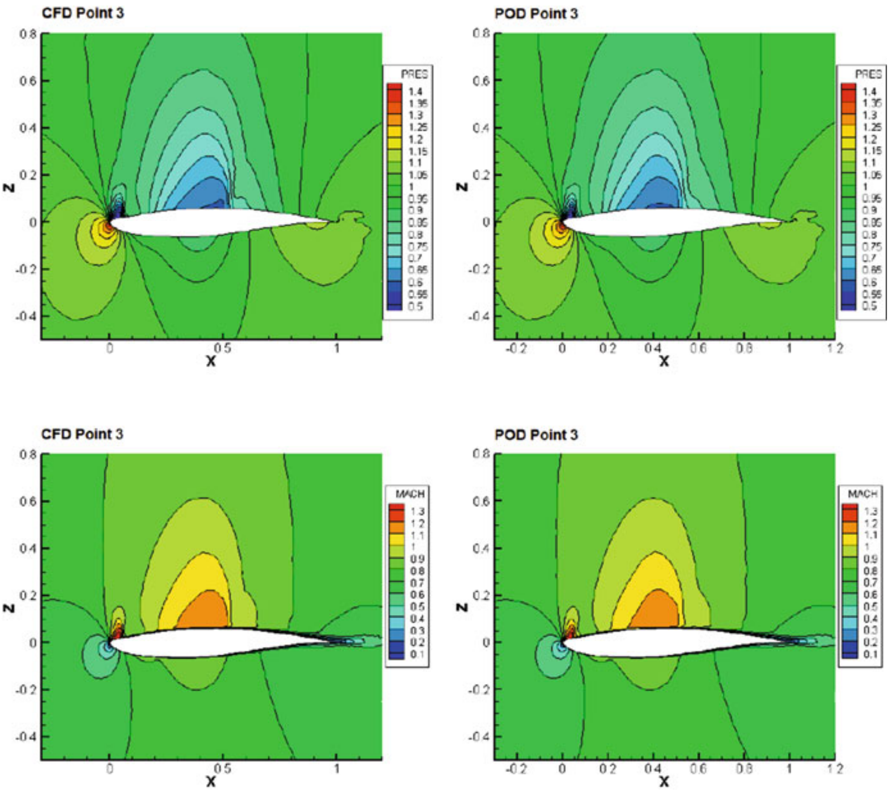


Fig. 17 RAE 2822 surrogate model. CFD and POD/ROM comparison for reconstruction point 4: pressure (*top*) and Mach (*bottom*) fields

Table 6 RAE 2822 surrogate model

	e % (c_L)	e % (c_L)
Point 1	0.3	0.1
Point 2	0.2	1.5
Point 3	0.4	0.4
Point 4	1.0	2.6

Errors on the lift and drag coefficients calculation

showed good agreement with the CFD reference solution. Further investigations could be interesting especially in the transonic case, testing, for example, different initial snapshot distributions.

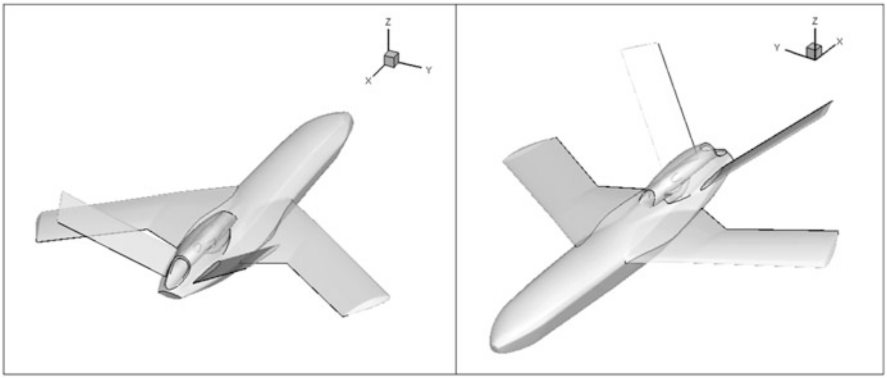


Fig. 18 Aircraft surrogate model. Visualization of the aircraft geometry

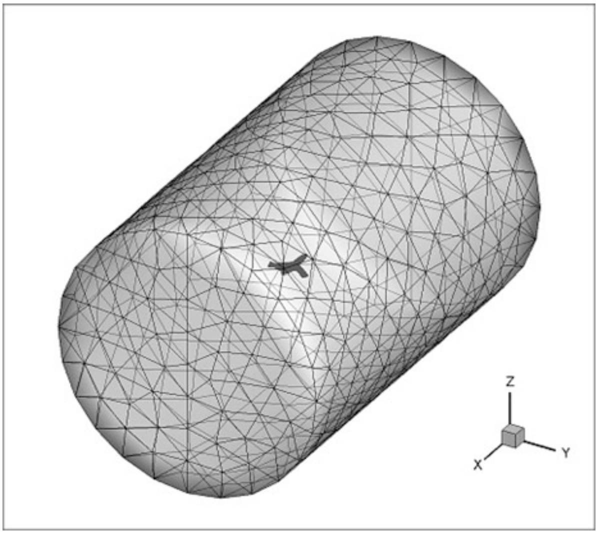


Fig. 19 Aircraft surrogate model. Visualization of the computational domain

Acknowledgements The authors want to thank Dott. Nicola Ceresola from Alenia Aermacchi and Ing. Giuseppe Rizza for their contribution to the computations of the test cases of Sects. 4.2 and 4.3.

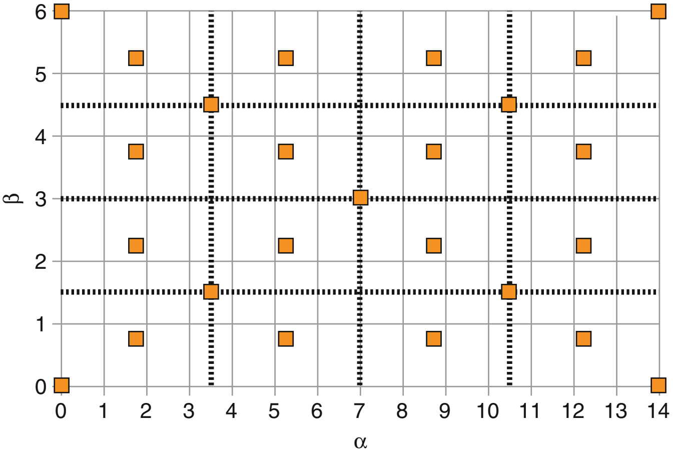


Fig. 20 Aircraft surrogate model. Snapshot distribution

Table 7 Aircraft surrogate model

Test point	α	β	e % (c_L)	e % (c_D)
Point 1	4.7	5.8	0.02	11.22
Point 2	3.4	0.3	0.41	11.21
Point 3	7.6	2.2	0.08	1.75
Point 4	10.1	5.2	0.03	1.80
Point 5	12.0	1.4	0.43	0.44
Point 6	3.6	5.3	0.24	9.7
Point 7	8.4	4.4	0.04	0.07
Point 8	10.9	1.4	0.14	0.53
Point 9	12.7	1.4	0.09	1.73
Point 10	1.2	4.9	5.26	7.54
Point 11	4.6	5.7	0.003	8.32
Point 12	4.0	4.2	0.12	1.40
Point 13	12.7	1.7	0.15	1.53
Point 14	8.0	1.0	0.13	0.71
Point 15	6.7	0.7	0.001	2.91
Point 16	8.6	4.9	0.09	0.36
Point 17	1.3	5.1	6.19	4.03
Point 18	12.2	3.8	0.07	0.18
Point 19	3.1	4.0	0.17	0.18
Point 20	7.1	0.5	0.09	4.32

Results of the 20 random points

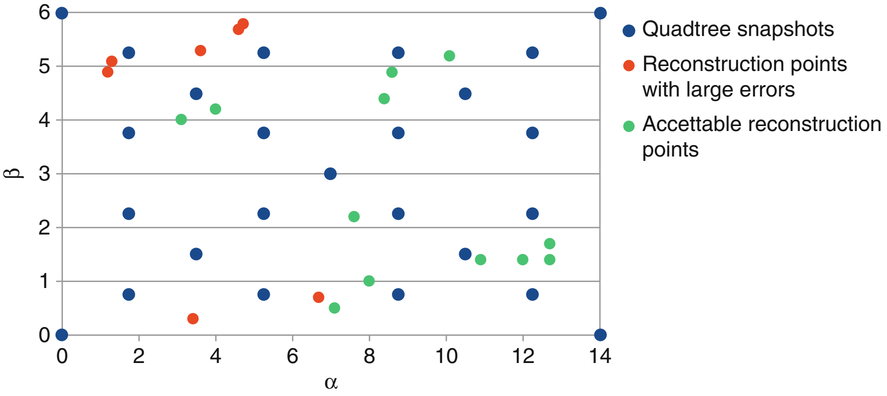


Fig. 21 Aircraft surrogate model. Large error points distribution

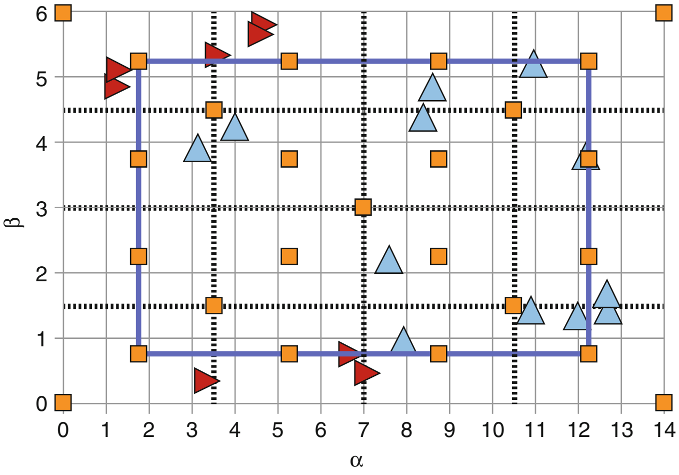


Fig. 22 Aircraft surrogate model. Inner area of the parameter space

References

1. Alonso, J.J., LeGresley, A.P.: Airfoil design optimization using reduced order models based on proper orthogonal decomposition. In: AIAA 2000 2545 (2000)
2. Bracconnier, T., et al.: Toward an adaptive POD/SVD surrogate model for aeronautic design. Comput. Fluids **40**, 195–209 (2011)
3. Breene, L.A.: Quadtrees and hypercubes: grid embedding strategies based on spatial data structure addressing. Comput. J. **36**(6), 562–569 (1993)
4. Broomhead, D.S., Lowe, D.: Multivariable functional interpolation and adaptive networks. Complex Syst. **2**, 321–355 (1988)
5. Buhmann, M.D.: Radial Basis Functions: Theory and Implementations. Cambridge University Press, Cambridge (2003)
6. Farin, G.E., Hansford, D.: The Essentials of CAGD. AK Peters, Ltd., Natick (2000)
7. Han, Z.H., Zimmermann, R., Görtz, S.: A new cokriging method for variable-fidelity surrogate modeling of aerodynamic data. In: 48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, Orlando (2010)
8. Mifsud, M.: Reduced-order modelling for high-speed aerial weapon aerodynamics. Ph.D. Thesis, Cranfield University (2009)
9. Myers, R.H., Montgomery, D.C., Anderson-Cook, C.M.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley, New York (2009)
10. Othman, N., Kanazaki, M.: Surrogate model of aerodynamic model toward efficient digital flight. In: Asia-Pacific International Symposium on Aerospace Technology (2014)
11. Sirovich, L.: Turbulence and the dynamics of coherent structures, part 1: coherent structures. Q. Appl. Math. **45**(3), 561–571 (1998)
12. Tan, B.T.: Proper orthogonal decomposition extensions and their applications in steady aerodynamics. Master Thesis, High Performance Computation for Engineered Systems, Singapore-MIT Alliance (2003)

Trust Region Filter-SQP Method for Multi-fidelity Wing Aerostructural Optimization

Ali Elham and Michel J.L. van Tooren

Abstract A trust region filter-SQP method is used for wing multi-fidelity aerostructural optimization. Filter method eliminates the need for a merit function, and subsequently a penalty parameter. Besides, it can easily be modified to be used for multi-fidelity optimization. A low fidelity aerostructural analysis tool is presented, that computes the drag, weight, and structural deformation of lifting surfaces as well as their sensitivities with respect to the design variables using analytical methods. That tool is used for a mono-fidelity wing aerostructural optimization using a trust region filter-SQP method. In addition to that, a multi-fidelity aerostructural optimization has been performed, using a higher fidelity CFD code to calibrate the results of the lower fidelity model. In that case, the lower fidelity tool is used to compute the objective function, constraints, and their derivatives to construct the quadratic programming subproblem. The high fidelity model is used to compute the objective function and the constraints used to generate the filter. The results of the high fidelity analysis are also used to calibrate the results of the lower fidelity tool during the optimization. This method is applied to optimize the wing of an A320 like aircraft for minimum fuel burn. The results showed about 9 % reduction in the aircraft mission fuel burn.

1 Introduction

According to industry criteria for aircraft design, the drag prediction accuracy using numerical methods should be within one drag count (one, ten, thousandth of the drag coefficient) [23]. To confirm a need for such a level of accuracy, Meredith [17] showed that one drag count is equal to the weight of one passenger for a long-haul aircraft. Similarly, 1 % error in wing structural weight estimation of the same class

A. Elham (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: a.elham@tudelft.nl

M.J.L. van Tooren
University of South Carolina, Columbia, SC, USA
e-mail: vantooren@cec.sc.edu

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_10

aircraft is equal to the weight of four to five passengers. The need for such a high level of accuracy forces the designers to use high fidelity, physics based analysis for aerostructural analysis, design, and optimization of aircraft. The traditional design methods based on empirical, statistic based methods do not satisfy the required level of accuracy.

On the other hand, using high fidelity methods for aerostructural optimization requires massive computational power [10, 15], that is a serious barrier against using high fidelity aerostructural optimization in early design stages. As a solution multi-fidelity optimization techniques are used to keep the level of accuracy similar to the results of the high fidelity analysis methods, while reduce the computational cost of the optimization. Alexandrov et al. [1] presented a model management framework for multi-fidelity aerodynamic shape optimization of lifting surfaces based on a trust region algorithm. In that model a lower fidelity tool is used for the optimization, while a higher fidelity tool is occasionally, but systematically, called to monitor the optimization process. March and Willcox [13] suggested a multi-fidelity optimization framework based on a trust region algorithm, in which the gradient of the objective function is computed using the low fidelity model, but the algorithm is provably converges to the solution of the high fidelity model. The same algorithm is used by Elham [3] for aerodynamic shape optimization of lifting surfaces, where an adjoint quasi-three-dimensional (Q3D) model is used for prediction of the wing drag and its derivatives, and a three-dimensional CFD tool is used to calibrate the results of the Q3D model.

Besides the framework for model management in a multi-fidelity optimization, the choice of a proper algorithm for numerical optimization is important. Aerostructural optimization of lifting surfaces, or the whole aircraft in general, involves hundreds to thousands of design variables, and tens to hundreds of constraints. Besides, computing the objective function and the constraints required execution of CFD and FEM analysis, which takes considerable amount of time. Therefore the gradient based optimization algorithms are the most efficient algorithms for solving such problems [16]. In an optimization using a gradient based algorithm, in order to achieve a quadratic rate of convergence, an underlying Newton iteration is required, which is the basics of sequential quadratic programming (SQP). The SQP algorithms are based on iteratively solving a quadratic model of the objective function and linear models of the constraints. The SQP approach has been used in both line search and trust region frameworks [18]. Many of the SQP methods, such as SNOPT [9], use a merit function, which combines the objective function and the constraints. An ℓ_1 penalty function and the augmented Lagrangian function are popular choices for the merit function. Selection of a proper merit function and the method used for updating the penalty parameter is a challenge [18]. Fletcher and Leyffer [7] proposed a method to eliminate the need for a merit function in an SQP algorithm. They suggested a *filter* that rejects the unacceptable solutions. In that so-called *trust region filter-SQP algorithm*, no merit function is constructed and the filter is applied to the objective function and the (norm of the) constraints separately.

In this research a modified version of the filter method proposed by Fletcher and Leyffer is applied to multi-fidelity aerostructural optimization of lifting surfaces.

In the next section, the aerostructural analysis tool used in this research is briefly explained. Then in Sect. 3, the trust region filter-SQP method is discussed in details. Eventually in Sect. 4 the filter method is applied to both a mono-fidelity and a multi-fidelity wing aerostructural optimization.

2 Aerostructural Analysis

The FEMWET tool presented by Elham and van Tooren [5] is used for wing aerostructural analysis. FEMWET is based on a Q3D aerodynamic analysis method, which is combined with a finite beam element model of the structure. In the Q3D approach an inviscid vortex lattice method (VLM) is combined with a viscous 2D airfoil analysis code for prediction of the wing total viscous drag. The idea behind the Q3D approach is to avoid using a high fidelity 3D CFD solver, but still predict drag with the same level of accuracy. In the proposed method the wing drag is decomposed into the induced drag and the parasite drag. To compute the wing total drag, first a VLM is executed to compute the lift distribution over the wing as well as the wing induced drag using a Trefftz plane analysis. The VLM code is connected to an FEM to automatically deform the VLM mesh based on the wing structural deformation. Then several sections along the span are analyzed using a 2D airfoil analysis CFD code. The parasite drag is computed based on the pressure and friction drag of the 2D sections. In order to perform the 2D analysis, several corrections including the corrections due to sweep, induced angle of attack, and wing structural deformation are applied to the section geometry as well as the flow conditions. Details of this method are presented in reference [5].

The wing box structure is modeled using four equivalent panels: the upper panel, including the wing box upper skin, stringers, and spars caps; the lower panel including the lower skin, stringers, and spar caps; and two vertical panels including the front and the rear spars webs. For finite element analysis of the wing box structure an equivalent Timoshenko beam is placed at the shear center of the wing box, see Fig. 1. Using this FEM model different failure criteria, referred to tension, compression, Euler buckling, and shear buckling in several wing box elements, located in (y_e, z_e) distance from the shear center (see Fig. 1), are calculated. For more details one can refer to [5].

When the Q3D aerodynamic solver is combined with the finite beam element model, four governing equations appear as follows:

$$R_1(X, \varGamma, U, \alpha, \alpha_i) = AIC\,\varGamma - RHS = 0 \tag{1}$$

$$R_2(X, \varGamma, U, \alpha, \alpha_i) = KU - F = 0 \tag{2}$$

$$R_3(X, \varGamma, U, \alpha, \alpha_i) = L - nW_{des} = 0 \tag{3}$$

$$R_4(X, \varGamma, U, \alpha, \alpha_i) = C_{l_{2d}} - C_{l_{vlm}} = 0 \tag{4}$$

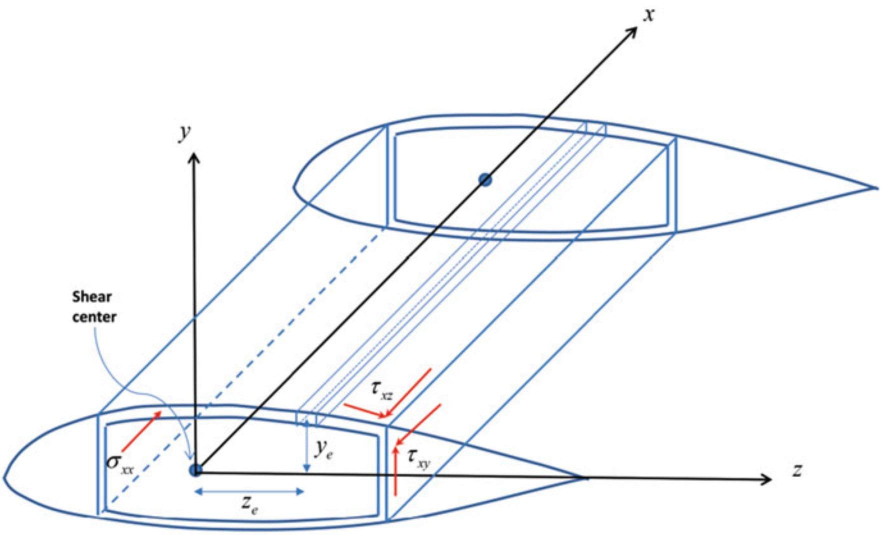


Fig. 1 Wing box panels element position w.r.t. the shear center

The first equation is the governing equation of the VLM and the second equation is the governing equation of the FEM. The third equation in fact indicates that the lift should be equal to the design weight multiplied by the load factor. The last equation is required to guarantee that the lift predicted by the VLM is the same as the lift computed by 2D section analysis at effective angle of attack. The effective angle of attack is the angle of attack that a 2D local section feels. Therefore the local downwash angle is required to compute the effective angle of attack. In such a system four sets of state variables are presented: the strengths of vortices in the VLM (Γ), the displacements in FEM (U), the global angle of attack (α), and the local downwash angles at each 2D section (α_i). For a given vector of design variables, X, the system of governing equations are solved using the Newton method:

$$\underbrace{\begin{bmatrix} \frac{\partial R_1}{\partial \Gamma} & \frac{\partial R_1}{\partial U} & \frac{\partial R_1}{\partial \alpha} & \frac{\partial R_1}{\partial \alpha_i} \\ \frac{\partial R_2}{\partial \Gamma} & \frac{\partial R_2}{\partial U} & \frac{\partial R_2}{\partial \alpha} & \frac{\partial R_2}{\partial \alpha_i} \\ \frac{\partial R_3}{\partial \Gamma} & \frac{\partial R_3}{\partial U} & \frac{\partial R_3}{\partial \alpha} & \frac{\partial R_3}{\partial \alpha_i} \\ \frac{\partial R_4}{\partial \Gamma} & \frac{\partial R_4}{\partial U} & \frac{\partial R_4}{\partial \alpha} & \frac{\partial R_4}{\partial \alpha_i} \end{bmatrix}}_{J} \begin{bmatrix} \Delta\Gamma \\ \Delta U \\ \Delta\alpha \\ \Delta\alpha_i \end{bmatrix} = - \begin{bmatrix} R_1(X, \Gamma, U, \alpha, \alpha_i) \\ R_2(X, \Gamma, U, \alpha, \alpha_i) \\ R_3(X, \Gamma, U, \alpha, \alpha_i) \\ R_4(X, \Gamma, U, \alpha, \alpha_i) \end{bmatrix} \tag{5}$$

To perform the Newton iteration, the matrix of the partial derivatives J is required. All the elements of that matrix are computed by a combined use of analytical methods and automatic differentiation (AD). The Matlab toolbox Intlab [21] is used for AD. More details of the sensitivity analysis are presented in [5]. In order to compute the sensitivity of the output (e.g., wing drag or structural failure loads) the coupled adjoint method [11] is used, where the total derivative of any function of interest I with respect to a design variable x is presented as follows:

$$\frac{dI}{dx} = \frac{\partial I}{\partial x} - \lambda_1^T\left(\frac{\partial R_1}{\partial x}\right) - \lambda_2^T\left(\frac{\partial R_2}{\partial x}\right) - \lambda_3^T\left(\frac{\partial R_3}{\partial x}\right) - \lambda_4^T\left(\frac{\partial R_4}{\partial x}\right) \tag{6}$$

in which $\lambda = [\lambda_1 \;\; \lambda_2 \;\; \lambda_3 \;\; \lambda_4]^T$ is the adjoint vector and computed using the following equation:

$$\begin{bmatrix} \frac{\partial R_1}{\partial \Gamma} & \frac{\partial R_1}{\partial U} & \frac{\partial R_1}{\partial \alpha} & \frac{\partial R_1}{\partial \alpha} \\ \frac{\partial R_2}{\partial \Gamma} & \frac{\partial R_2}{\partial U} & \frac{\partial R_2}{\partial \alpha} & \frac{\partial R_2}{\partial \alpha} \\ \frac{\partial R_3}{\partial \Gamma} & \frac{\partial R_3}{\partial U} & \frac{\partial R_3}{\partial \alpha} & \frac{\partial R_3}{\partial \alpha} \\ \frac{\partial R_4}{\partial \Gamma} & \frac{\partial R_4}{\partial U} & \frac{\partial R_4}{\partial \alpha} & \frac{\partial R_4}{\partial \alpha} \end{bmatrix}^T \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} \frac{\partial I}{\partial \Gamma} \\ \frac{\partial I}{\partial u} \\ \frac{\partial I}{\partial \alpha} \\ \frac{\partial I}{\partial \alpha_i} \end{bmatrix} \tag{7}$$

The FEMWET tool is also able to compute the aileron effectiveness and its sensitivity with respect to the design variables. The aileron effectiveness is defined as $L_{\delta_{\text{elastic}}}/L_{\delta_{\text{rigid}}}$. The parameter L_δ is the derivative of the wing rolling moment w.r.t the aileron deflection angle. This parameter is an important requirement for aircraft performance and is strongly affected by the wing stiffness. Low wing (mainly torsional) stiffness may result in a poor roll performance or even aileron reversal. Therefore for a wing aerostructural optimization a constraint on the aileron effectiveness seems necessary.

Elham and van Tooren [5] performed some analysis and aerostructural optimization to verify the results of the FEMWET tool. In order to verify the accuracy of the wing drag prediction, the results of Q3D aerodynamic solver were compared to the results of a higher fidelity CFD code called MATRICS-V [25]. The MATRICS-V flow solver is based on fully conservative full potential outer flow in quasi-simultaneous interaction with an integral boundary layer method on the wing. The MATRICS-V tool was developed by NLR and has been validated using wind tunnel test as well as the flight test results for Fokker 100 aircraft, see Figs. 2 and 3. Therefore the drag of the Fokker 100 wing predicted by the Q3D solver was compared to the drag predicted by MATRICS-V, see Fig. 4. From this figure one can observe a high level of accuracy for drag prediction using the Q3D solver.

In order to validate the wing weight and structural deformation predicted by FEMWET, Elham and van Tooren [5] performed an aeroelastic wingbox optimization of Airbus A320-200 wing. The optimization was subject to five different load cases to evaluate the structural failure with respect to tensile and compressive loads, buckling, fatigue, and aileron effectiveness. The total wing weight predicted by FEMWET is equal to 8791 kg, which is very close to the actual wing weight of A320-200 equal to 8801 kg [19]. The wing twist deformation predicted by FEMWET was compared to the actual wing twist deformation under 1 g load (obtained from [19]). The results show an acceptable level of accuracy for the tool, see Fig. 5.

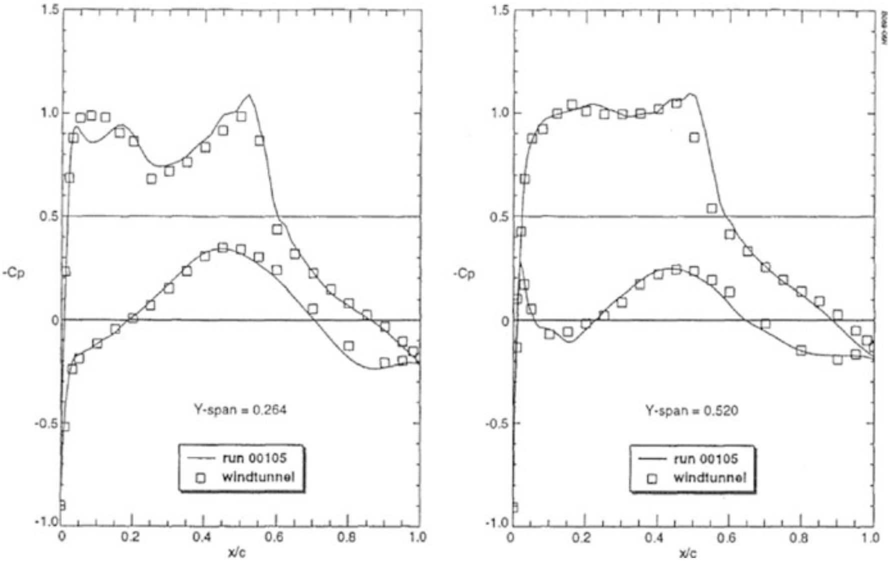


Fig. 2 Comparison of MATRICS-V and wind tunnel measured chordwise pressure distribution on two wing sections of Fokker 100 wing/body configuration at $M_\infty = 0.779$, $\alpha = 1.03°$, $Re_\infty = 3 \times 10^6$ source: NLR [25]

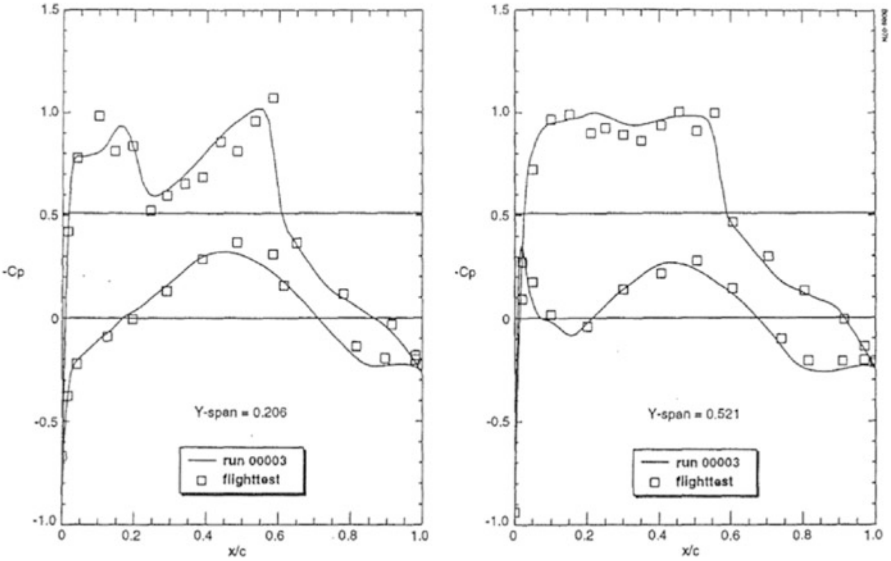


Fig. 3 Comparison of MATRICS-V and in flight measured chordwise pressure distribution on two wing sections of Fokker 100 wing/body configuration at $M_\infty = 0.775$, $\alpha = 1.0°$, $Re_\infty = 35 \times 10^6$ source: NLR [25]

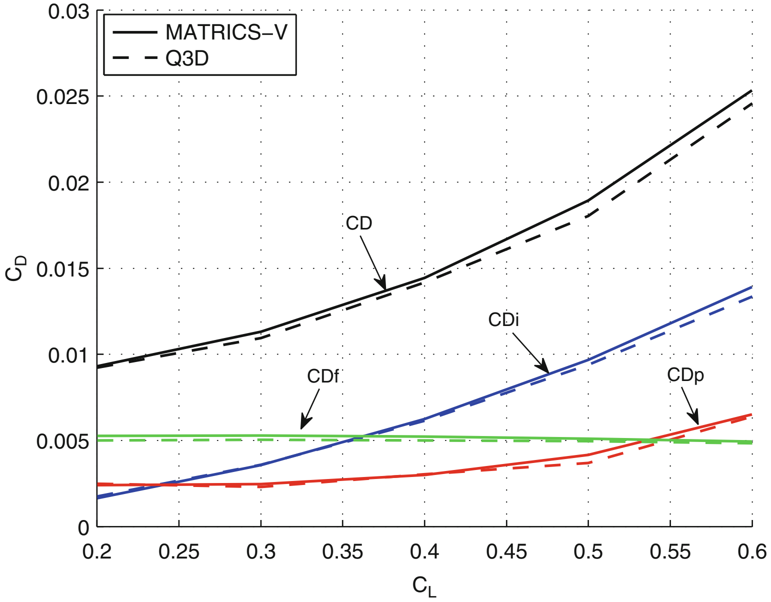


Fig. 4 Comparison of computed drag by the MATRICS-V and Q-3D solvers for cruise condition (1 g loaded wing and M = 0.75) [5]

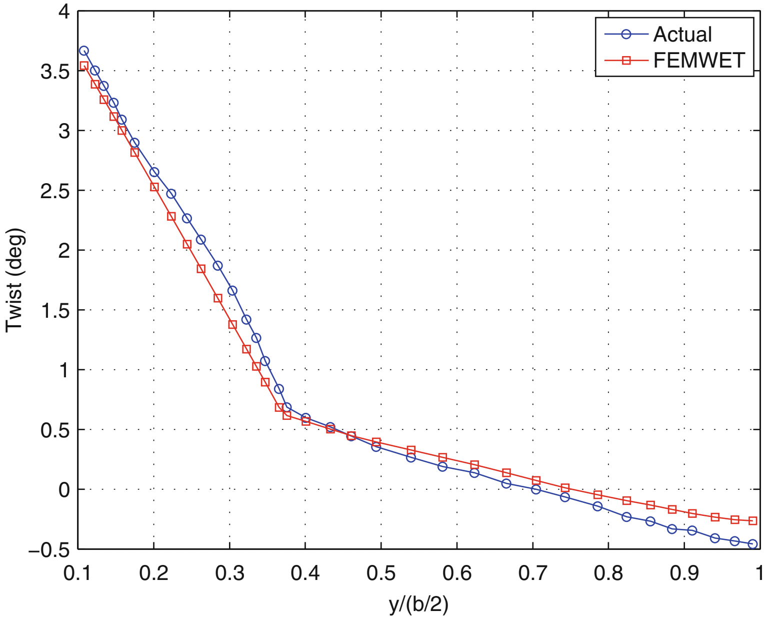


Fig. 5 A320-200 wing twist under 1 g load [5]

3 Trust Region Filter-SQP Method

A general optimization problem is defined as follows:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ s.t. \quad & c(x) = 0 \end{aligned} \tag{8}$$

Only equality constraints are considered here, however, inequality constraints can be easily defined as equality constraints using slack variables and adding simple bounds. Using an SQP approach, Eq. (8) is solved by solving the following quadratic problem (QP) iteratively:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\, s^T H(x_k)\, s + g(x_k)^T s \\ s.t. \quad & c(x_k) + A(x_k)s = 0 \end{aligned} \tag{9}$$

where H is the Hessian matrix of the Lagrangian function, g and A are the gradient of the objective function and the constraints, respectively. Most of the available SQP algorithms use the solution of the QP as a search direction, then find a step length that minimizes a one-dimensional problem, which results in a sufficient decrease of a merit function. The merit function combines the objective function and the constraints in a single function. One-dimensional optimization algorithms (such as polynomial interpolation) [24] or line search algorithms [18] are used to find the step length. An alternative to this approach is the use of trust region algorithms [2]. Trust region methods define a region around the current point, where the approximations of the objective function and constraints are trusted. The radius of the trust region plays the role of the step length, so the trust region algorithms find the search direction and the step length simultaneously. However the need for a merit function is still there.

As mentioned earlier, definition of a merit function and consequently a method for updating an associated penalty parameter is a challenge. Some aspects of the difficulties associated with the choice of the merit function and the penalty parameter are discussed in [7]. The filter method presented by Fletcher and Leyffer [7] eliminates the need for a merit function in an SQP algorithm. The concept of the filter is based on the two goals of a constrained optimization problem: minimizing the objective function and minimizing the constraint violation. So in the method proposed by Fletcher and Leyffer a filter is used to only accept the solutions that are not dominated by the Pareto front between the objective function and the constraint violation, see Fig. 6. If ϑ is the 2-norm of the (equality) constraints, the pair (f_1, ϑ_1) is said to dominate (f_2, ϑ_2) if and only if both $f_1 \leq f_2$ and $\vartheta_1 \leq \vartheta_2$. Defining $\mathscr{F}$ as the filter, that includes a set of pairs (f_j, ϑ_j) such that no pair dominate any other, a pair (f, ϑ) is acceptable to $\mathscr{F}$ if it is not dominated by any pair in the filter.

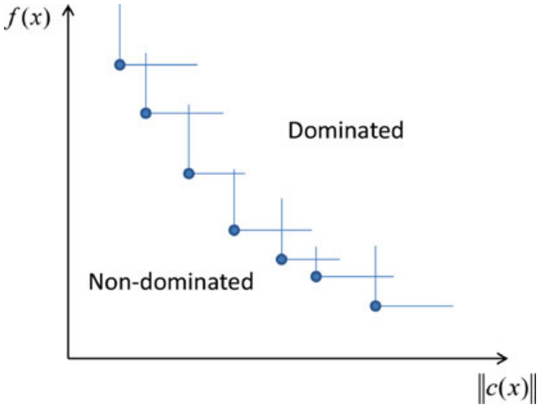


Fig. 6 Dominated and non-dominated points according to the filter

Algorithm 1 Basic trust region filter-SQP algorithm

```
1: Choose $x_0$, $\Delta_0$ and set $k = 0$
2: Solve the trust region quadratic problem (TRQP):
```

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\, s^T H(x_k)\, s + g(x_k)^T s \\ s.t. \quad & c(x_k) + A(x_k)s = 0 \\ & \|s\| \leq \Delta_k \end{aligned}$$

```
3: if TRQP is infeasible perform a constraint restoration and go to 2, otherwise continue.
4: set $x_{k+1} = x_k + s$
5: if $(f(x_{k+1}), \vartheta(x_{k+1}))$ is acceptable to the filter, then accept $x_{k+1}$, add $(f(x_{k+1}), \vartheta(x_{k+1}))$ to
   the filter, remove the dominated points from the filter and increase the trust region radius if
   possible. Else reject $x_{k+1}$ and reduce the trust region radius.
6: if the solution is not converged go to 2.
```

In a trust region filter-SQP method, the QP shown in Eq. (9) is solved within a trust region, then the solution is checked by the filter. If the filter rejects the solution, the radius of the trust region is reduced. A common problem in trust region methods is that the QP may have no feasible solution if the radius of the trust region is small. In such cases a constraint *restoration* is required. The idea of constraint restoration is to minimize $\vartheta(x)$ starting from the current iteration. A basic trust region filter-SQP algorithm is shown in Algorithm 1.

In order to prove the convergence of the trust region filter-SQP algorithm, a small envelope is required around the current filter, in which no point is accepted. This envelope in fact enforces a sufficient decrease in the objective function and the constraint. According to Fletcher et al. [8] a point is acceptable to the filter if:

$$\text{either} \quad f \leq f_j - \gamma\vartheta_j \quad \text{or} \quad \vartheta \leq \beta\vartheta_j \quad \text{for all j} \in \mathscr{F} \tag{10}$$

The proof of convergence of such an algorithm is given in [8]. A more refined trust region filter-SQP method is presented by Conn et al. [2], although in that algorithm a composite step optimization is used, where first in a normal step the norm of the constraints is minimized within the trust region and then in a tangential step the objective function is reduced. In this research the algorithm presented in [2] is modified to use the original SQP method presented by Fletcher and Leyffer [7] instead of a composite step optimization. This algorithm is presented in Algorithm 2.

Conn et al. [2] suggested the following values for the constants in Algorithm 2:

$$\gamma_0 = 0.5 \quad \gamma_1 = 2 \quad \eta_1 = 0.01 \quad \eta_2 = 0.9$$
$$\gamma_\vartheta = 10^{-4} \quad k_\vartheta = 10^{-4}$$

Algorithm 2 Trust region filter-SQP method

```
1: Choose $x_0$, $\Delta_0$, $\Delta_{max}$, $\Delta_{min}$, $\eta_1$, $\eta_2$, $\gamma_0$, $\gamma_1$, $\gamma_\vartheta$, $k_\vartheta$
2: Initialize the Hessian as $H_0 = I$
3: Compute the value and the gradient of the objective function and the constraints.
4: Solve the trust region quadratic programming (TRQP) to find $s$:

       minimize  $m_k(s) = \frac{1}{2} s^T H(x_k) s + g(x_k)^T s$
       $s.t.$    $c(x_k) + A(x_k)s = 0$
                 $|s| \leq \Delta_k$

5: If the TRQP does not have a feasible solution, then solve the restoration problem:

       minimize  $\vartheta(x_k + s) \equiv \|c(x_k + s)\|$

6: Evaluate the objective function and the constraints at $x_k + s$.
7: Check if the new point is acceptable to the filter, i.e., if:

       $F(x_k + s_k) < F_j - \gamma_\vartheta \vartheta(x_k + s_k)$  or  $\vartheta(x_k + s_k) < (1 - \gamma_\vartheta)\vartheta_j$  forall $(F_j, \vartheta_j) \in \mathscr{F}$

8: If $x_k + s_k$ is not acceptable to the filter, then set $x_{k+1} = x_k$ and $\Delta_{k+1} = min(\Delta_{min}, \gamma_0 \Delta_k)$ and
   go to 4.
9: If $x_k + s_k$ is acceptable to the filter and:

       $m_k(x_k) - m_k(x_k + s_k) < k_\vartheta \vartheta_k^2$  and  $\rho_k \equiv \frac{F(x_k) - F(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} \geq \eta_1$

   then set $x_{k+1} = x_k + s_k$ , update the Hessian using, i.e., the BFGS method and continue, else
   set $x_{k+1} = x_k$ and $\Delta_{k+1} = min(\Delta_{min}, \gamma_0 \Delta_k)$ and go to 4.
10: Add $F_k$ and $\vartheta_k$ to the filter, remove the dominated pairs and update the trust region radius as
    follows:

       $\Delta_{k+1} = \begin{cases} \Delta_k & \text{if} \rho_k \in [\eta_1, \eta_2), \\ max(\Delta_{max}, \gamma_1 \Delta_k) & \text{if} \rho_k \geq \eta_2. \end{cases}$

11: If the optimization has not converged to go 3.
```

In some of the trust region methods, as in the one suggested by Conn et al. [2], a ℓ_2 trust region subproblem is used. However in the filter method suggested by Fletcher and Leyffer [7] a ℓ_∞ trust region subproblem is used to ensures that the subproblem remains tractable as a QP. In this research instead of ℓ_2 or ℓ_∞, the trust region is defined to keep the absolute value of s lower than the trust region radius, see Algorithm 2. It has an advantage when the algorithm is applied to a wing aerostructural optimization. This advantage is explained in Sect. 4. In such an approach the trust region for a 2D case is a rectangle instead of a circle for ℓ_2 or a square for ℓ_∞.

To check the algorithm an analytical optimization test case is used as follows:

$$\begin{aligned}
&min\ e^{x_1x_2x_3x_4x_5}\\
&s.t.\ x_1^2+x_2^2+x_3^2+x_4^2+x_5^2-10=0\\
&\quad x_2x_3-5x_4x_5=0\\
&\quad x_1^3+x_2^3+1=0\\
&\quad lb\le x\le ub\\
&\quad ub=[2.3,2.3,3.2,3.2,3.2],\quad lb=-ub
\end{aligned}\tag{11}$$

The global minimum for this function is $x^*=[-1.717143, 1.595709, 1.827247, -0.7636413, -0.763645], f(x^*)=0.0539498$. The optimization was started from an initial point of $x_0=[1,1,1,1,1]$ and the proposed filter method found the optimum design vector of $x^*=[-1.7171, 1.5957, 1.8273, -0.7636, -0.7636]$ and $f(x^*)=0.0539$. The history of the objective function and norm of the constraints are shown in Fig. 7.

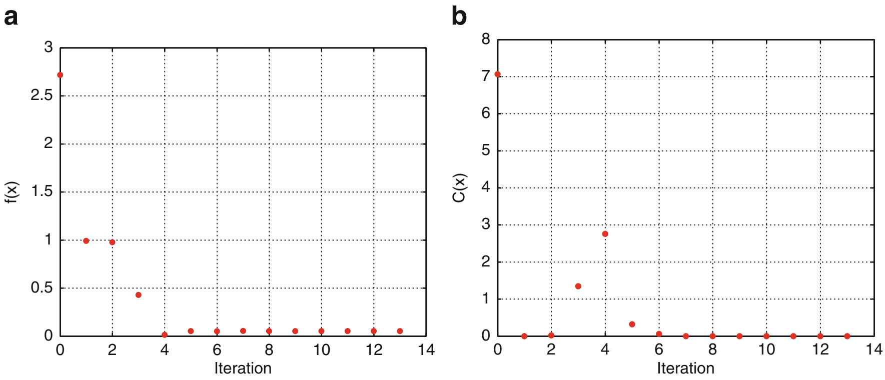


Fig. 7 History of the analytical test case optimization. (**a**) Objective function. (**b**) Maximum constraint violation

4 Wing Aerostructural Optimization

4.1 Problem Formulation

An A320 like aircraft wing is considered as a test case, see Fig. 8. The aircraft mission fuel weight is considered as the objective function. It is computed using the method presented by Roskam [20]. In that method the fuel weight of the cruise phase of the flight is computed using the Breguet method, and the fuel weights of the other segments of the mission, e.g., take-off, climb, etc., are computed using some statistical factors. In order to use the Breguet equation, the aircraft lift over drag ratio during the cruise is required. The lift and drag of the elastic wing during the cruise is computed using the FEMWET for an average aircraft all up weight equal to $W_{ave} = \sqrt{MTOW \times (MTOW - W_{fuel})}$ as suggested by Torenbeek [22]. The drag of the other aircraft components such as fuselage and tail is assumed to be constant.

The design variables are categorized into four groups. The first group includes the design variables describing the wing planform geometry. Six design variables are used for that purpose: the wing root chord C_r, span b, taper ratio λ, leading edge sweep angle Λ, twist angle at kink ϵ_k, and twist angle at tip ϵ_t. The wing aileron geometry was defined using three parameters. The start and end position of the aileron was fixed at 75 and 95 % of the wing span. The aileron chord was fixed as 25 % of the local wing chord. The second group of design variables defines the wing airfoil shape. The airfoil shapes at eight wing spanwise positions are parameterized using Chebyshev polynomials. 160 design variables, shown as G in Eq. (12), are used to control the airfoils shapes. The third group includes the design variables defining the wing box structure. The thickness of the wing box four equivalent panels in 10 spanwise positions are defined as design variables, 40 in total. As mentioned earlier the aircraft fuel weight is defined as the objective function. The fuel weight is a function of the aircraft total weight, which is a function

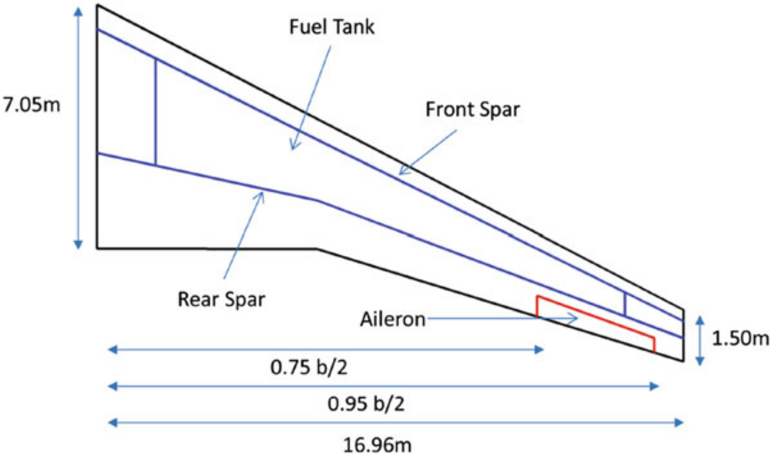


Fig. 8 Planform and wing box dimensions

Table 1 Load cases for wing aerostructural optimization

Load case	Type	H [m]	M	n [g]	q [Pa]
1	Pull up	7500	0.89	2.5	21,200
2	Pull up	0	0.58	2.5	23,900
3	Push over	7500	0.89	−1	21,200
4	Gust	7500	0.89	1.3	21,200
5	Roll	4000	0.83	1	29,700
6	Cruise	11,000	0.78	1	10,650

of the fuel weight. Also for wing structural analysis the aircraft MTOW is required which is a function of the fuel weight and wing structural weight. So in order to avoid any iteration during the optimization two surrogate variables for aircraft fuel weight and maximum take-off weight are added to the design vector as the fourth group of the design variables.

The aerostructural optimization is subject to several constraints. A series of constraints are used to avoid any structural failure. Five load cases are used for wing box structural analysis as shown in Table 1. Load cases number 1 to 3 are used for analyzing the wing box failure under tension, compression, and shear loads. Load case number 4 is used to simulate fatigue in the wing lower panel and the load case number 5 is used to simulate the aileron effectiveness. Load case 6 is used for wing drag prediction during the cruise.

In order to reduce the total number of the constraints on structural failure, the Kreisselmeier–Steinhauser (KS) function [12] is used to aggregate the constraints. All the failure constraints due to the 5 load cases are aggregated into 22 constraints using the KS function. A constraint is defined to keep the aircraft roll moment due to aileron deflection ($L_\delta = dL/d\delta$) higher or equal to the L_δ of the original wing. Another constraint is used to keep the wing loading lower or equal to the wing loading of the initial wing. Finally two consistency constraints are defined for the two surrogate design variables. The wing aerostructural optimization is formulated as follows:

$$\begin{aligned}
min \quad & W^*_{fuel}(X) \\
& X = [C_r, b, \lambda, \Lambda, \epsilon_{kink}, \epsilon_{tip}, G_i, t_{u_j}, t_{l_j}, t_{fs_j}, t_{rs_j}, W^*_{fuel}, MTOW^*] \quad i = 1..160, \;\; j = 1:10 \\
s.t. \quad & \mathrm{KS}_{\mathrm{failure}_k} \leq 0 \quad k = 1..22 \\
& \frac{L_{\delta_0}}{L_\delta} - 1 \leq 0 \\
& \frac{MTOW/S_w}{MTOW_0/S_{w_0}} - 1 \leq 0 \\
& \frac{W_{fuel}}{W^*_{fuel}} - 1 = 0 \\
& \frac{MTOW}{MTOW^*} - 1 = 0 \\
& X_{lower} \leq X \leq X_{upper}
\end{aligned} \tag{12}$$

4.2 Aerostructural Optimization Using the Trust Region Filter-SQP Method

In the first step a mono-fidelity optimization was performed. In that approach the aerostructural analysis method presented in Sect. 2 is used to predict the elastic wing drag and deformation. All the inequality constraints in Eq. (12) are changed to equality constraints by using slack variables.

In the second step a multi-fidelity optimization has been performed. March and Willcox [14] modified the composite step filter method presented by Conn et al. [2] to be used in a multi-fidelity optimization. The same approach as suggested by March and Willcox is used here to modify Algorithm 2 to be used for multi-fidelity optimization. In the modified algorithm, the Q3D aerodynamic analysis (connected with the FEM) is used as the low fidelity model and the MATRICS-V CFD code is used as the high fidelity tool. MATRICS-V is a 3D CFD code, which provides more accurate results than the Q3D method, but the computational time of running MATRICS-V is higher than Q3D. Besides, no analytical sensitivity analysis method is implemented in MATRICS-V. The MATRICS-V has been used for aerodynamic optimization using finite differencing for sensitivity analysis [4] with limited number of design variables. However increasing the number of design variables and coupling the aerodynamic solver with an FEM for fully coupled aerostructural optimization makes the use of finite differencing almost impossible. Therefore in the current research the low-fidelity aerostructural analysis tool is used to generate the TRQP, since that tool can compute the required sensitivities using analytical methods. Then the filter is applied based on the results of the MATRICS-V code. The drag predicted using the low fidelity model is calibrated using the results of the high fidelity model. Three calibration factors are defined for three different drag components (the induced drag, C_{D_i}, the pressure drag, C_{D_p},and the friction drag, C_{D_f}) as follows:

$$\begin{aligned} k_{cd_i} &= \frac{C_{D_{i_{\mathrm{high}}}}}{C_{D_{i_{\mathrm{low}}}}} \\ k_{cd_p} &= \frac{C_{D_{p_{\mathrm{high}}}}}{C_{D_{p_{\mathrm{low}}}}} \\ k_{cd_f} &= \frac{C_{D_{f_{\mathrm{high}}}}}{C_{D_{f_{\mathrm{low}}}}} \end{aligned} \tag{13}$$

After each iteration, if the new point is acceptable to the filter, the calibration factors are updated and the next iteration is performed. This guarantees that at the end of the optimization the drag predicted by the low fidelity model is the same as the drag predicted by the high fidelity model. Since a surrogate variable is used for the aircraft fuel weight, which is the objective function, the variable level of fidelity does not affect the objective function. The value of the object function is always the value of the surrogate fuel weight in the design vector. However the actual fuel weight is computed inside the constraint function in order to generate the consistency constraint on the fuel weight. Therefore two different values of the consistency constraint on the fuel weight and also the MTOW (which is a function of the fuel weight) are computed, one using the drag computed by the low fidelity model and one by using the drag predicted by the high fidelity model. The trust region filter-SQP method algorithm used for such a multi-fidelity optimization is shown in Algorithm 3.

Algorithm 3 Multi-fidelity trust region filter-SQP method

```
1: Choose $x_0$, $\Delta_0$, $\Delta_{max}$, $\Delta_{min}$, $\eta_1$, $\eta_2$, $\gamma_0$, $\gamma_1$, $\gamma_\vartheta$, $k_\vartheta$
2: Initialize the Hessian as $H_0 = I$
3: Compute the value and the gradient of the objective function and the constraints using the low
   fidelity tool.
4: Solve the trust region quadratic programming (TRQP) to find $s$:

       $\text{minimize} \quad m_k(s) = \frac{1}{2} s^T H(x_k)\, s + g(x_k)^T s$
       $s.t. \quad c(x_k) + A(x_k)s = 0$
       $|s| \leq \Delta_k$

5: If the TRQP does not have a feasible solution, then solve the restoration problem:

       $\text{minimize} \quad \vartheta_{low}(x_k + s) \equiv \|c(x_k + s)\|$

6: Evaluate the objective function, the constraints and their derivatives at $x_k + s$ using the low
   fidelity method. Also evaluate the objective function and constraints using the high fidelity
   method at $x_k + s$.
7: Check if the new point is acceptable to the filter, i.e., if:

       $F(x_k + s_k) < F_j - \gamma_\vartheta \vartheta_{high}(x_k + s_k) \quad \text{or} \quad \vartheta_{high}(x_k + s_k) < (1 - \gamma_\vartheta)\vartheta_{high_j} \quad \text{forall } (F_j, \vartheta_{high_j}) \in \mathscr{F}$

   where $\vartheta_{high}$ is the norm of the high fidelity constraints.
8: If $x_k + s_k$ is not acceptable to the filter, then set $x_{k+1} = x_k$ and $\Delta_{k+1} = min(\Delta_{min}, \gamma_0 \Delta_k)$ and
   go to 4.
9: If $x_k + s_k$ is acceptable to the filter and:

       $m_k(x_k) - m_k(x_k + s_k) < k_\vartheta \vartheta^2_{high_k} \quad \text{and} \quad \rho_k \equiv \frac{F(x_k) - F(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} \geq \eta_1$

   then set $x_{k+1} = x_k + s_k$ , update the Hessian using, i.e., the BFGS method, update the
   calibration factors of the low fidelity model based on the results of the high fidelity model
   and continue, else set $x_{k+1} = x_k$ and $\Delta_{k+1} = min(\Delta_{min}, \gamma_0 \Delta_k)$ and go to 4.
10: Add $F_k$ and $\vartheta_{high_k}$ to the filter, remove the dominated pairs and update the trust region radius
    as follows:

       $\Delta_{k+1} = \begin{cases} \Delta_k & \text{if} \rho_k \in [\eta_1, \eta_2), \\ max(\Delta_{max}, \gamma_1 \Delta_k) & \text{if} \rho_k \geq \eta_2. \end{cases}$

11: If the optimization has not converged go to 3.
```

As mentioned before in the TRQP, instead of ℓ_2 or ℓ_∞, a bound around the absolute value of s is defined. So Δ in TRQP is a vector with the same length as s. It helps to define different values of Δ for different design variables. As one can see in Eq. (12) different groups of design variables are used for a wing aerostructural optimization. Although all the design variables are normalized to have the same order of magnitude, defining the same trust region radius for all of them does not seem logical. For example, assuming the trust region radius allows 10 % change in the design variables ($\Delta = 0.1$ for a design vector normalized with the initial values of the design variables). The effect of 10 % change in the thickness of the wing box skin at wing tip on the aircraft fuel weight is not the same as the effect of 10 % change in the aircraft MTOW on the aircraft fuel weight. By defining Δ as a vector, some design variables are allowed to have a larger change and some are more tightly limited.

The history of the mono-fidelity optimization is shown in Fig. 9. Figure 10 shows the history of the multi-fidelity optimization. Figure 11 shows the planform of the initial wing compared to the planform of the optimized wing using both the mono-fidelity and the multi-fidelity optimizations. The shape of the airfoils in several spanwise position and the pressure distribution on those airfoils are shown in Figs. 12 and 13, respectively. The characteristics of the optimized aircraft are summarized in Table 2.

From Table 2 one can observe that the multi-fidelity optimization resulted in slightly less fuel weight reduction (about 0.7 % less than mono-fidelity optimization). The drag of the wing optimized using the multi-fidelity approach is slightly higher than the drag of the wing optimized using only the low fidelity model. It can be explained by looking at Fig. 4, showing that the low fidelity model slightly underestimates the drag comparing to the high fidelity model. The wing optimized using the multi-fidelity approach has slightly higher sweep than the one optimized using the mono-fidelity method, therefore it has slightly higher wing structural weight. In general the results of the multi-fidelity optimization are more realistic,

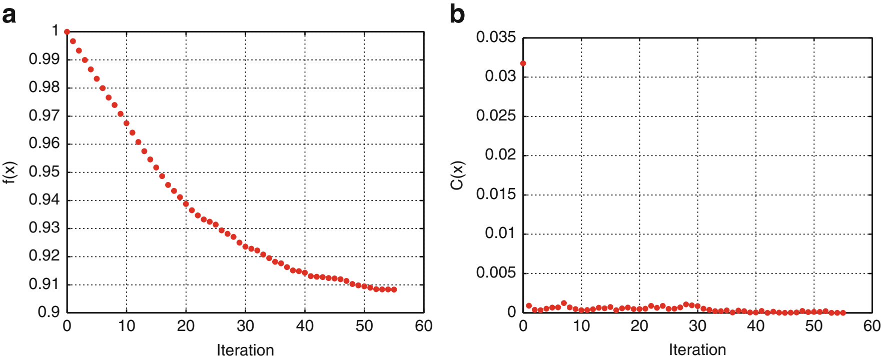


Fig. 9 Optimization history of the mono-fidelity wing aerostructural optimization. (**a**) Objective function. (**b**) Maximum constraint violation

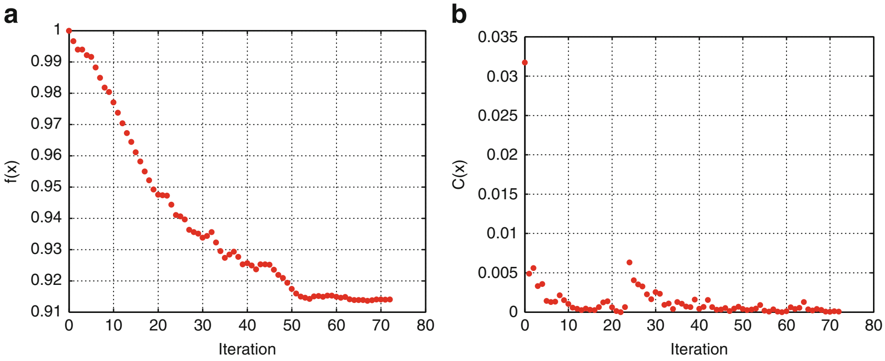


Fig. 10 Optimization history of the multi-fidelity wing aerostructural optimization. (**a**) Objective function. (**b**) Maximum constraint violation

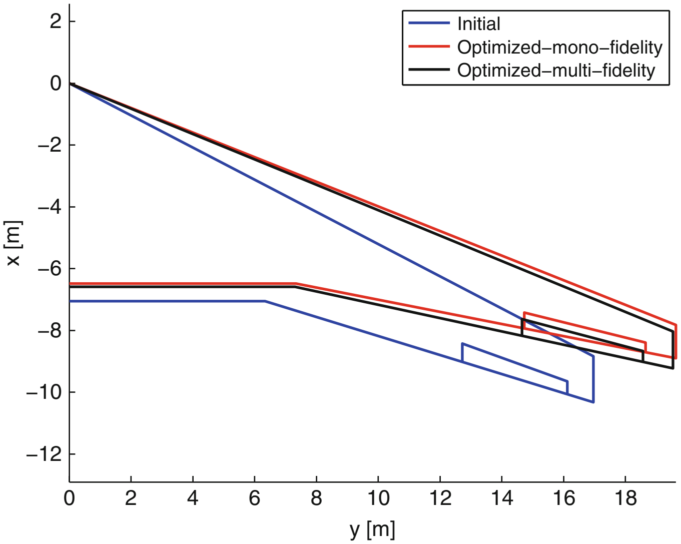


Fig. 11 Planform shapes of the initial and the optimized wings

since a more accurate drag analysis is used. However the results of the mono-fidelity optimization are quite similar to the results of the multi-fidelity optimization, that indicates a good level of accuracy of the low fidelity model. It should also be noted that the mono-fidelity optimization was about four times faster than the multi-fidelity optimization.

In both cases the optimizer moved toward a larger wing span and a lower leading edge sweep. The larger span results in a lower induced drag, but a higher wing structural weight. Lower sweep, on the other hand, results in a lower structural

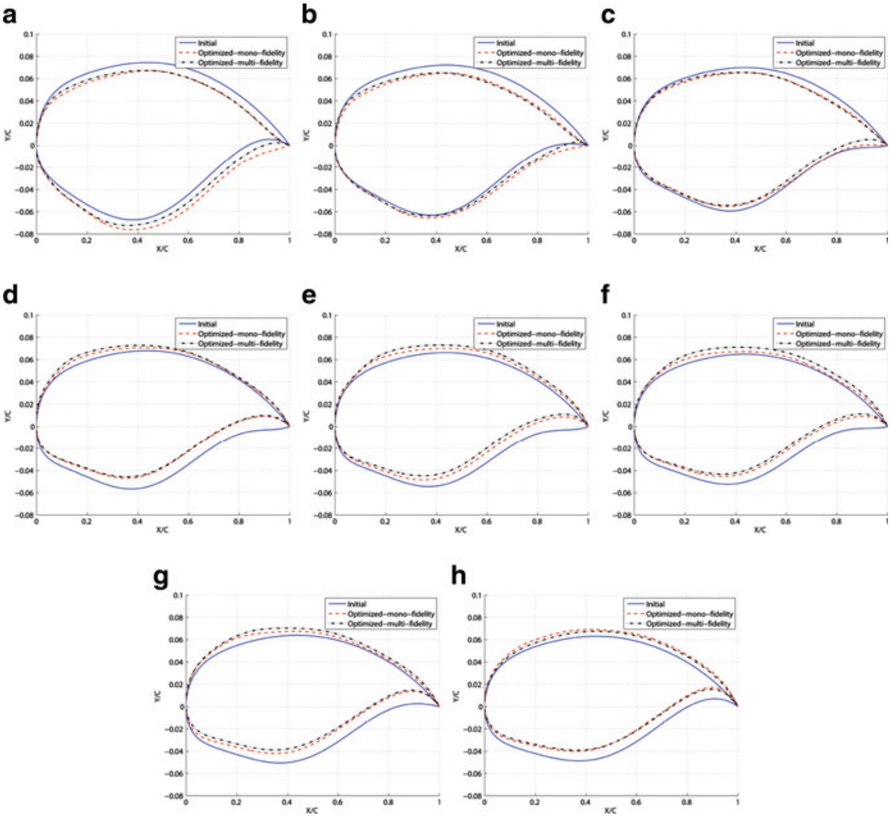


Fig. 12 The shape of the airfoils is several spanwise positions. (**a**) $y/b = 0$. (**b**) $y/b = 0.14$. (**c**) $y/b = 0.29$. (**d**) $y/b = 0.43$. (**e**) $y/b = 0.57$. (**f**) $y/b = 0.71$. (**g**) $y/b = 0.86$. (**h**) $y/b = 1$

weight, but may increase the wave drag. However the optimizer managed to modify the airfoil shapes to eliminate the shock wave from the surface of the optimized wings, see Fig. 13.

The constraint on the aileron effectiveness is usually an active constraint in wing aeroelastic optimization. In fact to achieve higher aileron effectiveness, and consequently a higher value of L_δ, a higher wing stiffness (mainly torsional stiffness) is required for a given wing and aileron geometry. Increasing the torsional stiffness results in a higher structural weight. The study of Elham and van Tooren [6] showed that the wing structural weight increases quadratically by increasing the required value for aileron effectiveness. The optimizer in this research, both in the mono-fidelity and multi-fidelity cases, moved toward a more flexible wing to reduce the wing structural weight. The initial wing has a tip vertical deflection of 1.48 m and tip twist of $-3.8°$ under 2.5 g pull up load, while the optimized wing (using multi-fidelity method) has a tip vertical deflection of 1.77 m and tip twist of $-3.9°$. The aircraft roll requirement was satisfied by increasing the aileron arm and the aileron

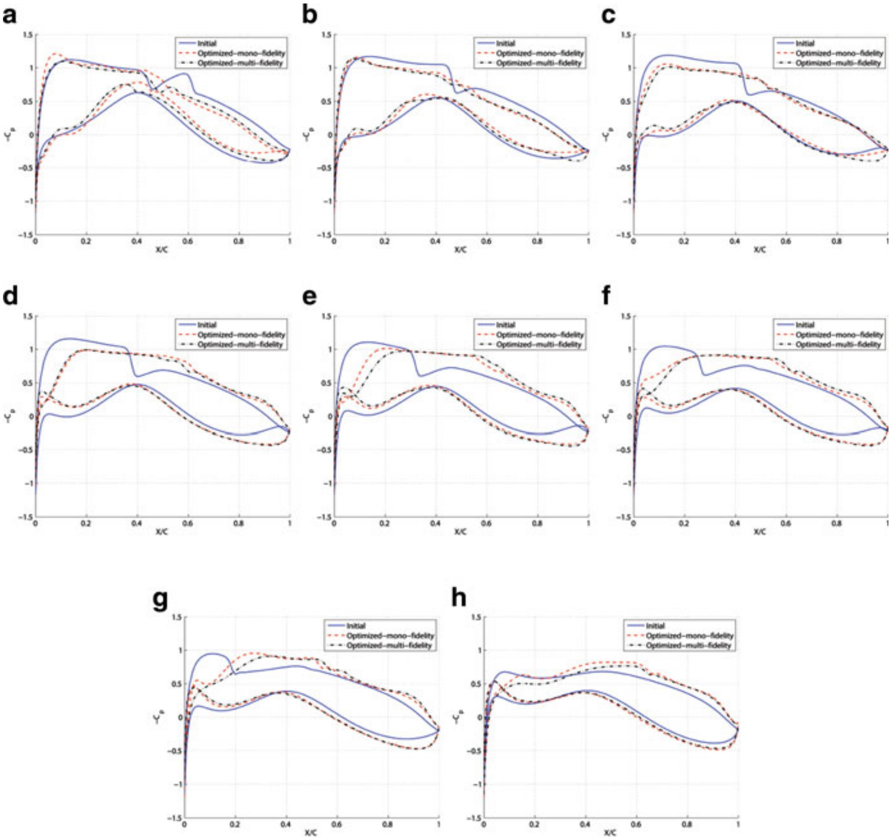


Fig. 13 Pressure distribution on airfoils in several spanwise positions. (**a**) $y/b = 0$. (**b**) $y/b = 0.14$. (**c**) $y/b = 0.29$. (**d**) $y/b = 0.43$. (**e**) $y/b = 0.57$. (**f**) $y/b = 0.71$. (**g**) $y/b = 0.86$. (**h**) $y/b = 1$

Table 2 Characteristics of the initial and the optimized aircraft

	MTOW	W_{fuel}	W_{wing}	C_D
Initial	1	1	1	1
Optimized—mono-fidelity	0.9813	0.9053	1.0370	0.7598
Optimized—multi-fidelity	0.9847	0.9120	1.0520	0.7830

surface (both were resulted from a larger span). The larger aileron area and arm allowed to keep the value of L_δ higher than the required (4.04×10^4 for the optimized wing vs 3.80×10^4 for the initial wing) with a lower value of the aileron effectiveness (0.43 for the optimized wing vs 0.53 for the initial wing) that resulted from a lower wing stiffness.

5 Conclusions

A trust region filter-SQP method is used for wing multi-fidelity aerostructural optimization. The algorithm allows to combine a lower fidelity model, that predicts the sensitivity of the objective function and the constraints, with a higher fidelity model, that is more accurate but more expensive to be executed. The low fidelity model is used to generate the TRQP subproblem. The high fidelity model is used to generate the filter and also to calibrate the results of the low fidelity model. Using that approach a high fidelity CFD tool that does not provide the sensitivities can be used for a gradient based optimization. In addition to that, the aerodynamic solver is coupled with a structural solver for a fully coupled aerostructural optimization.

A mono-fidelity as well as a multi-fidelity wing aerostructural optimization has been performed using the proposed algorithm. The results showed about 9 % reduction in the aircraft fuel weight. The optimizer found the optimum planform shape, airfoil shape as well as the wing box structure to achieve that amount of reduction in the fuel weight.

In this study only a high fidelity aerodynamic solver was combined with the low fidelity aerostructural analysis tool. As the next step a high fidelity FEM can be combined with the low fidelity tool as well. In that approach both the aerodynamic and structural analysis can be performed using two different levels of fidelity. There will be no need for modifying the proposed algorithm. The same algorithm can be used, but this time the constraints related to the wing structure will be computed using two different tools.

References

1. Alexandrov, N.M., Lewis, R.M., Gumbert, C.R., Green, L.L, Newman, P.A.: Approximation and model management in aerodynamic optimization with variable-fidelity models. J. Aircr. **38**(6), 1093–1101 (2001)
2. Conn, A.R., Gould, N.I.M, Toint, P.L.: Trust-Region Methods. MPS-SIAM Series on Optimization, 959 p. Society for Industrial and Applied Mathematics, Philadelphia, PA (2000)
3. Elham, A.: Adjoint quasi-three-dimensional aerodynamic solver for multi-fidelity wing aerodynamic shape optimization. Aerosp. Sci. Technol. **41**, 241–249 (2015)
4. Elham, A., van Tooren, M.J.L.: Effect of wing-box structure on the optimum wing outer shape. Aeronaut. J. **118**(1199), 1–30 (2014)
5. Elham, A., van Tooren, M.J.L.: Coupled adjoint aerostructural wing optimization using quasi-three-dimensional aerodynamic analysis. In: 16th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 22–26 June 2015, Dallas, TX, AIAA 2015-2487
6. Elham, A., van Tooren, M.J.L.: Tool for preliminary structural sizing, weight estimation, and aeroelastic optimization of lifting surfaces. Proc. IMechE Part G: J. Aerosp. Eng. **230**(2), 280–295 (2016)
7. Fletcher, R., Leyffer, S.: Nonlinear programming without a penalty function. Math. Program. Ser. A **91**, 239–269 (2002)
8. Fletcher, R., Leyffer, S., Toint, P.L.: On the convergence of a filter-SQP algorithm. SIAM J. Optim. **13**(1), 44–59 (2002)

9. Gill, P., Murray, W., Saunders, M.: SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev. **47**(1), 99–131 (2005)
10. Kenway, G.K.W., Martins, J.R.R.A.: Multipoint high-fidelity aerostructural optimization of a transport aircraft configuration. J. Aircr. **21**(1), 144–160 (2014)
11. Kenway, G.K.W., Kennedy, G.J., Martins, J.R.R.A.: Scalable parallel approach for high-fidelity steady-state aeroelastic analysis and adjoint derivative computations. AIAA J. **52**(5), 935–951 (2014)
12. Kreisselmeier, G., Steinhauser, R.: Systematic control design by optimizing a vector performance indicator. In: Cuenod, M.A. (ed.) IFAC Symposium on Computer Aided Design of Control Systems. Pergamon Press, Oxford (1980)
13. March, A., Willcox, K.: Convergent multifidelity optimization using Bayesian model calibration. In: 13th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 13–15 Sept 2010, Fort Worth, TX, AIAA 2010-9198
14. March, A., Willcox, K.: A robust approach to aerostructural design. In: 3rd Aircraft Structural Design Conference, Oct 2012. Royal Aeronautical Society, Delft (2012)
15. Martins, J.R.R.A., Alonso, J.J., Reuther, J.J.: High-fidelity aerostructural design optimization of a supersonic business jet. J. Aircr. **41**(3), 523–530 (2004)
16. Martins, J.R.R.A., Alonso, J.J., Reuther, J.J.: A coupled-adjoint sensitivity analysis method for high-fidelity aero-structural design. Optim. Eng. **6**, 33–62 (2005)
17. Meredith, P.T.: Viscous phenomena affecting high-lift systems and suggestions for future CFD development. AGARD TR-94-18415- 04-01, Sept 1993
18. Nocedal, J., Wright, S.J.: Numerical Optimization, 664 p. Springer, New York (2000)
19. Obert, E.: Aerodynamic Design of Transport Aircraft, p. 638. IOS Press, Amsterdam (2009)
20. Roskam, J.: Airplane Design, Part I: Preliminary Sizing of Airplanes. DARcorporation, Lawrence, Kan (1986)
21. Rump, S.M., INTLAB - Interval Laboratory. In: Developments in Reliable Computing, pp. 77–104. Kluwer, Dordrecht (1999)
22. Torenbeek, E.: Advanced Aircraft Design, Conceptual Design, Analysis and Optimization of Subsonic Civil Airplanes, p. 410. Wiley, West Sussex (2013)
23. van Dam, C.: Aircraft design and the importance of drag prediction. In: CFD-Based Aircraft Drag Prediction and Reduction, vol. 2, pp. 1–37. von Karman Institute for Fluid Dynamics, Rhode-St-Genese (2003)
24. Vanderplaats, G.N.: Multidisciplinary Design Optimization, 477 p. Vanderplaats Research and Development Inc., Monterey, CA (2007)
25. van der Wees, A., van Muijden, J., van der Vooren, J.: A Fast and Robust Viscous-Inviscid Interaction Solver for Transonic Flow About Wing/Body Configurations on the Basis of Full Potential Theory, AIAA Paper 1993-3026, July 1993

Tensegrity Rings for Deployable Space Antennas: Concept, Design, Analysis, and Prototype Testing

Pier Luigi Ganga, Andrea Micheletti, Paolo Podio-Guidugli, Lucio Scolamiero, Gunnar Tibert, and Valfredo Zolesi

Abstract In this paper we describe a tensegrity ring of innovative conception for deployable space antennas. Large deployable space structures are mission-critical technologies for which deployment failure cannot be an option. The difficulty to fully reproduce and test on ground the deployment of large systems dictates the need for extremely reliable architectural concepts. In 2010, ESA promoted a study focused on the pre-development of breakthrough architectural concepts offering superior reliability. This study, which was performed as an initiative of ESA Small Medium Enterprises Office by Kayser Italia at its premises in Livorno (Italy), with Università di Roma TorVergata (Rome, Italy) as sub-contractor and consultancy from KTH (Stockholm, Sweden), led to the identification of an innovative large deployable structure of tensegrity type, which achieves the required reliability because of a drastic reduction in the number of articulated joints in comparison with non-tensegrity architectures. The identified target application was in the field of large space antenna reflectors. The project focused on the overall architecture of a deployable system and the related design implications. With a view toward

P.L. Ganga (✉) • V. Zolesi
Kayser Italia srl, Livorno, Italy
e-mail: p.ganga@kayser.it; v.zolesi@kayser.it

A. Micheletti
Dipartimento di Ingegneria Civile e Ingegneria Informatica, University of Rome "Tor Vergata", Rome, Italy
e-mail: micheletti@ing.uniroma2.it

P. Podio-Guidugli
Accademia Nazionale dei Lincei, Rome, Italy

Department of Mathematics, University of Rome "Tor Vergata", Rome, Italy
e-mail: ppg@uniroma2.it

L. Scolamiero
European Space Agency, ESA ESTEC, Noordwijk, The Netherlands
e-mail: Lucio.Scolamiero@esa.int

G. Tibert
Department of Aeronautical and Vehicle Engineering, KTH , Stockholm, Sweden
e-mail: tibert@kth.se

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_11

verifying experimentally the performance of the deployable structure, a reduced-scale breadboard model was designed and manufactured. A gravity off-loading system was designed and implemented, so as to check deployment functionality in a 1-g environment. Finally, a test campaign was conducted, to validate the main design assumptions as well as to ensure the concept's suitability for the selected target application. The test activities demonstrated satisfactory stiffness, deployment repeatability, and geometric precision in the fully deployed configuration. The test data were also used to validate a finite element model, which predicts a good static and dynamic behavior of the full-scale deployable structure.

List of Symbols

n	Number of bars in the tensegrity prism (TP)
a	Lower TP radius
b	Upper TP radius
h	TP height
φ	TP twist angle (for short, the twist)
h^*	"Overlap" between two successive stages of a symmetric Snelson tower/Snelson ring
γ	a/b ratio between the TP radii
δ	h^*/h ratio between TP height and Snelson tower/Snelson ring overlap
H_s	Stowed height of the deployable tensegrity ring

1 Introduction

Large deployable space antenna reflectors (LDRs), with diameters between 4 and 25 m, are required in several mission types, particularly in the telecommunication domain, but also for Earth observation, deep-space missions, and radio-astronomy [9].

To be deployed once in orbit, reflectors with diameter in excess of 4–5 m must have a foldable structure for compatibility with the launcher's available storage. When associated with the need for extreme deployment reliability, the demanding mechanical, thermal, and radio-frequency requirements of the as-deployed reflector result in very challenging, multidisciplinary design issues. As a consequence, very few companies specialize in the production of such large reflectors, most of them being based in the USA (Northrop-Grumman, Harris Corporation) and subject to US exporting regulations.

Due to the emerging market of small and microsatellites and the stringent storage requirements dictated by small launcher fairings, the foldability requirement may be imposed also to much smaller reflectors (down to 2 m in diameter); therefore, LDR technology is a potential candidate for much larger a class of antenna reflectors.

With a view toward reducing dependence on non-EU suppliers, ESA is pursuing developments in this domain. In particular, within the frame of an initiative of the ESA Small Medium Enterprises Office (http://www.esa.int/SME/), the study of a potentially breakthrough technology has been undertaken, whose goal was to conceive a deployable large antenna reflector of intrinsically high reliability. A concept validation by testing a reduced-scale breadboard model has been performed.

This paper reports the outcome of the above mentioned activities, namely the conception of an innovative large deployable structure based on “tensegrity” principles, currently being protected by an international patent filing [17].

2 Large Space Deployable Reflectors

The need for LDRs of 4–25 m in diameter is well established [9]; in fact, the market goes beyond pure telecommunication missions (still the major users of such technology), and spans from Earth observation, navigation, and deep-space missions, to radio-astronomy.

Operative radio-frequency bands go from the lowest P-band frequencies up to L-S, Ku and higher, and finish with Ka, the band reserved for small-diameter reflectors, typically about 5 m in diameter. Several 12-m reflectors have already successfully flown, and recent missions have embarked and successfully deployed reflectors up to 18 and 22 m in diameter.

To comply with the demanding radio-frequency needs, as-deployed shape accuracy and high stability in operational conditions (for the entire operational life) are required. To limit the overall reflector mass, high-stability/low-density materials and technologies are utilized, with large use of carbon fiber reinforced plastics (CFRP) for rigid structural members. Subtler radio-frequency phenomena (known as passive inter modulation products—PIMP) pose even stricter requirements on candidate materials, processes, and thermo-mechanical design solutions.

2.1 LDR Classification and State of the Art

Here follows a brief overview and classification of the state-of-the-art LDR architectures and the relative mission applications [9].

2.1.1 Mesh Reflectors

This is by far the most successful technology, based on a tension-truss concept and metal tricot mesh as reflective layer. The mesh is knitted with gold plated tungsten or molybdenum wires (15–25 μm diameter), and it needs to be subjected to a tension between 5 and 10 g/cm both to achieve adequate electrical contact between wires and to prevent PIMP problems. To provide the required mesh tension and shape accuracy simultaneously is a challenging task, due to the detrimental phenomenon known as “pillow effect” (or “facet effects”) which occurs when controlling the mesh surface in a limited number of points. Among mesh reflectors, two different mechanical and deployment architectures dominate the market: the *peripheral expandable ring* and the *hinged radial ribs* architectures.

In the former architecture, the peripheral expandable ring applies tension to two paraboloidal triangular networks. The two tensile networks provide shape and pretension to the underlying radio-frequency reflective mesh layer (this is the case, e.g., for the AstroMesh reflector by AstroAerospace, now Northrop Grumman, Fig. 1, [20]). LDRs belonging to this category have been supplied by Northrop Grumman and flown on the following telecommunication missions: Inmarsat-4, Alphasat (Inmarsat-XL), Thuraya-1,-2,-3, MBSAT, and SMAP for an Earth Observation mission.

Hinged radial ribs reflectors are based on multiple, rigid, radial compressive elements to apply tension to the radio-frequency reflective mesh. Cable nets and cable ties provide the mesh with the required shape (Fig. 2). Harris has supplied reflectors based on this architectures for the following telecommunications missions: SkyTerra-1, TerreStar-1 (now EchoStar-1) ICO-G1, MexSat-1, and MexSat-2.

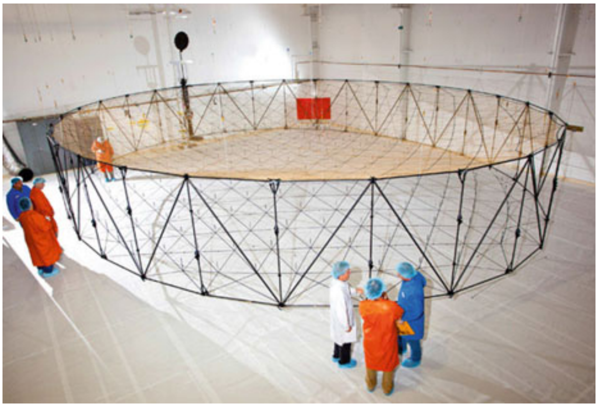


Fig. 1 Astro-Mesh 14 by 11 m peripheral expandable ring concept (image by Northrop Grumman via spacenews.com/northrop-unit-delivers-alphasat-1-xl-reflector/)

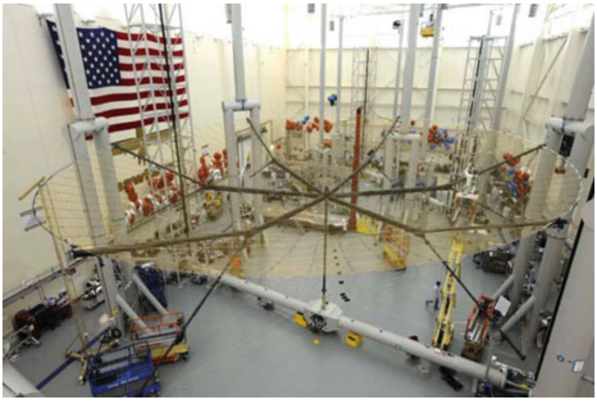


Fig. 2 TerreStar 18 m diameter reflector (image by Harris Corporation, harris.com/harris/whats_new/TerreStar-reflector.jpg)

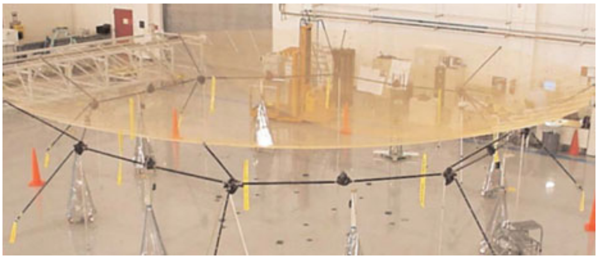


Fig. 3 Hoop-truss reflector (image by Harris Corporation via www.propagation.gatech.edu/ECE6390/project/Fall2010/Projects/group4/comm.html)

Variations to these two basic architectures do exists, e.g., the "Hoop-Truss reflector," also called "Double pantograph ring" or "Conical pantograph ring," in Fig. 3 (ESA patents: 568–596, US patent US 9153860 B2) but there are no known flight applications to date.

Europe is very active in the domain of metallic mesh based LRDs, and aims to achieve technological independence from non-European suppliers in this strategic domain. Among other activities, ESA is performing technology research and development activities in the domain of basic metallic-mesh materials [16] and in the domain of alternative deployment architectures, focusing on dimensional scalability and modularity of the concepts, so as to cover diameters from 5 to 18 m and more while maintaining cost effectiveness of the final product [16]. Many reduced-scale (typically, 4 m in diameter) ground demonstrators have been produced.

It is worth noticing that very recently ESA has selected the 12 m LDR for the 7th Earth Explorer mission: BIOMASS, set for launch in 2020 [3]. The selected deployable reflector technology falls in the domain of hinged radial rib reflectors, supplied by Harris Corporation.

2.1.2 Membrane and Inflatable Reflectors

These reflectors consist of a thin membrane of metallized polyimide films. They need first to be inflated, in order to achieve the required shape and surface accuracy, then to be made rigid (via thermal or UV curing of associated resin systems) to maintain shape and stiffness during their operational life. Notably, the ESA 12 m diameter inflatable rigidizable reflector (Fig. 4), and the JPL L.Garde 14 m inflatable antenna demonstrator, were flown in 1996 on board the STS-77 mission (Fig. 5). The major limit of this technology is the modest achievable surface accuracy, the

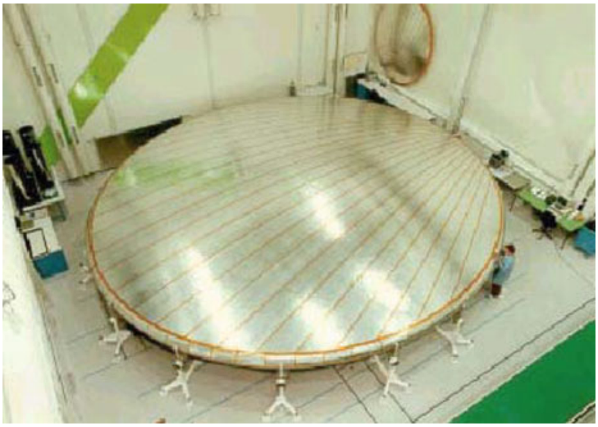


Fig. 4 ESA—Contraves—RUAG (CH) inflatable space rigidized structure (image by Contraves via www.thespaceoption.com/cres_mcbc.php)

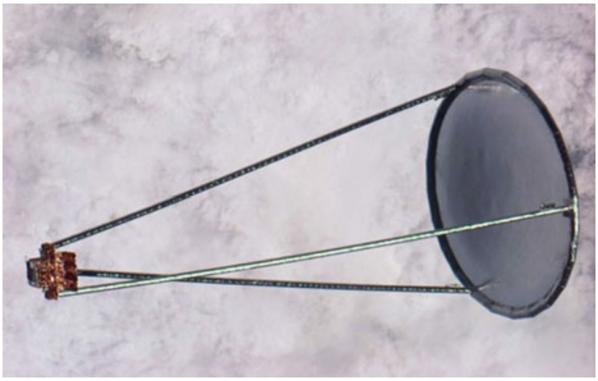


Fig. 5 Inflatable antenna experiment—NASA JPL—L. Garde (image by L.Garde, www.lgarde.com/deployable-antennas.php)

reason why reflectors of this kind are still at the level of prototypes and technology demonstrators, and can only be utilized at lower range of radio-frequency bands.

2.1.3 Shell-Membrane Reflectors

These reflectors are based on a triaxially woven carbon-fiber fabric, reinforcing a space-qualified silicon matrix (CFRS). Developed by the Technical University of Munich—LLB, this material allows for full foldability of the reflecting surface, preserving the high dimensional stability and radio-frequency properties of the carbon-fiber layers. The main advantage with respect to classical metallic-mesh reflectors is that there is no need of tensioning a mesh, and hence no "pillow effect." Developments are on-going in Europe [2], although there has been no flight mission application to date.

2.1.4 Largely Deformable Shell Reflectors

These reflectors provide a very elegant and mass efficient solution. They are very popular although currently limited in diameter size (no more than 6–8 m). This class of reflectors is well represented by the "Spring Back Antenna" from Hughes Space and Communication Group, now Boeing Satellite Systems Inc., for data relay satellite missions (Fig. 6).

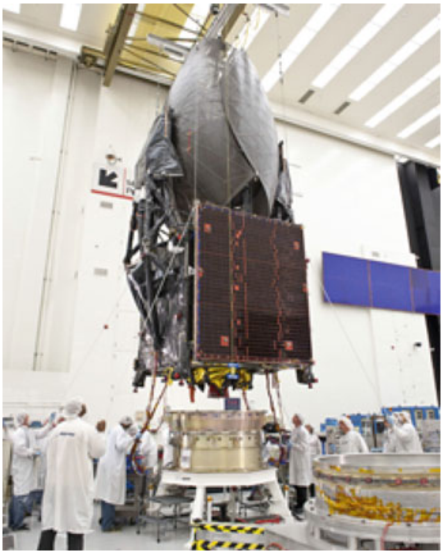


Fig. 6 Boeing Satellite System (former Hughes Space and Communication) Spring Back Antenna reflector (image by Boeing Satellite Systems via www.nasa.gov/topics/technology/features/tdrs-upgrade.html)

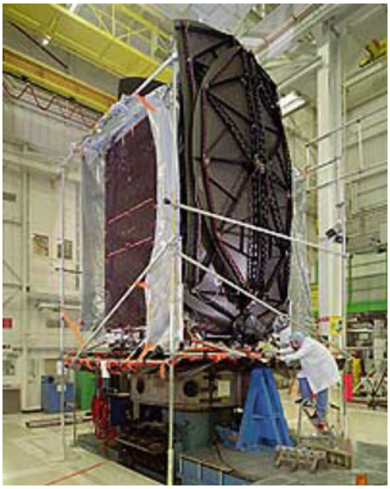


Fig. 7 Boeing Satellite System (former Hughes Space and Communication) foldable rigid panels reflector (image by Boeing Satellite System via www.spacedaily.com/news/xm-radio-01c.html)

2.1.5 Solid Surface Reflectors

The best example of this architecture is the XM Satellite Radio antenna reflectors from Hughes Space and Communication International (Fig. 7), now Boeing Satellite System Inc. Also in this case, diameters do not exceed 6–8 m; their surface accuracy is better than that of deformable shell reflectors, and they also allow for "surface shaping" features (*ad hoc* deviations from a nominal paraboloid) improving radio-frequency performances.

2.2 LDR State-of-the-Art Assessment

What makes antenna reflectors unique in terms of design challenges is the need for extreme deployment reliability: a deployment failure would most of the times result in mission loss, an unacceptable option.

Several concepts have been studied worldwide to combine the reflector-specific set of multidisciplinary requirements and the fundamental need of an absolutely reliable deployment. However, the very specialized competencies required, and the amount of investment necessary to develop/qualify reliable products, have resulted in very few companies offering commercially qualified units, the most prominent being Harris Corporation [5] and Northrop-Grumman [13].

The experience gained by the major large reflector suppliers notwithstanding, the deployment of such items is always a critical step in a mission scenario. Indeed, the typical structure to be deployed consists of a large number of interconnected rigid elements. As a consequence, a large number of mechanical joints (either simply revolute or telescopic, or motorized, joints) are necessary to fold the structure when in launch configuration and to deploy it in orbit.

Mechanical joints/hinges, and “mechanisms” in general, are typically sources of reliability concern, in that they may induce localized failures. The starting point of the development presented in this paper is that a system with a minimal number of joints would have optimal reliability, because of the low number of single-point failure sources.

The possibility of using a structural architecture of “tensegrity” type, where mechanical joints are in principle totally absent from the design, was then considered. The idea of using “tensegrity”-type structures for large antenna applications is not new, and in fact it has been the subject of a related patent [19]. However, it is our opinion that the new ideas we conceived in the course of our study, and the new design features we introduced, make the final design original and unique, so much so as to deserve an international patent filing [17].

In the following sections, we shall describe the technical features of the structural architecture we propose, as well as its validation by means of the realization of a scaled model breadboard and a test campaign.

3 Tensegrity Structure Description

3.1 Definition

Tensegrity structures (TS) were first conceived by the artist Kenneth Snelson [4, 18] in 1948. In the 1960s, Snelson began to build a number of outdoor sculptures, which made tensegrities worldwide popular among architects and engineers because of their innovative structural concept. Indeed, when an architect or a structural engineer looks at a realization of Snelson’s, he observes that:

- TSs are pre-stressed spatial frameworks whose elements are bars and cables;
- cables form a connected set, i.e., tens(ile int)egrity;
- bar ends never touch (floating compression).

In addition, TSs possess the important *form-finding property*, to be described in Sect. 3.3.

3.2 The Tensegrity Prism

A regular n-bar tensegrity prism (TP) is a cyclic-symmetric structure with an n-fold axis of cyclic-symmetry, always taken vertical. As shown in Fig. 8, a TP can have

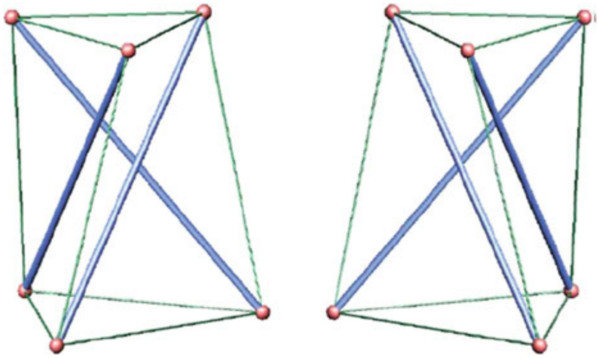
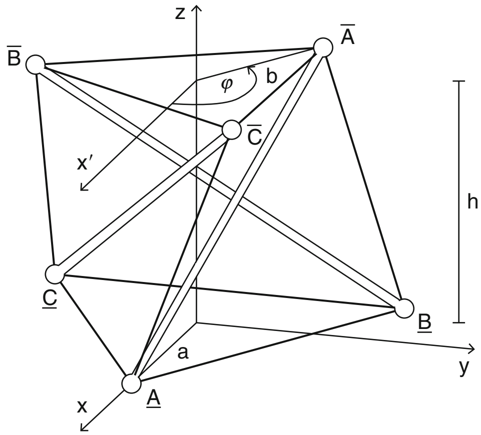


Fig. 8 The simplest three-dimensional TS: two tensegrity prisms with opposite chirality

Fig. 9 The TP parameters



two different orientations. The geometry of a TP can be identified by means of five parameters (Fig. 9):

- the number of bars n,
- the lower "radius" a,
- the upper "radius" b,
- the height h,
- the twist angle φ (for short, the *twist*).

3.3 Form-Finding Property

As observed by Oppenheim and Williams [14], form-finding (FF) is a property that becomes evident when we try to build a TS by hand. Let us suppose that we have what is necessary to assemble one of the systems in Fig. 8, all elements having a fixed length. Once all but one connections between elements are realized, we notice that the so obtained partial assembly has no stiffness, and that there are many possible configurations with slack cables. If the last element is a cable (a bar), its length is determined when we try to decrease (increase) the distance between the two nodes to be connected. As soon as that distance reaches a minimum (maximum) value, the system takes its shape. If we force the two nodes to get closer (farther), then the system acquires a self-stress state with the last element in tension (compression). Figure 10 illustrates the FF property in the simplest case. With this example in mind, we can state the FF property as follows: "*Given an N-elements tensegrity system, if the lengths of* $(N - 1)$ *elements are fixed, then a stable equilibrium configuration is obtained when the last cable (bar) has minimal (maximal) length.*"

For a fixed topology, i.e., once a collection of nodes connected by bars and cables is chosen, it is possible to pass from one stable configuration to another simply by changing the lengths of two or more elements.

Due to the FF property, a tensegrity system is stable only for a restricted set of configurations. For example, in the system in Fig. 10, stable configurations are those with the three nodes collinear. The problem of finding the set of stable configurations for a given tensegrity system, referred to as the *form-finding problem*, has been extensively studied in the literature [6, 21].

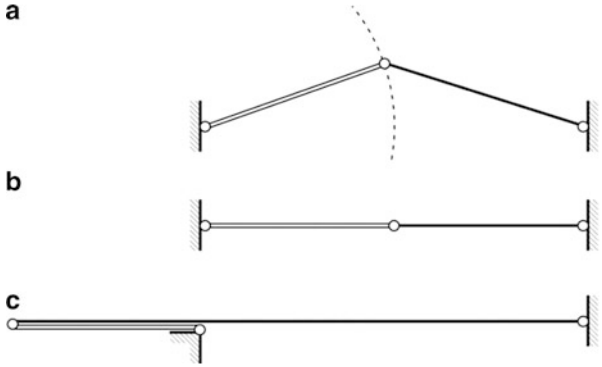


Fig. 10 The form-finding property for a system composed of two elements. The double-line element has fixed length; the single-line element has variable length. The central node can only be on the dashed circumference shown in (**a**). On progressively shortening (lengthening) the single-line element, configuration (**b**), (**c**) is reached; on further shortening (lengthening) the same element, the system is found in a self-stress state with that element in tension (compression)

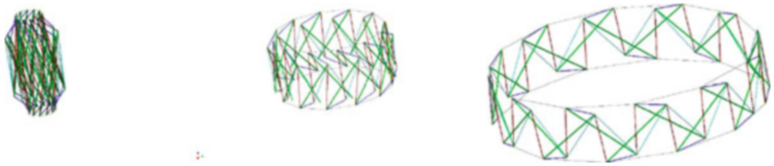


Fig. 11 TS theoretical ring in different equilibrium configurations

3.4 Tensegrity Deployable Structures

The FF property of tensegrity systems and their related capability to change shape suggest to have recourse to such systems when it is desirable to have deployable or variable-geometry structures, or "smart" structures, some elements of which serve as sensors and actuators. By actuating cables and/or bars, a TS can pass from one equilibrium configuration to another through a continuous path of equilibrium configurations (Fig. 11 shows a TS ring in different equilibrium configurations). Due to the absence of hinges between bars, the mechanical behavior of a floating-compression system can be predicted with better accuracy than for conventional hinged systems.

3.5 Tensegrity Rings for Space Structures

The first studies of ring-shaped TS's appear to be performed by Burkhardt [1] in 2003; a tensegrity torus was analyzed by Peng et al. in 2006 [15] and by Yuan et al. in 2008 [22]. The deployable tensegrity ring that we here consider has been presented for the first time in [23]. The same kind of tensegrity ring has been studied in [7].

The tensegrity ring (TR) concept is suitable for disc- or ring-shaped space structures. Since bars are not connected to each other, none of the usual hinge mechanisms are present in TS's: freedom in spatial orientation and relative motion of bars during deployment is granted by the flexibility of the interconnecting cables. The absence of mechanical joints reduces the possible failure modes of the deployable system, thus increasing its overall reliability, a fundamental requirement for this type of space technology; in addition, this feature permits an especially tight and compact stowage of the structure. Finally, as is the case with conventional pin-jointed trusses, none of the individual members is bent, sheared, or twisted.

We named the tensegrity ring we developed for the present application "Snelson ring" (SR). SR is a TR with the same graph as a two-level Snelson tower. To obtain a Snelson tower, we "superimpose" a number of tensegrity prisms (TP) (as shown in Fig. 12) by repeating the following sequence of steps:

1. we take two prisms with opposite orientations;
2. we remove the lower cables of the upper prism;

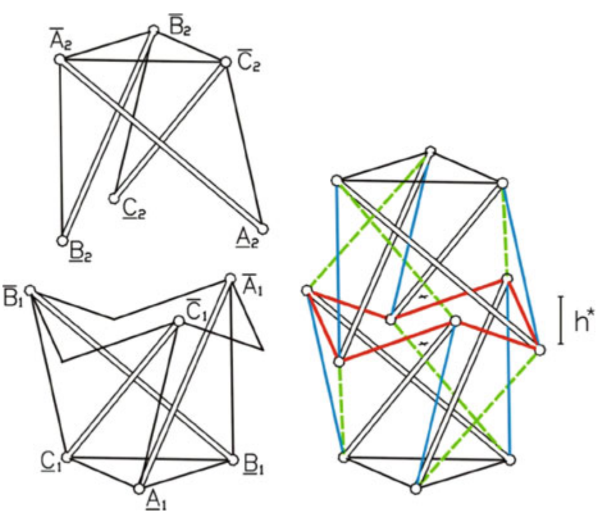


Fig. 12 Superposition of two TPs to obtain a two-level Snelson tower

3. we connect the lower nodes of the upper prism with the middle points of the upper base cables of the lower prism;
4. we add $2n$ additional cables (in green in Fig. 12).

In an SR, we distinguish four groups of cables according to position, in such a way that symmetrically placed cables belong to the same group. The cables in these groups are named as follows:

- “verticals,” when they connect bars of the same TP;
- “diagonals,” when they connect bars of different TPs;
- “saddles,” when they belong to both TPs;
- “polygonals,” when they form base polygons.

In Fig. 12, verticals, diagonals, and saddles are depicted, respectively, in blue, green, and red.

The geometry of symmetric Snelson towers can be identified by six parameters, namely the above-defined five parameters (n, a, b, h, φ) of a typical TP plus a new parameter:

h^* is the “overlap“ between stages (see Fig. 12).

Note h^* is null when saddles lay on the same horizontal plane. Three additional geometric properties are used to characterize a deployable SR:

δ is the overlap ratio (h^*/h) between the Snelson tower/Snelson ring overlap

and the TP height;

γ is the ratio (a/b) between the two radii of a typical TP;
H_s is the stowed height of the deployable tensegrity ring.

The form-finding condition for TP and SR has been obtained in the literature by different authors; here we make use of the conditions given in [10]. For TPs, the form-finding condition:

$$\varphi = \frac{\pi}{n} + \frac{\pi}{2} = \varphi_0$$

involves only n and φ. For $\theta := \varphi - \varphi_0$, the form-finding condition for SRs:

$$\delta^2(\gamma \sin\varphi_0 + \sin\theta) + \delta(\gamma - \gamma \sin\varphi_0 + 2\sin\varphi_0 \sin\theta - 2\sin\theta) + \\ -2\sin\varphi_0 \sin\theta + 2\sin\theta \;=\; 0 \tag{1}$$

involves the full set of parameters, namely n, φ, γ, and δ. In Fig. 13, δ is plotted versus φ for various values of γ and for $n = 6$ and $n = 12$. We see that ring-shaped Snelson towers obtain for small twists and large overlaps.

3.6 Deployment Strategy

A TR can be deployed by changing the length of some of its elements so as to obtain the desired change in shape from stowed to deployed configurations. For the SR considered for the present application, it was chosen to change the length of a subset of cables, while keeping constant the lengths of the remaining cables and of all the bars. In order to have a slow, smooth, and controllable deployment process, all the cables in the TR have to be kept in tension.

SRs have good properties with regard to their use as deployable structures. We found that SR can easily be deployed by lengthening the polygonal cables while shortening the vertical cables, as shown in Fig. 14. SRs have the important property of being *super-stable* [11], a feature that other types of TR lack. Super-stability implies that a structure is stable independently of the self-stress level and of its elastic properties, so that finding admissible deployment paths is simpler. It is worth noticing that super-stability of SRs does not depend on n.

The adopted deployment strategy consists of two phases:

- change in configuration, from folded to deployed (Deployment Phase 1);
- final pre-stressing, to reach a prescribed stress level in the system (Deployment Phase 2).

In the present study, Deployment Phase 1 has been simulated numerically by means of the procedure detailed in [12]; a custom-made finite element code has been used for the simulation of Deployment Phase 2. The feasibility of both phases has been verified in advance with the aid of small-scale models.

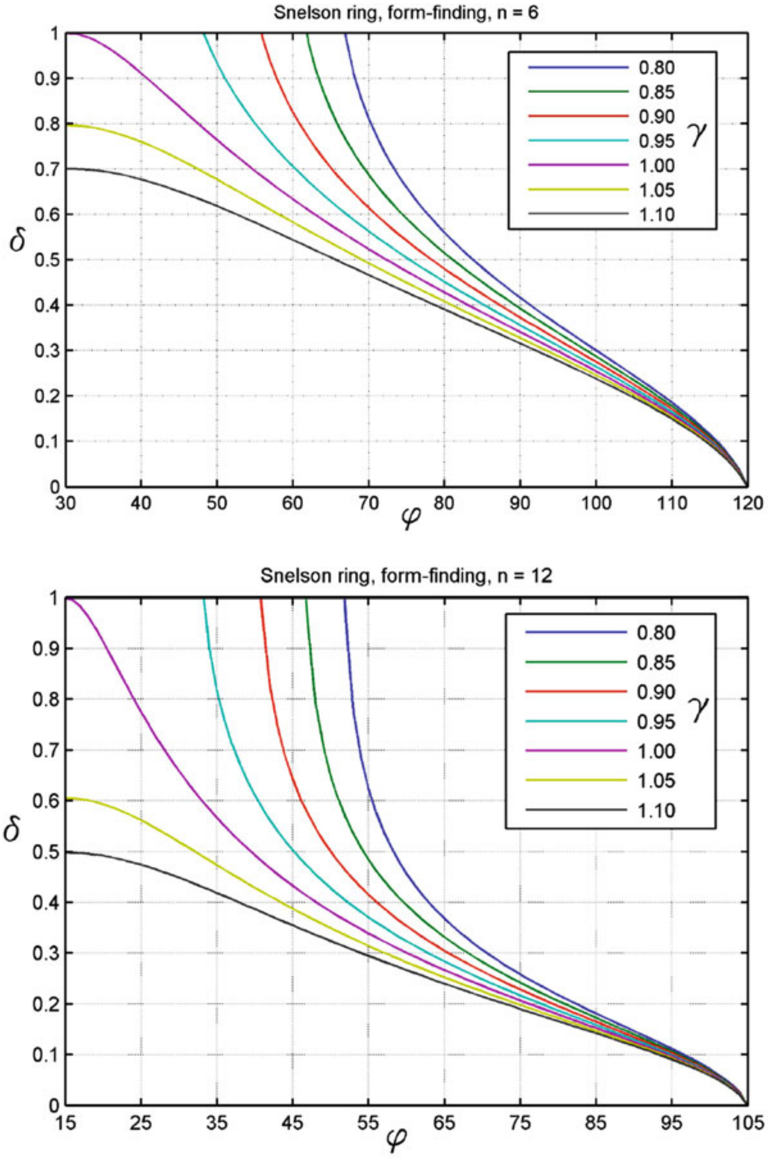


Fig. 13 Relation between δ and φ (°) for various values of γ, $n = 6, 12$

3.6.1 Deployment Phase 1

During Deployment Phase 1, the change of configuration is obtained by changing only the lengths of two groups of cables: the polygonal cables lengthen, the vertical cables shorten. Figure 14 shows the stowed and the deployed configuration of a

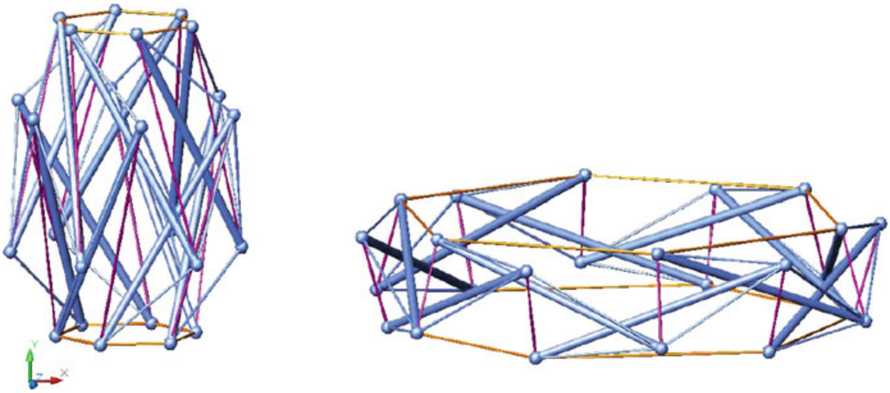


Fig. 14 A hexagonal TR, folded and deployed

hexagonal TR, one obtained from the other in this way (purple cables are shortened during deployment; orange cables are lengthened).

3.6.2 Deployment Phase 2

Due to the FF property, the pre-stress can be induced in the structure by acting on few cables only. These cables can be conveniently chosen among those not involved in Deployment Phase 1, on keeping in mind that the corresponding actuators are due to apply a large force to obtain a small change in length.

4 Tensegrity Space Structure Design

A deployable tensegrity ring of Snelson type (SR) was identified as the main structure in a tensegrity space structure (TSS) to be designed consistent with the following specifications, among others:

- Function: Deployable Antenna Reflector
- Operating frequency: from 6 to 14 GHz
- Reflective Mesh tension: 5 N/m
- Reflector diameter: 12 m
- Stowed height: about 4.4 m
- Stowed diameter: about 2.4 m (excluding the reflector-to-boom interface)
- Mass budget: 57 kg or less (excluding the spacecraft boom)
- Eigenfrequency (deployed, not including boom): 1.2 Hz (min), 1.5 Hz (target).

These specifications are compliant with the typical launcher's mechanical interface (i.e., stowed dimensions) and the typical deployed-to-folded diameter ratio.

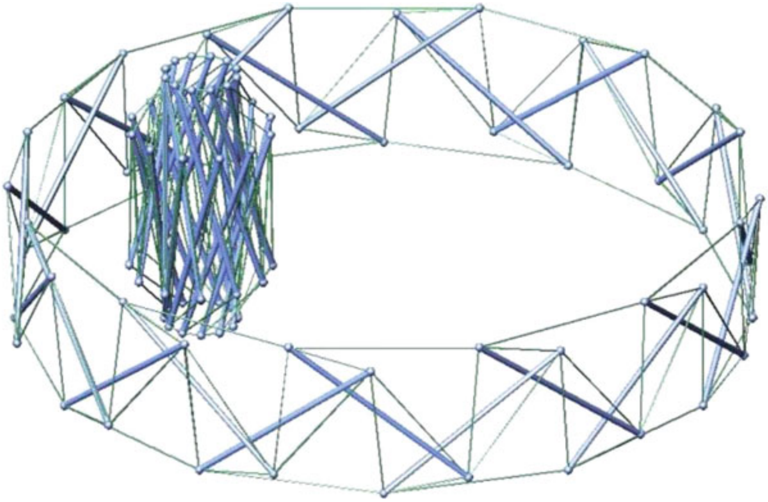


Fig. 15 TSS deployable tensegrity ring model

4.1 Tensegrity Ring Analysis

We considered a reflector of 12 m in diameter and 2.6 m in height, so as to accommodate the two paraboloids with a gap for the central tension-tie ($2.6 = 2 \times 1.25 + 0.1$). A parametric analysis of the SR was performed in the absence of the inner tension truss (also called web in the present document) (Fig. 15).

The FF analysis presented in Sect. 3.5 showed that suitable configurations have a small twist φ and a large overlap ratio δ. Note that it is not possible to have $\delta \geq 1$, since this would require that some cables take a compressive or null stress; moreover, having $\gamma > 1$ causes problems with regard to the clearance between bars. Given these constraints, we focused on those configurations having δ near but not greater than 1. Figure 16 shows a closer view of the form-finding solutions for $\gamma = 0.96, 0.98, 1.00$, and $n = 12$.

To pick a convenient set of geometric parameters, we looked at deployability; in particular, we computed an approximate value for the stowed height H_s, as the sum of the lengths of one bar and one diagonal cable. We did this because in the stowed configuration these elements, which are kept almost parallel to the vertical axis, span the height of the SR. The computed values are plotted in Fig. 17 for the same values of γ and n as before. These plots shows that only for $\gamma = 1$ the stowed height requirement can be fulfilled. However, a precise computation with the procedure given in [12] gives smaller values of H_s, and by taking $\gamma = 0.98$ the resulting stowed height is $H_s = 4.53$ m.

Next, we looked at the clearance between bars, D_b, computed as the distance between their axes. Figure 18 shows that the clearance becomes very small as φ

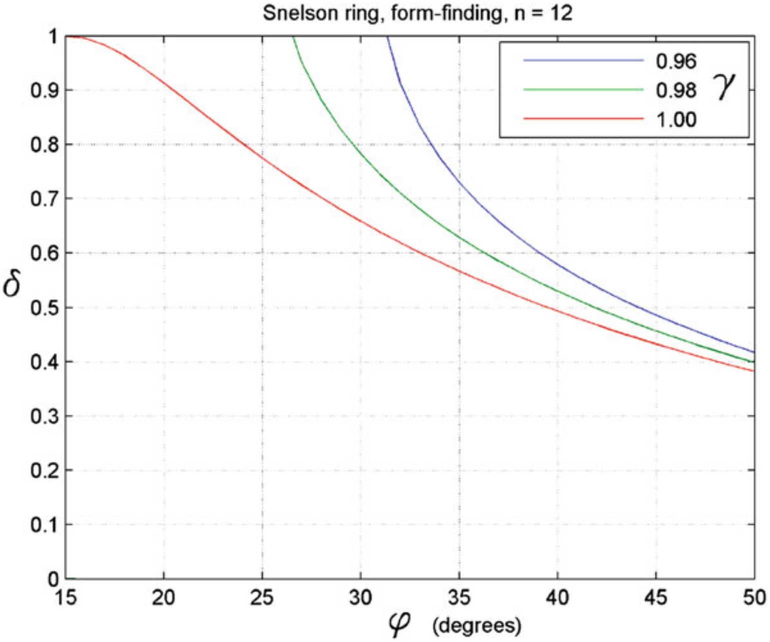


Fig. 16 A closer view of the form-finding solutions for an SR, $n = 12$

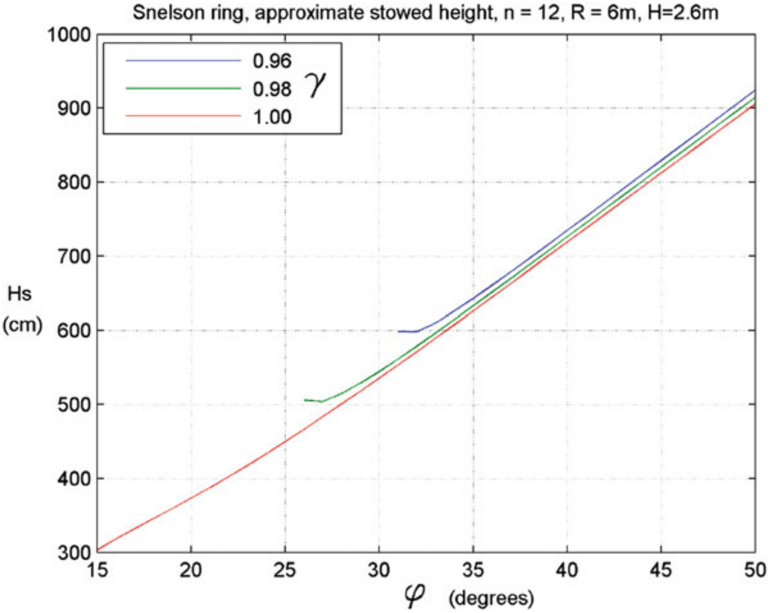


Fig. 17 Approximate stowed height versus φ for various values of γ, $n = 12$

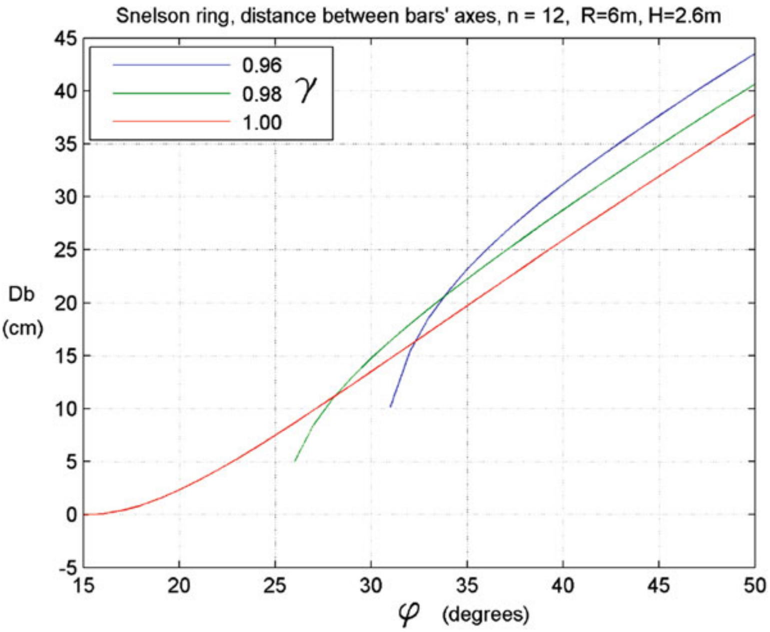


Fig. 18 Clearance between bars, $n = 12$

gets closer to the range of interest. Note that the behavior of both D_b and H_s does not change much when increasing n. Lastly, with a view toward dimensioning bars and cables, we checked bar lengths (Fig. 19) and stresses (Fig. 20).

All in all, the parameters chosen in order to provide a compact stowage of the ring, while maintaining good structural performances, are the following:

$$\varphi = 28^\circ\ , \quad \gamma = 0.98\ ,$$

with, we recall, $n = 12$, a deployed diameter of 12 m, a deployed height of 2.6 m, and a stowed height of 4.53 m. Figure 15 shows such an SR, both folded and deployed. Figure 21 shows the height versus the base radius of the reflector during deployment. The plot of the clearance between bars versus the base radius during deployment in Fig. 22 shows that the clearance decreases monotonically and reaches a minimum value of about 0.11 m in the deployed configuration.

Remark. To match a web with six-fold symmetry, parameter n should be a multiple of 6. We found that, with $n = 6$, a 12 m reflector based on an SR cannot satisfy the stowage-height requirement, because bars would be excessively long; however, reflectors of smaller radius can have $n = 6$ and be conveniently designed in the same way described above. On the other hand, an SR with $n = 18$ would be too complex a structure for a 12 m reflector, requiring a larger number of actuators than an SR with $n = 12$; in addition, such an SR would also be a more, and possibly too much, flexible structure.

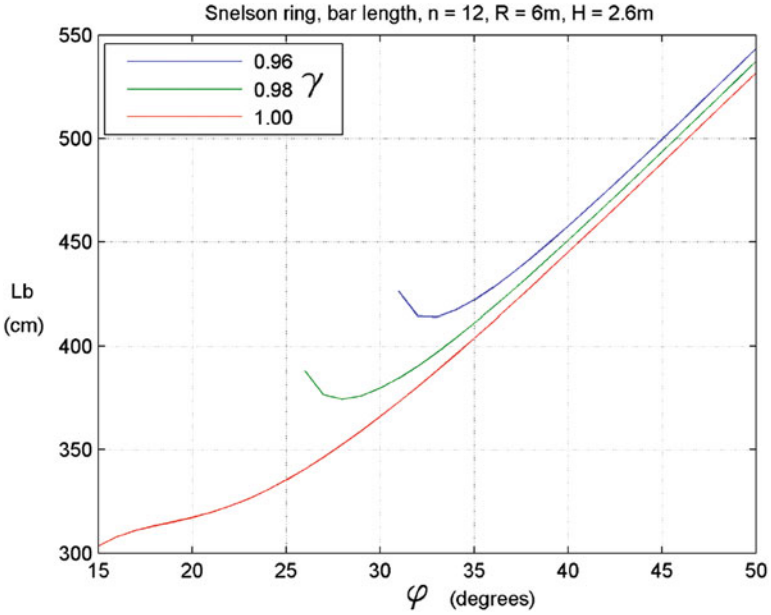


Fig. 19 Bar length, $n = 12$

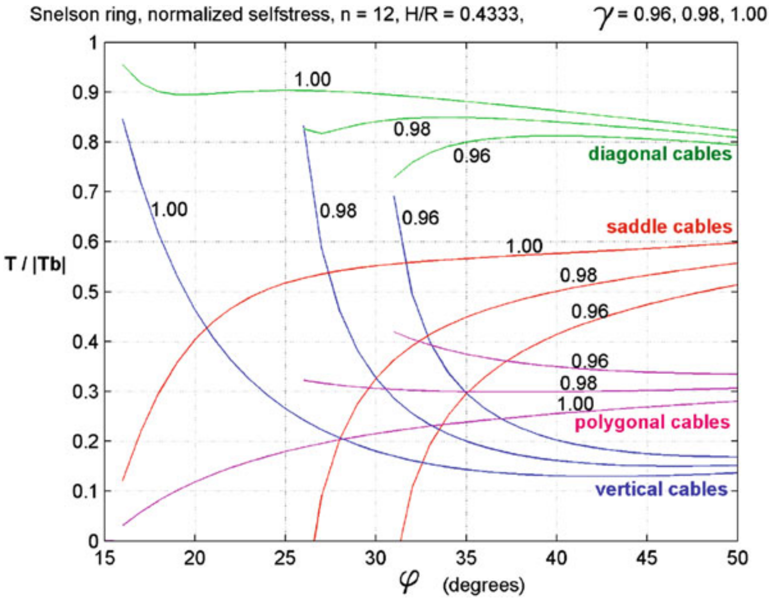


Fig. 20 Axial forces in cables, normalized with respect to the compressive force in bars

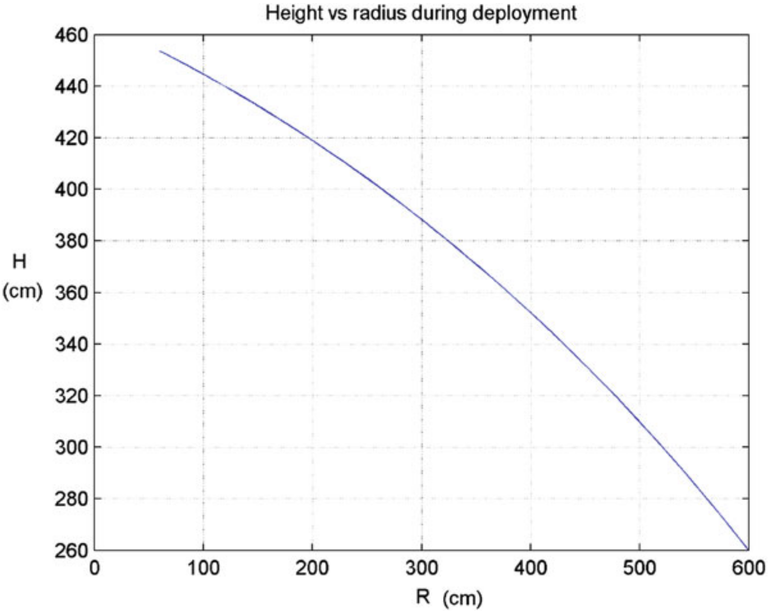


Fig. 21 Design of the reflector: height versus radius during deployment

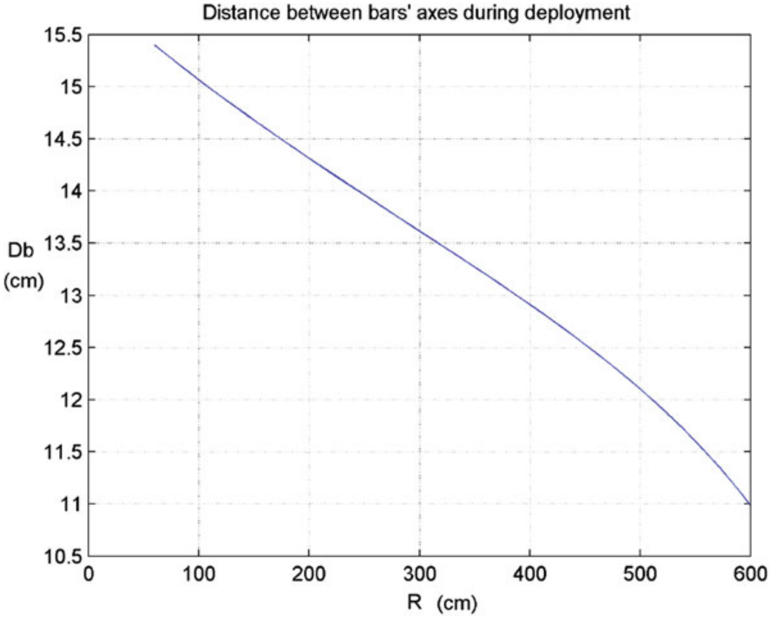


Fig. 22 Design of the reflector: clearance between bars during deployment

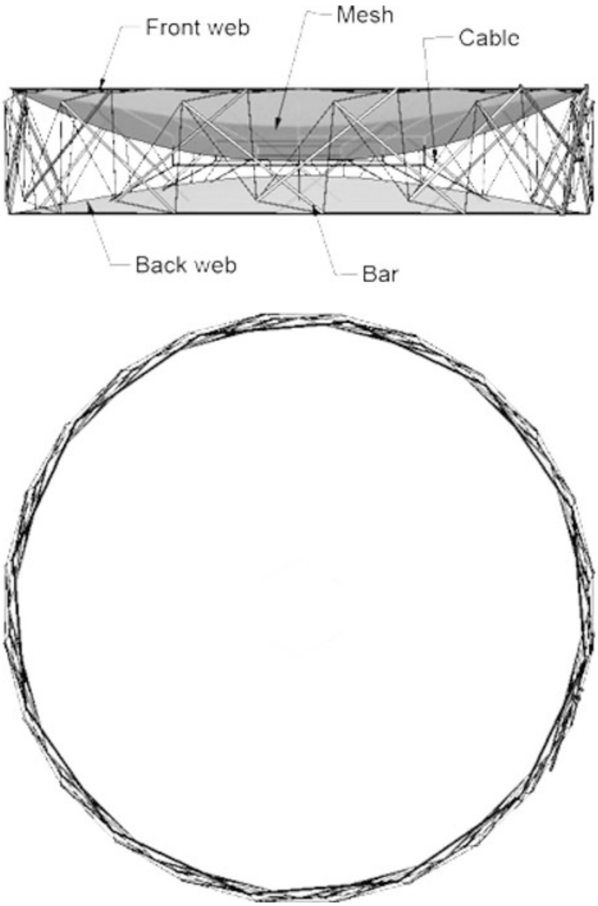


Fig. 23 TSS Flight Model deployed (TSS-to-Boom I/F not shown in the picture)

4.2 *Flight Model Design*

A preliminary design of the Flight Model (FM) of the TSS was performed, with the aim of investigating the expected physical and structural properties of the TSS when materials easily available on the market are used. The Flight Model is composed of the following elements: cables, bars, front and back web (in light gray in Fig. 23), reflective mesh (in heavy gray), deployment actuation system, tensioning actuation system, and TSS-to-Boom (spacecraft) apparatus I/F. The Flight Model is 12 m in diameter and 2.6 m in height in its deployed configuration, 2.33 m in diameter and 4.53 m in height when folded. All the 24 TSS TR bars are of the same fixed length. The overall calculated mass is 58 kg, including all the above mentioned elements and an additional 10 % margin to take into account unavoidable uncertainties at this stage of design. The front and back webs are fastened to the top and bottom

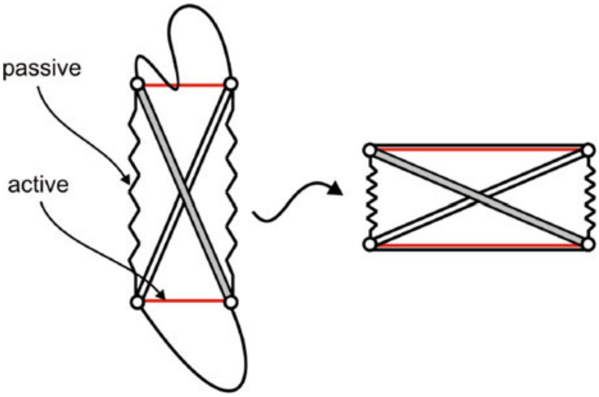


Fig. 24 A simple 2D structure exemplifying Deployment Phase 1. In an SR, the active and passive elements are, respectively, the hoop cables and the vertical cables. Polygonal cables have fixed length; they are slack before completion of Deployment Phase 1

polygons of the TR; moreover, they are linked to each other by means of tension elements, called tension ties. The reflecting mesh is fastened to the top web by means of tension elements distributed all over its surface, so as to give it the required working shape. The TR is composed of groups of cables identified as specified in Sect. 3.5 and shown in Fig. 25. Notice the additional group consisting of two continuous cables, henceforth referred to as the hoop cables, running in parallel to the top and bottom polygons, whose service function is explained below. Recall that some of the cables maintain a fixed length both in stowed and in deployed configuration (except of course for the modest lengthening due to tension), while the other cables change their length during deployment: precisely, vertical cables shorten, hoop cables lengthen. A vertical cable is shortened by pulling it inside a bar tube, by means of the deployment passive actuator described below; the cable portion remaining outside the bar after shortening is visible in Fig. 25.

The two hoop cables, the one running through the top-polygon nodes the other running through the bottom-polygon nodes, are lengthened by unwinding them from pulleys driven by electrical motors (the deployment active actuators) with controlled speed. Their function is to regulate the deployment speed during Deployment Phase 1: at the end of Deployment Phase 1, they become slack and have no structural role in the fully deployed configuration. On the contrary, polygonal cables are slack during Deployment Phase 1 and become in tension at the end of Deployment Phase 1 (see Fig. 24 for an illustration of such deployment strategy in a simple 2D example). They inherit the structural role of the hoop cables, starting from Phase 2 of deployment and, later, in the fully deployed configuration (polygonal and hoop cables appear overlapped in Fig. 25). Finally, diagonal and saddle cables are always in tension, both in the folded and deployed configurations and during deployment).

Deployment is implemented by the actuation systems mentioned above. There are 24 passive deployment actuators (one inside each tubular bar), which pull

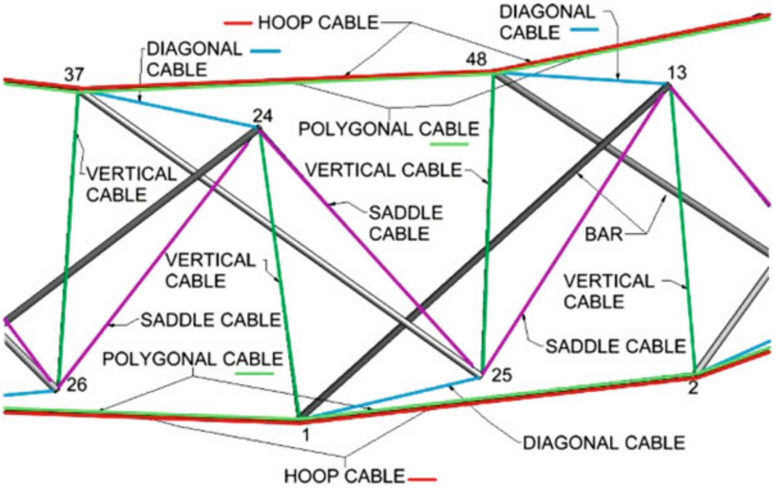


Fig. 25 TSS cables nomenclature (*close-up view* of a portion of the TSS)

vertical cables inside the tubes; by means of pre-loaded springs, they provide the force needed during Phase 1. Each of the two active deployment actuators consists of a rotating electrical motor and a pulley, where a hoop cable is coiled in the folded configuration they unwind the hoop cables, making sure that deployment proceeds in a smooth way by controlling the deployment speed (in their absence, Phase 1 would last only a few seconds, due to the action of the pre-loaded springs, and this could cause uncontrolled perturbations not only of the TSS but also of the spacecraft). Phase 1 ends when passive actuators have come to the end of their strokes, locking devices have reached the locked position, and hoop cables are completely unwound (the locking devices fix the position inside the bars of the endpoints of vertical cables, henceforth keeping their length fixed). At the end of Phase 1, the TSS has shape and dimensions close to the final ones; however, its stiffness is still low, because the cables do not have the design tension yet: specific actuators take care of this, during Phase 2. The three tensioning actuators are mounted 120° apart in the top polygon, so as to apply the required tension to three of the diagonal cables, and hence to all the dependent cables. Tensioning actuators apply tension by reducing the distance between the points to which the diagonal cables are fastened. As a consequence, during Deployment Phase 2 the TSS geometry is slightly modified.

The TSS Flight Model is attached to the spacecraft boom by means of an interface structure denoted by I/F, consisting of a plate (where the boom is attached) and three arms connected to three nodes of the TSS. Three cylindrical hinges and three spherical hinges are used to connect the arms to the plate and to the TSS (see the sketch in Fig. 26). The two active deployment actuators that unwind the hoop cables during Phase 1 are also mounted on the I/F.

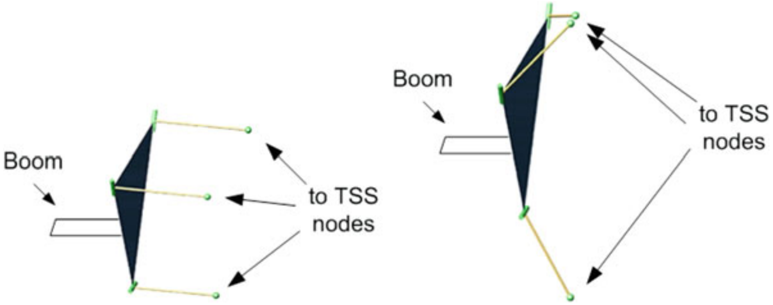


Fig. 26 TSS-to-Boom I/F. *Left*: deployed configuration; *right* folded configuration

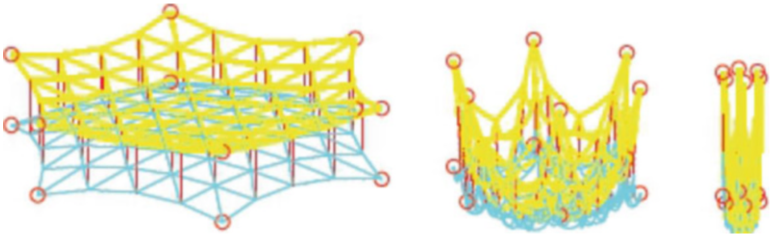


Fig. 27 Simulation of the RF mesh supporting web configuration (*red lines* represent the tension ties). *Left*: fully deployed ($D = 12\,\mathrm{m}$). *Center*: partially folded ($D = 6\,\mathrm{m}$). *Right*: fully folded ($D = 0.6\,\mathrm{m}$)

An important role in the TSS functions is assigned to the reflective mesh and to the web. The material of choice for the radio-frequency (RF) reflective surface must have low density and be easily foldable into a compact shape. The most common surface material for space reflectors of moderate precision is a mesh knitted from metallic or synthetic fibers plated with RF reflective material. The mesh must be sufficiently compliant to match without wrinkling the web's double-curvature surface. As the most recent studies suggest [8], 5 N/m is a mesh-tension value sufficient for operating frequencies up 14 GHz. Since earlier studies also find this value suitable, we selected it as the nominal tension in the reflective mesh of our antenna. The relevant web configuration was analyzed (for the dimensions of triangle sides and web tension, see Fig. 27).

The tension-truss concept requires that the triangulated web is put under tension by loads approximately perpendicular to the surface of the antenna. The tension-truss concept is used in several antennas, currently operating in orbit. Its main advantage is that the geometric accuracy of the paraboloidal surface can be increased without any need to change the configuration of the supporting ring structure. The configuration of tension ties for the TSS was analyzed (e.g., axial/non-axial tension ties), and deployment simulations were performed. Our analyses suggested to avoid non-axial tension ties. To conform to the no-elongation and easy-tensioning requirements, a tension-tie configuration was identified and studied for a five-ring web assembly. This solution, which is in our opinion the simplest one, can be adopted also for a larger number of rings.

Mesh folding and stowage is critical and should be studied in detail, as for state-of-the-art large reflectors. Mesh development activities include tests to characterize mesh mechanical properties including tendency to self-adhesion. The absence of external mechanical joints is considered advantageous also in relation to reducing the risk of mesh entanglement. The launch regime will be addressed by designing a suitable hold-down and release system for the deployable boom plus reflector dish assembly. There will be primary hold-down mechanisms to hold-down the deployable boom to the spacecraft lateral panel, and secondary hold-down mechanisms to restrain the reflector dish in its folded state and release it when boom deployment is completed.

In Europe, ESA [9] has already pre-qualified a deployable boom system with associated motorized deployment mechanisms and hold-down release system for a large reflector antenna of 12 m aperture. The problem of designing the mechanical connection between the reflector dish and the deployable boom has been addressed and included in the present development.

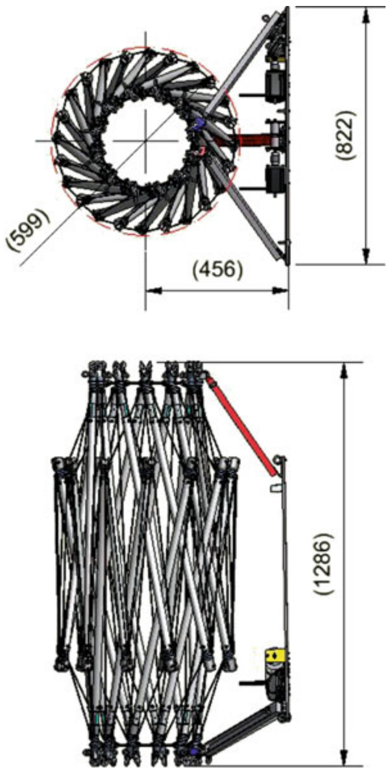
4.3 Breadboard Design

We performed a detailed design of a breadboard (BB) having all of the main structural features of the Flight Model described above. The TSS BB consists of the following main components: cables, bars, a simplified web consisting of radial cables, deployment actuation system, tensioning actuation system, and TSS-to-Boom I/F (Figs. 28 and 29). The BB is a scaled version of the TSS flight concept, designed according to the following rules:

- the polygon has the same number of sides (12) as the flight concept;
- the scaling ratio 1-to-4 applies to the overall deployed dimensions;
- the dimensions of the components (e.g., joints, cable, and bar cross-sections) may not be equally scaled.

The rigid parts of the BB were made mainly of aluminum and stainless steel; for cables, Vectran$^{\text{TM}}$ was used; cable terminals were realized with the use of thimbles and ferrules. Bars are composed of a tube and two joints, one for each bar end. The two joints of a bar are obtained by assembling machined parts, and include the interfaces between that bar and all the relative cables. Each bar includes, inside the tube, a passive deployment actuator, used to shorten a vertical cable. Such an actuator pulls inside the bar a portion of the cable, shortening the cable portion external to the bar. During deployment, the cable is retracted into the bar, so that the distance between the two bars connected by that cable is reduced (for this reason, such a cable is also referred to as a shortening cable). The 24 passive actuators inside the bars provide, by means of compression springs, the force needed to deploy the structure in the course of Phase 1. Each passive actuator includes a locking

Fig. 28 TSS BB folded dimensions, in mm (the web is not shown in this figure)

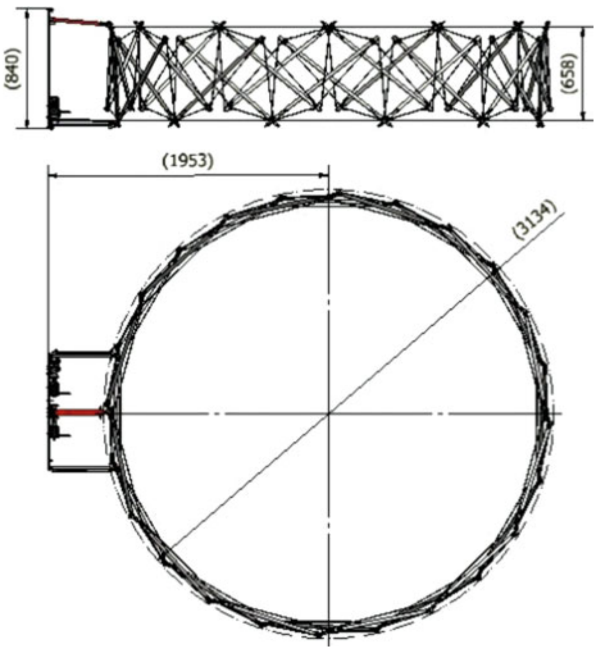


device, which is needed to lock the shortening cable (vertical cable) into position and to fix its length, after Phase 1 is completed. The two joints located at a bar's ends are different, because the cables they join have different roles, and also because there is a cable that enters the bar tube at only one of its ends and is pulled by the passive actuator during deployment. The two joints are called joint A and joint B, with the cable being retracted into joint B.

The BB web consists of two sets of radial cables, joining the top-polygon vertices with the top-polygon center point and the bottom-polygon vertices with the bottom-polygon center point. Two discs collect, respectively, the top radial and the bottom radial cables; they are connected by an elastic member called the tension tie (see Fig. 30).

The BB was provided with a gravity compensation system (GCS), to reduce as much as possible gravity effects during deployment. The GCS is composed of an aluminum plate, called GCS plate, fixed to the ceiling of the laboratory, and of the cables by which the BB is attached to the GCS plate. 12 out of 24 of the BB bars are attached to the GCS plate. The three tensioning actuators and the TSS-to-Boom I/F are also attached to the GCS plate. The TSS-to-Boom I/F is attached to the GCS plate by means of three cables. GCS cables are composed of series of springs and a rope cable (of the same material used for the BB cables). The

Fig. 29 TSS BB deployed dimensions, in mm (the web is not shown in this figure)



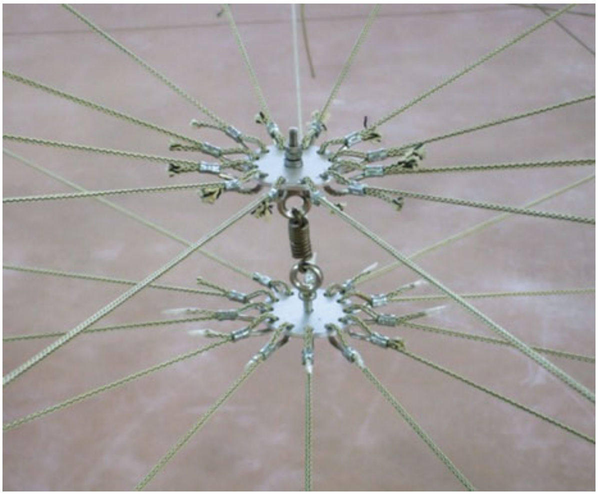


Fig. 30 TSS BB Web. The single tension tie includes a spring

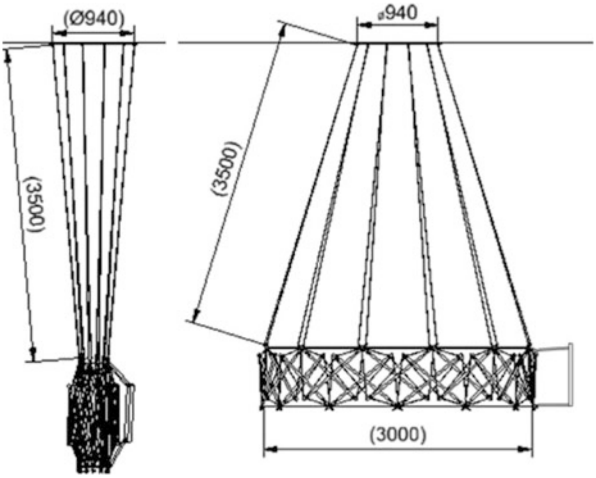


Fig. 31 TSS BB attached to the GCS. *Left*: folded, *right*: deployed

number and the elastic properties of the springs are selected so as to decouple the natural frequency due to GCS cables from the natural frequency of the TSS ring (in particular, the springs that equip the suspension cables provide a natural frequency of about 0.5 Hz in the vertical direction). Figure 31 shows the BB attached to the GCS; the relevant reference dimensions are indicated; it is also shown how the TSS-to-Boom I/F modifies its shape on unfolding. In the unfolded configuration, the horizontal component of the GCS constraining force applied to the BB ring is about 20 % (peak value) of the vertical one. A moving mass is used to compensate the radial component of the TSS-to-Boom I/F weight force.

5 Breadboard Test Campaign

A test campaign was performed on the breadboard described above, including:

1. BB Geometry and Shape Test;
2. BB Performance Test (deployment and folding-up);
3. BB Structural Test (stiffness);
4. BB Stop-and-Go Test.

Figures 32 and 33 show the TSS BB attached to the GCS cables; the TSS-to-Boom I/F is visible on the left.

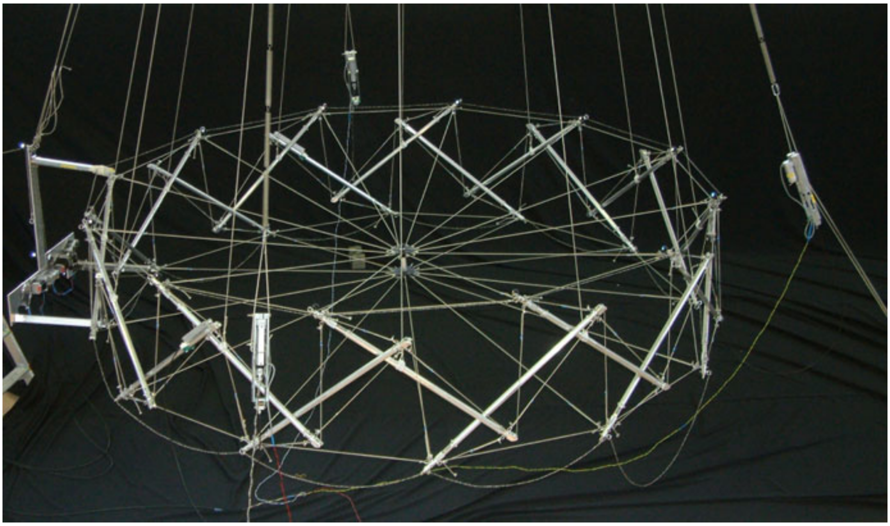


Fig. 32 TSS BB attached to the GCS. Deployed configuration—*top-side view*

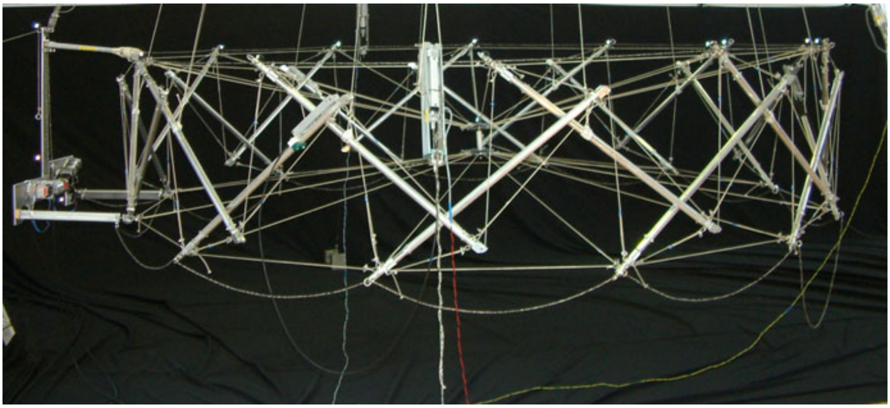


Fig. 33 TSS BB attached to the GCS. Deployed configuration—*side view*

5.1 BB Geometry and Shape Test

This test was aimed at measuring the geometrical-shape repeatability of the structure in the deployed configuration. The position of some points of the deployed structure after different stowing/deployment sequences was measured and the relevant differences in position between one stowing/deployment sequence and the others were calculated (post-processing). Three folding/deployment sequences were performed and the geometry data acquired (three repetitions). A total station (laser measurement) was used to acquire the position of 15 markers placed on the BB. The data were elaborated in two ways:

- Calculating the distances of all the marker pairs and the relevant statistics (mean and standard deviation). The calculated mean of the standard deviation for markers located on the top polygon's sides was 0.34 mm.
- Calculating by orthogonal regression the fitting planes for markers placed on the TR top polygon's nodes. For the point distances from the fitting plane calculated for the three acquisitions, this elaboration showed a variance between 0.03 and 0.36 mm and a standard deviation between 0.16 and 0.6 mm.

All in all, the test showed a good repeatability of the folding/deployment process.

5.2 Breadboard Performance Test

The aim of this test was to verify that the deployment of the structure worked smoothly, with no bar and/or cable entanglements. Five folding/deployment complete sequences were performed. One entanglement only occurred, during sequence no. 4, due to the wrong folding of one of the cables that prevented complete deployment.

5.3 Breadboard Structural Test and analysis

The aim of this test was to measure the natural frequencies of the BB. Two triaxial accelerometers were placed on the structure and the response to in-plane and out-of-plane perturbations of the ring was recorded. The in-plane perturbation was introduced by means of a rope passing through diametrically opposite bars ends of the top and bottom polygons. The rope length was such as to reduce the diametral distance of the connected bar ends (i.e., ring forced to an elliptical shape). The rope was then cut causing the perturbation in the radial direction. The out-of-plane perturbation was introduced by constraining a bar end of the ring structure to the ground, so as to force the ring into a cantilever-like bent shape. The rope was then cut, causing a bending-mode perturbation.

In addition to the 0.5 Hz design frequency of the GCS in vertical direction, the next eight measured frequencies were at 1.2, 2.7, 5.4, 7.8, 10.3, 11.1, 13, and 13.8 Hz. Figures 34 and 35 show the recorded power spectrum relevant to, respectively, in-plane perturbation and out-of-plane perturbation.

A structural analysis was performed before and after the test campaign. Besides the frequencies relevant for the GCS, the analysis indicated that two out-of-plane natural frequencies (at, respectively, 11.0 and 11.2 Hz) affected all nodes in a bending motion of the annular structure. Note that, as per visual inspection, the various types of modes are often coupled to each other, due to the fact that

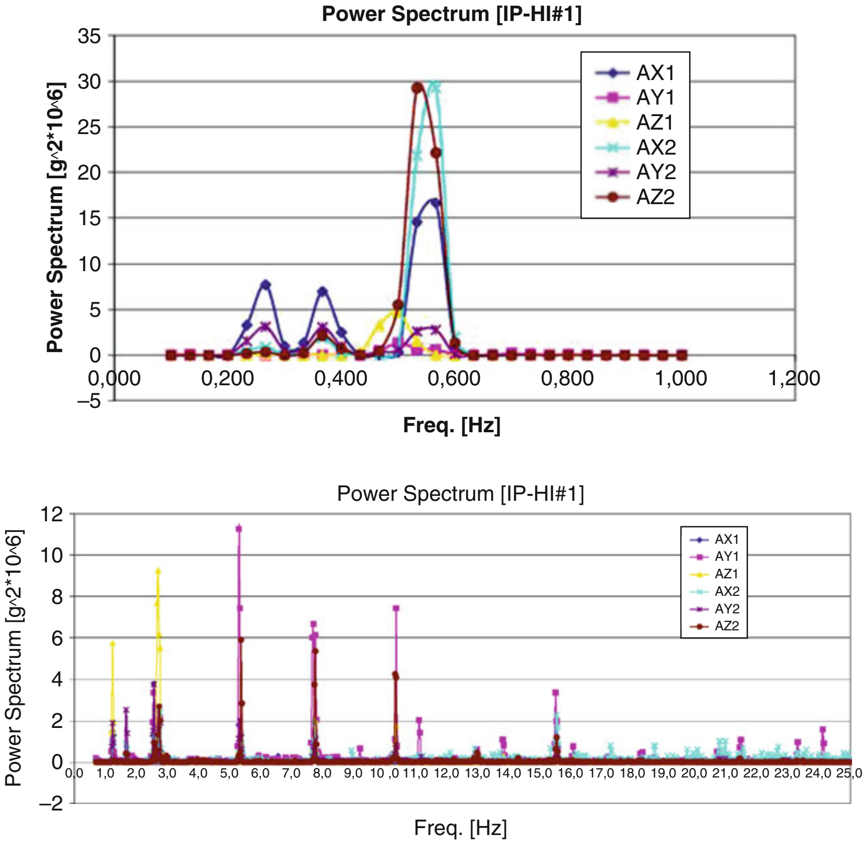


Fig. 34 TSS BB—recorded power spectrum vs. frequency. In-plane perturbation

frequencies are close to each other. This can be seen, for example, in Fig. 36 left, where the out-of-plane bending of the TSS is coupled with the transversal vibration of the GCS supporting the IF.

The in-plane modes involve intermediate nodes only, without affecting nodes at the vertices of the base polygons. In these modes, the motion of the intermediate nodes is directed radially in the horizontal plane. The 17 calculated frequencies are in the range between 6.7 and 14.1 Hz. The structural analysis also shows that the modes associated with the GCS correspond to the first peaks appearing in the power spectrum from the tests. The correspondence is quite clear for frequencies of about 0.5, 1.2, 2.7, and 5.4 Hz. The frequency of the first modes involving intermediate nodes (about 7, 8 and 10 Hz) are located in proximity of the peaks of the spectrum obtained from the tests. A correspondence between the frequency of the first out-of-plane bending mode at 11 Hz and relevant peak in the spectrum is also visible.

The results of the analysis are in a fairly good agreement with those of the test, even though the dynamic response of the BB can, in principle, be coupled with

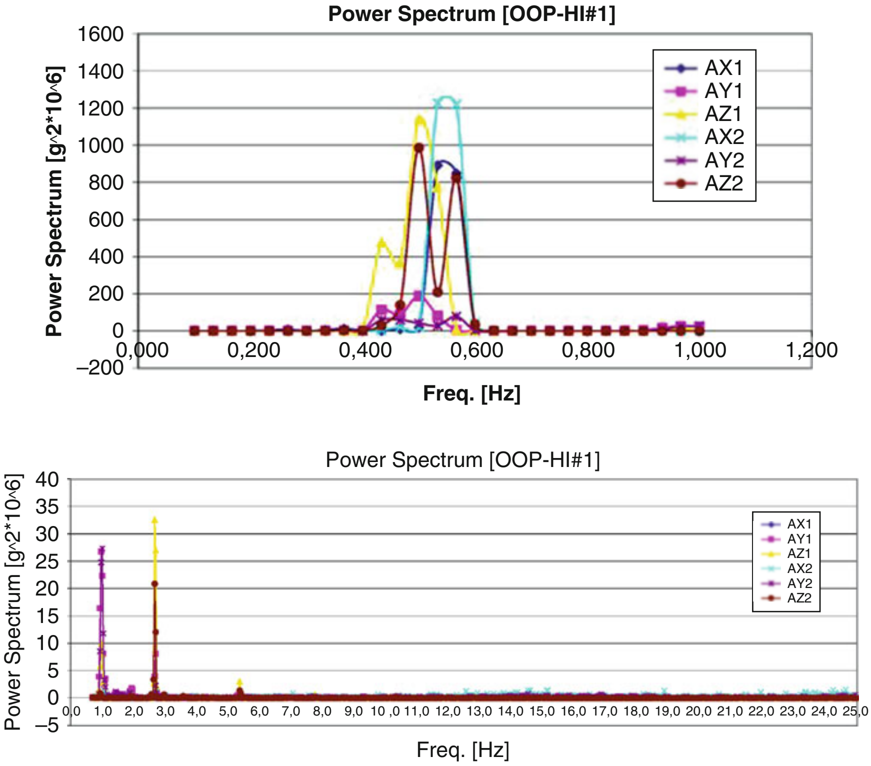


Fig. 35 TSS BB—recorded power spectrum vs. frequency. Out-of-plane perturbation

that of the GCS. A modal analysis in the absence of gravity was performed for both the BB and the FM. In both cases, the first mode is an out-of-plane cantilever-like bending mode, with frequency of 1.9 Hz for the BB and 2.1 Hz for the FM. In consideration of the fairly good agreement between tests and numerical simulations, these results show that the FM should have good dynamic performance, since its first natural frequency is not only higher than 1 Hz but also indeed far away from this value.

5.4 Breadboard Stop-and-Go Test

This test was aimed to demonstrate the capability of the TSS BB to complete deployment even if a stop occurs during deployment. The deployment was started and stopped after 30 s, before Phase 1 was completed (in nominal conditions, Phase 1 is completed in 2 min). After a 60 s stop, deployment was re-started until a successful completion, including Phase 2.

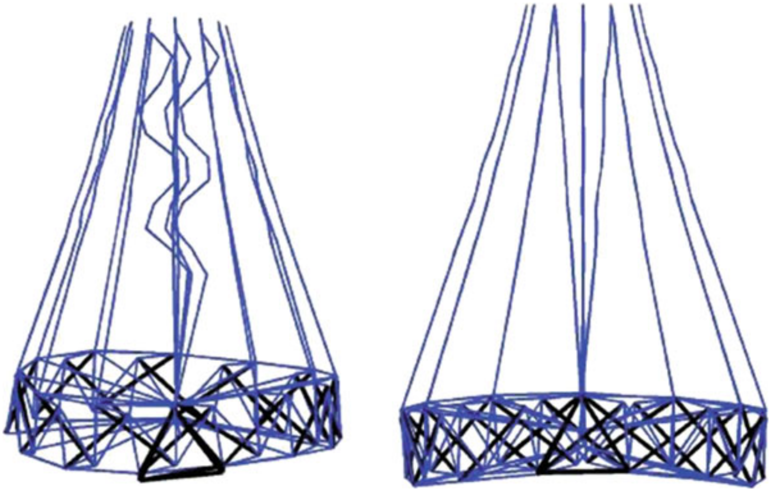


Fig. 36 *Left*: out-of-plane bending, coupled with transversal vibrations of the GCS, 11.0 Hz. *Right*: Out-of-plane bending, 11.2 Hz

6 Conclusions

The successful development of a new architectural concept of a large deployable reflector (about 12 m in diameter) for space applications has been achieved and presented in this paper. By exploiting "tensegrity" structural principles, a large deployable ring has been conceived, which does not include any mechanical joint or articulation between its rigid members, which are interconnected only by cables. Having no mechanical joints in the expandable ring, therefore eliminating a major potential source of single-point failures, constitutes a major advantage in terms of deployment reliability, a crucial requirement for such systems. The new architecture has been studied in detail, and a reduced-scale breadboard model (3 m in diameter) has been realized and tested to validate the main design features. Means to interface the expandable ring to the hosting spacecraft have been studied in detail, resulting in an innovative and very efficient solution. A suitable gravity off-loading system has been designed and implemented for the test campaign of the reflector breadboard. All major design assumptions and features have been validated during the test campaign, namely: deployment functionality (including "Stop-and-Go" deployment verification); deployment accuracy/repeatability; and stiffness in deployed configuration. The newly conceived architecture, which has been protected by an international patent filing, is a potential candidate for further development studies to reach higher technology readiness level (TRL) as well as, possibly, for an in-orbit deployment demonstration. ESA has established a roadmap to increase large deployment reflector TRL status [9] and a research and development activity has recently started for the development of a suitable RF mesh.

References

1. Burkhardt, B.: Synergetics gallery. http://bobwb.tripod.com/synergetics/photos/index.html. Cited Feb 2016
2. Datashvili, L., Maghaldadze, N., Endler, S.: Advances in mechanical architectures of large precision space apertures. In: Proceedings of the 13th European Conference on Spacecraft Structures, Materials & Environmental Testing, 1–4 Apr 2014, Braunschweig
3. ESA webpage on future missions http://www.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Earth_Explorers/Future_missions/About_future_missions. Cited March 2016
4. Gómez Jáuregui V.: Controversial origins of tensegrity. In: Domingo, A., Lazaro, C. (eds.) Proceedings of the IASS Symposium 2009, Valencia (2009)
5. Harris Corporation: Unfurlable Antenna Solutions. http://govcomm.harris.com/solutions/spacesystems/unfurlablemeshantennareflectors.aspx. Cited Feb 2016
6. Hernandez, S.J., Mirats Tur, J.M.: Tensegrity frameworks: Static analysis review. Mech. Mach. Theor. **43**, 859–881 (2008)
7. Li, B., Luo, A., Liu, R., Tao, J., Liu, H., Wang, L.: Configuration modeling and interior force analysis of deployable tensegrity. In: Proceedings of the 14th IFToMM World Congress, Taipei, 2015, pp. 126–131 (2015)
8. Lubrano, V., Mizzoni, R., Silvestrucci, F., Raboso, D.: PIM characteristics of the large deployable reflector antenna mesh. In: 4th International Workshop on Multipactor, Corona and Passive Intermodulation in Space RF Hardware, Noordwijk (2003)
9. Mangenot, C., Santiago-Prowald, J., van't Klooster, K., Fonseca, N., Scolamiero, L., et al.: Large Reflector Antenna Working Group, Executive Report, TEC-EEA/2010.636/CM, ESA, Estec, Noordwijk (2010)
10. Micheletti, A.: The indeterminacy condition for tensegrity towers, a kinematic approach. Revue Française de Génie Civil **7**, 329–342 (2003)
11. Micheletti, A.: Bistable regimes in an elastic tensegrity systems. Proc. R. Soc. A **469**, 20130052 (2013)
12. Micheletti, A., Williams, W.O.: A marching procedure for form-finding for tensegrity structures. J. Mech. Mater. Struct. **2**, 857–882 (2007)
13. Northrop-Grumman: Astromesh^{r}. http://www.northropgrumman.com/businessventures/astroaerospace/products/pages/astromesh.aspx. Cited Feb 2016
14. Oppenheim, I.J., Williams, W.O.: Tensegrity prisms as adaptive structure. In: Adaptive Structures and Material Systems, vol. 54, pp. 113–120. ASME, Dallas, TX (1997)
15. Peng, Z., Yuan, X., Dong, S.: Tensegrity torus. In: Proceedings of the IASS-ACPS 2006, Beijing (2006)
16. Scialino, G.L., Salvini, P., Migliorelli, M., Pennestrì, E., Valentini, P.P., van't Klooster, K., Santiago Prowald, J., Rodrigues, G., Gloy Y.: Structural characterization and modelling of metallic mesh material for large deployable reflectors. In: Proceedings of the Advanced Lightweight Structures and Reflector Antennas, 1–3 Oct 2014, Tbilisi, GA (2014)
17. Scolamiero, L., Zolesi, V.S., Ganga, P.L., Podio-Guidugli, P., Micheletti, A., Tibert, A.G., in the name of European Space Agency, International Patent for "A Deployable Tensegrity Structure, Especially For Space Applications", Patent Application: PCT/IB2012/051309, filed Mar 2012
18. Snelson, K.D.: personal webpage http://www.kennethsnelson.net. Cited Feb 2016
19. Stern, I.: US Patent for a "Deployable reflector antenna with tensegrity support and associated method", No. 6,542,132, Apr 1, 2003
20. Thomson, M.W., Marks, G.W., Hedgepeth, J.M.: US Patent for a "Light-weight reflector for concentrating radiation", No. 5,680,145, Oct 21, 1997
21. Tibert A.G., Pellegrino, S.: Review of form-finding methods for tensegrity structures. Int. J. Space Struct. **18**, 209–223 (2003)

22. Yuan, X., Peng, Z., Dong, S., Zhao, B.: A new tensegrity module - 'torus'. Adv. Struct. Eng. **11**, 243–252 (2008)
23. Zolesi, V., Ganga, P., Scolamiero, L., Micheletti, A., Podio-Guidugli, P., Tibert, A.G., Donati, A., Ghiozzi, M.: On an innovative deployment concept for large space structures. Proceedings of 42nd International Conference on Environmental Systems (ICES), AIAA 2012-3601, San Diego, CA, 15–19 July 2012

Data Analytic UQ Cascade

Bijan Mohammadi

Abstract This contribution gathers some of the ingredients presented at Erice during the third workshop on "Variational Analysis and Aerospace Engineering." It is a collection of several previous publications on how to set up an uncertainty quantification (UQ) cascade with ingredients of growing computational complexity for both forward and reverse uncertainty propagation. It uses data analysis ingredients in a context of existing deterministic simulation platforms. It starts with a complexity-based splitting of the independent variables and the definition of a parametric optimization problem. Geometric characterization of global sensitivity spaces through their dimensions and relative positions through principal angles between vector spaces bring a first set of information on the impact of uncertainties of the functioning parameters on the optimal solution. Joining the multi-point descent direction and probability density function quantiles of the optimization parameters permits to define the notion of directional extreme scenarios (DES) without sampling of large dimension design spaces. One goes beyond DES with ensemble Kalman filters (EnKF) after the multi-point optimization algorithm is cast into an ensemble simulation environment. This formulation accounts for the variability in large dimension. The UQ cascade continues with the joint application of the EnKF and DES leading to the concept of ensemble directional extreme scenarios which provides a more exhaustive description of the possible extreme scenarios. The different ingredients developed for this cascade also permit to quantify the impact of state uncertainties on the design and provide confidence bounds for the optimal solution. This is typical of inverse designs where the target should be assumed uncertain. Our proposal uses the previous DES strategy applied this time to the target data. We use these scenarios to define a matrix having the structure of the covariance matrix of the optimization parameters. We compare this construction to another one using available adjoint-based gradients of the functional. Eventually, we go beyond inverse design and apply the method to general optimization problems. The ingredients of the paper have been applied to constrained aerodynamic performance analysis problems.

B. Mohammadi (✉)
Instiut Montpellierain Alexander Grothendieck, Montpellier, France
e-mail: bijan.mohammadi@umontpellier.fr

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_12

1 Context

Our domain of interest is aerodynamic shape optimization. The questions of interest are

- can we propose an aircraft shape designed to have similar performances over a given range of some functioning parameters (to be formulated through the moments of a functional)?
- can we do that modifying as less as possible an existing mono-point optimization shape design loop?
- is it possible for the time-to-solution cost of this parametric shape design to remain comparable to the mono-point situation?

We consider a generic situation where the simulation aims at predicting a given quantity of interest $j(\mathbf{x}, \alpha)$ and there are a few functioning or operating parameters α and several design parameters $\mathbf{x}$ involved. The ranges of the functioning parameters define the global operating/functioning conditions of a given design. This splitting of the independent variables in two sets is important for the sequel.

We propose a cascade of ingredients to account for uncertainties avoiding any sampling of large dimensional spaces. A sampling will be only necessary for the functioning parameters $\mathbf{u}$ range leading to a multi-point optimization problem.

The literature on uncertainty quantification (UQ) is huge. In short, forward propagation aims at defining a probability density function (PDF) for j knowing those of $\mathbf{x}$ and α [22, 29, 43]. This can be done, for instance, through Monte Carlo simulations or a separation between deterministic and stochastic features using Karhunen–Loeve theory (polynomial chaos theory belongs to this class) [18, 19, 23, 54, 58]. Examples of shape optimization with polynomial chaos and surrogate models during optimization are given in [6, 7].

Backward propagation aims at reducing models bias or calibrating models parameters knowing the PDF of j (or other constraints and observations) [4, 20, 52]. This can be seen as a minimization problem and Kalman filters [28] give, for instance, an elegant framework for this inversion assimilating the uncertainties on the observations.

Our aim is to propose a geometric framework to address the curse of dimensionality of existing approaches related to the explosion of their computational complexity due to the sampling necessary to access probabilistic information, even if this can be improved with intelligent sampling techniques [3, 50]. The different ingredients presented here can be applied with either high-fidelity or reduced order models, when available [42, 45, 48, 53]. Low-order models are often used instead of the full models to overcome the computational complexity of UQ.

After the splitting of the independent variables mentioned above, we define a multi-point formulation to account for the variability on α. This is feasible because the size of α is assumed small. We define a global sensitivity space using the sensitivities of j with respect to $\mathbf{x}$ for the multi-point problem. Once this space built, we analyze its dimension. We previously showed how to perform this task and how to use this information for adaptive sampling [15, 31].

The next step is to analyze the impact of different modeling or discretizations on the results. Different models or solution procedures lead to different sensitivity spaces. Beyond their respective dimensions, principal angles between the respective sensitivity vector spaces permit to measure the deviation due to such changes. The dimensions of the spaces and the angles are interesting measures for both the epistemic and aleatory uncertainties. Indeed, suppose that, for a given model the dimensions of the sensitivity spaces remain unchanged when enriching the sampling of the functioning parameter range. This stability would be a first indication of a low level of sensitivity of the simulations with respect to this parameter. Once this is established, principal angles between subspaces permit to analyze both the impact of a given evolution of the modeling on the sensitivity spaces and an enrichment of our sampling. Eventually, constant dimension and low angles will clearly indicate a situation of low uncertainty.

These ingredients can be used in a context of multi-point robust analysis of a system to define worst-case scenarios for its functioning. To this end we combine a multi-point sensitivity with the probabilistic features of the control parameters through their quantiles [27, 34] to define the concept of directional extreme scenarios (DES) without a sampling of large dimension design spaces.

ensemble Kalman filters (EnKF) [1, 12, 13, 16, 17, 28] permit to go beyond the directional uncertainty quantification concept when accounting for the uncertainties in large dimension. They also permit backward uncertainty propagation assimilating the uncertainty on the functional and constraints during the design. We cast our multi-point optimization problem into the ensemble formulation. Joint application of the EnKF and DES leads to the concept of ensemble directional extreme scenarios (EDES) which provides a more exhaustive description of possible extreme scenarios.

Despite these approaches avoid the sampling of a large dimensional space the computation cost remains high and the procedures difficult to simply explain in engineering environments. We propose a low-complexity approach for the inversion of uncertain data where the target state $\mathbf{u}^*$ used in an inverse problem is uncertain. In this situation, we consider functional of the form $j(\mathbf{x}, \alpha, \mathbf{u}^*) = \|\mathbf{u}(\mathbf{x}, \alpha) - \mathbf{u}^*(\alpha)\|$ to reduce the distance between a model state $\mathbf{u}(\mathbf{x}, \alpha)$ and observations.

Targeting uncertain data is a realistic situation as the acquired data are usually uncertain. It is therefore interesting to be able to quantify the impact of this uncertainty on the inversion results. An important information will be the sensitivity of the design to a given level of uncertainty on the data at some location. Indeed, if this sensitivity is low, this would be an indication that a more accurate acquisition there is unnecessary.

Considering the target as uncertain is also interesting because we do not always have existence of a solution for an inversion problem as $\mathbf{u}^*$ is not necessary solution of the state equation making an exact or deterministic inversion pointless. Also, the approach permits to go beyond inversions based on least square minimization involving a mean state target.

Finally, the uncertainty in measurements is also an interesting way to account for the presence of variability in the state (e.g., due to the presence of turbulence in the flow). More generally, as the model and numerical procedures are by nature imperfect and partial, we can consider this uncertainty as a representation or estimation of these imperfections. These imperfections are even more present in inverse problems where one cannot afford the same level of resolution than for a single simulation. We therefore need to be able to quantify the impact of these weaknesses on the design. The approach presented here is therefore also useful to account for epistemic uncertainties related to possible model or solution procedures deficiencies.

Concerning the computational cost of these analyses, one can say that, when using the same calculation ingredients than in a high-fidelity simulation (i.e., without calling for low-order models or cheaper discretizations), the best calculation complexity one might think of for a simulation under uncertainty is when its cost is comparable to the deterministic situation. This is clearly unreachable except if all the extra effort can be achieved in a fully parallel manner and parallel to the initial deterministic calculation in order for the time to solution to remain unchanged when accounting for the presence of uncertainties. This is the case with Monte Carlo approaches. But these are quite expensive and do not take advantage of available simulation environments. In particular, when an adjoint-based optimization environment exists. Our proposal consists of upgrading existing platforms without abandoning what has been built for the deterministic situations and with keeping the time to solution unchanged in the presence of uncertainties with two sources of parallelism coming from the multi-point formulation to account for the uncertainties on the functioning parameters and from the EnKF formulation for those on the optimization variables and observation data.

2 Parametric Optimization

We are interested in a class of optimization problems where the cost function involves a functioning parameter α not considered as a design parameter:

$$\min_{\mathbf{x}\in\mathbf{O}_{ad}} j(\mathbf{x},\alpha), \quad \alpha \in \mathbf{I} \subset \mathbb{R}^n, \mathbf{O}_{ad} \subset \mathbb{R}^N. \tag{1}$$

where $\mathbf{x}$ is the design vector belonging to $\mathbf{O}_{ad}$ the optimization admissible domain. Usually, the functioning parameters (or operating conditions) α are just a few. On the other hand, the size N of $\mathbf{x}$ is usually large. Together, $\mathbf{x}$ and α fully describe our system and we have $n << N$. This splitting between functioning parameters (or operating conditions) and design variables is central to our discussion.

In [31, 32] we showed how to use multi-point optimization to address such optimization problem. The aim is to remove, during optimization, the dependency in α. This is done minimizing a functional $J(\mathbf{x})$ encapsulating this dependency expressed through $\mathbf{A} = \{j(\mathbf{x}, \alpha_k), \alpha_k \in \mathbf{I}_M\}$ over $\mathbf{I}_M$ a given sampling of $\mathbf{I}$:

$$J = \mathbf{J}(\mathbf{A}), \quad \text{such that} \quad \mathbf{G}(\mathbf{A}) \leq 0. \tag{2}$$

Several choices are possible for **J** and **G** to address the issue of robust design. For instance, following Taguchi's definition, one can look for minimal-variance design or only a given level for the variance. Indeed, a classical approach to extend single-point design and improve off-design points is to control μ mean performance and σ variance of the functional [51] as in first order second moment methods [30]. One can also look for information about the tails of the distributions which can be linked to the variance in the Gaussian framework and we use this relationship in quantile-based extreme scenarios.

Often it might be interesting to go beyond the first two moments and in particular consider the first four moments of j during the design. Going beyond the first two moments is important when the PDF of j deviates from a pure Gaussian distribution. Indeed, even with interval-based (with uniform PDF) or Gaussian entries there is no reason the PDF of the solution of a simulation to remain uniform or Gaussian.

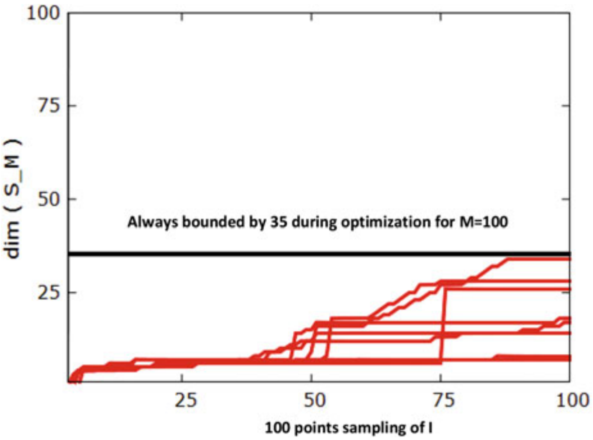
The third and fourth moments are the skewness γ and the kurtosis κ. One can consider that a robust design should favor symmetry in the distribution which means lower absolute value of skewness. For instance, in a Gaussian distribution we have $\gamma = 0$. Also, in a normal distribution the mean and median coincide and if a PDF is not too far from a normal distribution, the median will be near $\mu - \gamma\sigma/6$. Therefore, if $|\gamma| \to 0$ the PDF tends toward a normal distribution. This provides an inequality constraint on $|\gamma|$ as γ can be positive or negative. For a unimodal PDF a reduction of the skewness comes when the mean and the mode of the distribution converge toward each other at given standard deviation.

Concerning the fourth moment, a robust design should favor higher density near the mean which means higher kurtosis, but this is more subtle. Indeed, despite higher kurtosis means concentration of the probability mass around the mean, it could also imply thicker tails in the PDF. This means that more of the variance is the result of infrequent extreme deviations. We need therefore to define what we mean by robust design: acceptance of frequent modest deviations or acceptance of infrequent extreme ones. If operational security is a major concern the latter should be definitely avoided. Hence, a reasonable requirement would be to have a design reducing the initial kurtosis value: $\kappa \leq \kappa_0$ together with a constraint on the variance σ.

3 Gradients, Sensitivity Spaces, and Admissible Search Directions

Monte Carlo simulations permit to recover these moments with an error decreasing as $\sigma M^{-1/2}$ with M the number of functional evaluations and this rate is independent of n. But, for small n, classical numerical integration over-performs Monte Carlo

Fig. 1 Histories of Gram–Schmidt orthonormalization of $\{\nabla_x j(x, \alpha_k),\ \ \alpha_k \in \mathbf{I}_{100}\}$ during optimization. The dimension of the global search space S_M always remains below 35 which makes safe the choice of $M = 100$



simulations in terms of complexity based on the number of functional evaluations to recover at a given accuracy these moments. As we are interested by small values of n (typically $n = 2$ or 3 in our applications), this latter can therefore be preferred. Anyway, both Monte Carlo trials and numerical integration lead to the introduction of weighted sums over a M-point sampling $\mathbf{I}_M$ of I as estimators of the previous moments.

The linearity in the sums permits to access the gradient of the moments with respect to the control parameters $\mathbf{x}$ from the gradient of the functional at the sampling point α_k. These are four vectors in $S_M = Span\{\nabla_{\mathbf{x}} j(\mathbf{x}, \alpha_k), \alpha_k \in \mathbf{I}_M\} \subset \mathbb{R}^N$. In applications of interest N is large. However, we showed that often $\dim(S_M) << N$ [31–33]. This analysis also permits to a posteriori give confidence bounds on the choice of the sampling size M which should be clearly larger than $\dim(S_M)$. Figure 1 shows an example of this analysis during the optimization of the shape of an aircraft with $N = 5000$ and $M = 100$ showing that $\dim(S_M)$ always remains below 35 making 100 a safe choice and clearly smaller than the dimension of the optimization space. This is interesting as indeed, without other information and considering vector spaces theory, the size of the sampling should be larger than the dimension of the control space (i.e., $M = N + 1$).

Let us denote by $C_{i=1,..,3}$ the three constraints on the second, third, and fourth moments and let us consider the subspace $s_M = Span\{\nabla_{\mathbf{x}} C_{i=1,...,3}\} \subset \mathbb{R}^3 \subset \mathbb{R}^N$. Obviously $p = dim(s_M) \leq 3$. Let us denote by $\{\mathbf{q}_{i=1,...,p}\}$ an orthonormal basis for s_M obtained, for instance, orthonormalizing the three gradient vectors by the Gram–Schmidt procedure. The gradients G of the constraints can therefore be expressed as linear combination of q_i: $G = (\nabla_{\mathbf{x}} C_{i=1,...,3}) = P^{-1}(\mathbf{q}_{i=1,...,p})$ with P being the matrix expressing the coordinates of $\mathbf{q}$ in G.

With equality constraints, a descent direction d can be obtained writing the first order optimality condition stating that d needs to be orthogonal to s_M. Hence, using the local orthonormal basis $\{\mathbf{q}_{i=1,...,p}\}$, we consider d given by:

$$d = \nabla_{\mathbf{x}}\mu - \sum_{i=1}^{p} < \nabla_{\mathbf{x}}\mu, \mathbf{q}_i > \mathbf{q}_i. \tag{3}$$

Denoting by Π the matrix of the projection operator $< \nabla_{\mathbf{x}}\mu, q >$ we have

$$d = \nabla_{\mathbf{x}}\mu - (\Pi PG)^t \ PG = \nabla_{\mathbf{x}}\mu - (G^t P^t \Pi P)^t \ G = \nabla_{\mathbf{x}}\mu + \Lambda^t G,$$

with $\Lambda^t = (\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3$. We have $d \rightarrow 0$ with the optimization iterations converging. Stationarity in d therefore realizes the first order optimality condition for the Lagrangian $L = J + \Lambda^t C$.

With inequality constraints, the solution of our minimization problem needs to verify the first order KKT conditions [41]. But, the optimality condition for the Lagrangian will involve only positive Lagrange multipliers: $\Lambda \in \mathbb{R}^3_+$ and $\nabla_{\mathbf{x}}L = \nabla_{\mathbf{x}}J + \Lambda^t \nabla_{\mathbf{x}}C = 0$ with the complementarity condition $\Lambda^t C = 0$ meaning that $\lambda_i = 0$ if $C_i \leq 0$ and $\lambda_i > 0$ if $C_i = 0$ (i.e., C_i is an active constraint). To define d we follow what put in place for the equality constraints, but only considering active constraints gradients in the definition of s_M which is not anymore a subspace but a convex cone:

$$s_M = \{\mathbf{x} \mid \mathbf{x} = \sum_{i=1}^{3} \beta_i \nabla_{\mathbf{x}} C_i, \beta_i > 0 \mid C_i = 0\} \subset \mathbb{R}^3 \subset \mathbb{R}^N \tag{4}$$

At the solution, $\nabla_{\mathbf{x}}J$ is orthogonal to this cone. Before working on the cone, let us start defining a local orthonormal basis $\{\tilde{\mathbf{q}}_{i=1,\ldots,p}\}$ for $\tilde{s}_M$ from (4) both with $\beta_i \in \mathbb{R}$. This is therefore a subspace and the basis can be defined as previously with $p = dim(s_M)$. Now, $\mathbf{q}_i = \pm\tilde{\mathbf{q}}_i$ and the sign chosen such that $< \mathbf{q}_{i=1,\ldots,p}, \nabla_{\mathbf{x}}C_j >\geq 0$, if $C_j = 0$ for $j = 1, \ldots, 3$ (i.e., pointing inside the cone). Here, $\{\mathbf{q}_{i=1,\ldots,p}\}$ are therefore the generators of the cone s_M deduced from a basis of $\tilde{s}_M$. If the generators cannot be defined, the problem is found having no solution as at least two of the constraints are incompatible with the gradients parallel and pointing in opposite directions. These generators permit to define the admissible search direction d from (3) but taking into account that we only remove the nonadmissible contribution:

$$d = \nabla_{\mathbf{x}}\mu - \sum_{i=1}^{p} \chi_i < \mathbf{q}_i, \nabla_{\mathbf{x}}\mu > \mathbf{q}_i, \tag{5}$$

with $\chi_i = 0$ if $< \mathbf{q}_i, \nabla_{\mathbf{x}}\mu >\geq 0$ and $\chi_i = 1$ if $< \mathbf{q}_i, \nabla_{\mathbf{x}}\mu >< 0$.

4 A Multi-Point Descent Algorithm

Our aim is to use existing platforms. Hence, to compute the ingredients above we use an available single-point optimization environment which can be easily modified for parallel multi-point calculations. This platform involves a direct simulation chain linking the parameters $(\mathbf{x}, \alpha)$ to the state $\mathbf{u}$ solution of a state equation $F(\mathbf{u}(q(\mathbf{x}, \alpha)) = 0$ and its adjoint $\mathbf{v}$ and to a functional j:

- Given $\mathbf{x}_0, 0 < \rho, \mathbf{I}_M, p_{max}$,
- Optimization iterations $p = 1, \ldots, p_{max}$
  - 1-M parallel state equation solutions $F(\mathbf{u}(q(\mathbf{x}_p), \alpha_k)) = 0,\ \alpha_k \in \mathbf{I}_M$,
  - 2-M parallel evaluations of $j(\mathbf{x}_p, \alpha_k),\ \alpha_k \in \mathbf{I}_M$,
  - 3-M parallel solutions of the adjoint state $\mathbf{v}$ equation:

$$\mathbf{v}^t F_{\mathbf{u}}(\mathbf{u}(q(\mathbf{x}_p), \alpha_k)) = j_{\mathbf{u}}^t,\ \alpha_k \in \mathbf{I}_M,$$

  - 4-M parallel evaluations of $\nabla_{\mathbf{x}} j(\mathbf{x}_p, \alpha_k) = j_{\mathbf{x}} + (\mathbf{v}^t F_{\mathbf{x}})^t,\ \alpha_k \in \mathbf{I}_M$,
  - 5-define d the descent direction using (5),
  - 6-minimization using d: (e.g., $\mathbf{x}_{p+1} = \mathbf{x}_p - \rho d$),
  - Stop if a given stopping criteria is achieved.

In multi-criteria problems steps 2, 3, and 4 include the treatment of more than one functional inducing a different definition of the descent direction d to account for other constraints (mainly physical this time) than the moment-based ones mentioned above.

Despite the natural presence of parallelism due to the M independent evaluations of the state, functional, and its gradient, computational complexity remains an issue. We have shown previously how to reduce this effort optimizing the sampling size [31] together with the use of incomplete sensitivity concept in the evaluation of the gradients which permits to avoid the solution of the M adjoint equations [40]. This is particularly suitable when using black-box state equation solvers not providing the adjoint of the state variables.

Such minimization problems have brought new interest to descent methods and this not only because of their lower computational complexity compared to gradient free methods [24, 41, 47]. Indeed, beyond minimization, we saw that gradients are useful to see what should actually be the search space in a context of robust multi-point design [32, 33]. Hence, beyond individual gradient accuracy (i.e., at each of the sampling point), what is important in multi-point problems is the global search space defined by the ensemble of the gradient vectors. This means that one might tolerate higher error levels in each of the gradient defined at the different sampling point than for a single-point optimization situation as what is important is for the global search space to remain nearly unchanged. An interesting mathematical concept which permits to measure the deviation between two subspaces is the principal angles between subspaces.

5 Angles Between Subspaces

We use the mathematical concept of "principal angles" between subspaces in the Euclidean spaces (here $\mathbb{R}^N$) initially introduced by Jordan [26]. If the maximum principle angle between the two subspaces is small, the two are nearly linearly dependent. Geometrically, this is the angle between two hyperplanes embedded in a higher dimensional space.

Let us briefly recall the concept of principal angles and how to practically compute them [25, 55].

For simplicity, suppose A and B are two subspaces of dimension k of $\mathbb{R}^N$, $N \geq 2k$, although this is not a prerequisite to define the principal angles. The k principal angles $\{\theta_i, i = 1, \ldots, k\}$ are recursively defined as:

$$\cos(\theta_i) = \frac{< a_i, b_i >}{\|a_i\| \|b_i\|} = \max\{\frac{< a, b >}{\|a\| \|b\|} : a \perp a_m, b \perp b_m; m = 1, \ldots, i-1\},$$

where $a_j \in A$ and $b_j \in B$.

The principal angles θ_i are between 0 and $\pi/2$. This is an important point and will be used later to take advantage of the positivity of the cosine of the angles. The procedure finds unit vectors $a_1 \in A$ and $b_1 \in B$ minimizing the angle θ_1 between them. It then takes the orthogonal complement of a_1 in A and b_1 in B and iterates. This procedure is not useful in practice as computationally inadequate. We would like to be able to find the angles θ_i from the inner products $< a_i, b_j >$ of the elements of two bases of A and B [49]. This would be interesting in our multi-point optimization context where we can exhibit an orthonormal basis of the global search space for the multi-point optimization problem using Gram–Schmidt orthonormalization.

Now, let $\{a_i, i = 1, \ldots, k\}$ and $\{b_i, i = 1, \ldots, k\}$ be two arbitrary orthonormal bases for A and B. Orthonormal bases are easy to obtain through the Gram–Schmidt orthonormalization procedure. Consider M being the matrix of the projection operator Pr_A of B onto A defined by:

$$Pr_{_A}(b_i) = \sum_{j=1}^{k} < b_i, a_j > a_j, \quad M = (< b_i, a_j >)_{i,j}.$$

The principal angles can be linked to this operator [49] through:

$$M = G\Sigma H^t,$$

where G and H are orthogonal matrices and $\Sigma = diag(\cos(\theta_i))$.

As G and H are orthogonal matrices, this is a singular vector decomposition of M. G and H are unknown at this point. But, we will show that we do not need them to get the θ_i. Otherwise, the approach will be again computationally useless.

We recall that the columns of G are the left-singular vectors of M and eigenvectors of MM^t and the columns of H are the right-singular vectors of M and eigenvectors of M^tM. And most important that $\cos^2(\theta_i)$ are the eigenvalues of $Pr_A^t Pr_A$ which writes in matrix form as: $M^tM = (G\Sigma H^t)^t(G\Sigma H^t) = H\Sigma^2 H^t$ with $\Sigma^2 = diag(\cos^2(\theta_i))$.

Therefore, to find the principal angles between subspaces A and B, knowing an orthonormal basis in each subspace, one should calculate M and find the eigenvalues of M^tM and take the square root of them. This last operation is valid as the angles are between 0 and $\pi/2$, and their cosine therefore always positive.

We presented the approach for subspaces of the same dimension k, but it is not necessary for the two subspaces to be of the same size in order to find the angles between them. We need $N \geq 2k$ to be able to exhibit two orthogonal subspaces. If $N < 2k$, some principal angles necessarily vanish and for $N = k$ they all vanish. This procedure is still valid if the subspaces have different dimensions. The projection operator can be defined as well as its transpose and the eigenvalues of M^tM are real as this is a symmetric square matrix.

In our optimization applications we always proceed first with a reduction in size of the search space using a sampling reduction size algorithm [31]. This makes the calculation of the whole set of eigenvalues feasible in terms of calculation complexity. However, if this is out of reach, one can evaluate the bounds on the angles to see the global pertinence of our reduced order models and gradient approximations. This can be done without an exact calculation of the all eigenvalues. It is sufficient to use the Gershgorin circle theorem to find these bounds as every eigenvalue of M^tM lies within at least one of the Gershgorin discs $D((M^tM)_{ii}, R_i)$ centered on $(M^tM)_{ii}$ and with radius $R_i = \sum_{j\neq i} |(M^tM)_{ij}|$. And because M^tM is symmetric, the eigenvalues being real, we only consider the intersection of the discs with the x-axis. Alternatively, the largest and smallest principal angles can be found using iterative power and inverse power methods applied to M^tM.

One should, however, be aware that these bounds might not be sufficiently sharp to discriminate between two reduced order models and decide, for instance, which one is more adequate for sensitivity analysis. Figure 2 shows a typical sketch. It represents principal angles calculated between a first subspace generated by the exact gradients of a transport model for a ten points sampling of one of the functioning parameters of the model and two subspaces generated by the sensitivities derived from two approximations of this model for the same sampling. Details of the models can be found in [34]. But the modeling problem is not of a main concern here. What is important is that if one only considers the first and last principal angles, model M_2 is found being a better approximation to be used in a linearization procedure. However, with the whole spectrum in hand the picture is quite different and M_1 appears to be more suitable if one intends to use this reduced order model for sensitivity analysis.

Principal angles between multi-point search spaces are interesting to measure the pertinence of sensitivity definitions based on different models or numerics. Indeed, the design will be unaffected by a reduction in the model's complexity if the search

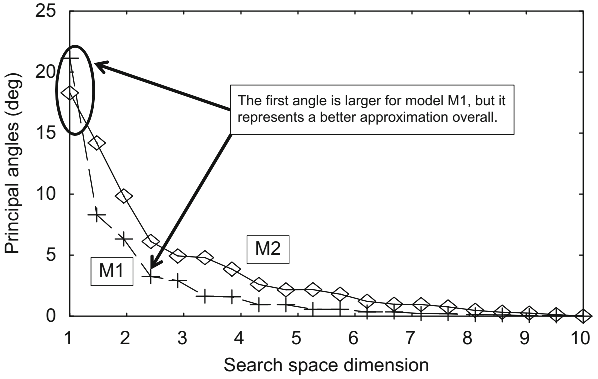


Fig. 2 Principal angles between the subspaces generated by an exact gradient calculation and the linearization of two reduced order models. This permits to quantify the pertinence of an approximation from the whole spectrum. Model M_1 is found to be a better approximation even with a first principal angle slightly larger than with M_2

subspaces, generated by the gradients at the sampling points of the functioning parameter interval and their approximations, remain the same. This is therefore an original quantification tool for epistemic uncertainties.

6 Inversion for Incertain Data

Let us expand the class of problem introduced in Sect. 2 to the following situation:

$$\min_{\mathbf{x}\in\mathbf{O}_{ad}} j(\mathbf{x},\alpha,\mathbf{u}^*), \quad \mathbf{u}^* \in \mathbb{R}^p, \alpha \in \mathbf{I} \subset \mathbb{R}^n, \mathbf{O}_{ad} \subset \mathbb{R}^N. \tag{6}$$

$\mathbf{u}^*$ represents either measurements or state estimations. It is a vector of random variables. We are interested in functionals j of the form:

$$j(\mathbf{x},\alpha,\mathbf{u}^*) = \tilde{j}(\mathbf{x},\alpha) + \frac{1}{2}\|\Pi\mathbf{u}(\mathbf{x},\alpha) - \mathbf{u}^*(\alpha)\|^2. \tag{7}$$

The first term is what has been discussed up to now. Operator $\Pi : \mathbb{R}^N \to R^p$ (typically a linear interpolation operator) makes the state available at data locations. Inverse problems are in this class [44, 52]. This formulation also permits to see the state as uncertain as a whole with Π the identity operator. One can also introduce zoning techniques (as shown in Fig. 6) to discriminate through the level of confidence one might have on the state evaluation following the variability one observes in practice (experimental or in flight). It is indeed well known that the flow distribution is quite stable in the cockpit and over the first and business class siting area where the flow is nearly potential. On the other hand, flow variability increases spanwise (easy to see from wings tips motions) and also toward the tail of the aircraft (flying coach once makes this easy to understand). These are due, among others, to separated turbulent flows instabilities and fluid–structure interactions which are more difficult to predict and the state is therefore more "uncertain" there.

To summarize, we assume the components of $\mathbf{u}^*$ independent, uncertain, and given by their Gaussian PDF, for instance, $\mathscr{N}(\mu_i, \sigma_i^2),\quad i = 1, \ldots, p$ with mean μ_i and variance of σ_i^2. $Cov_{\mathbf{u}^*}$ is therefore a diagonal matrix.

The simplest way to measure the effect of these uncertainties on the inversion result is again to proceed with Monte Carlo simulations. This implies a sampling of the variation domain of the data consistent with their PDF. This means we proceed with M independent inversions for M data sets defined by independent choices compatible with the PDF of $\mathbf{u}^*$ given by:

$$\mathscr{N}(\mu_i, \sigma_i^2) \rightarrow (\mathbf{u}_i^*)^m, \quad i = 1, \ldots, p, \quad m = 1, \ldots, M.$$

These independent inversions will produce M optimal control parameters $\mathbf{x}_{opt}^m,\ m = 1, \ldots, M$ from which statistical moments can be defined (typically the mean and variance) with again a rate of convergence in $M^{-1/2}$ independent of p. Such generation of scenarios is already very demanding when involving only a direct simulation chain. In our problem, each of the scenarios involves an inversion, each requiring several solutions of the direct and adjoint problems. This complexity makes that this approach is clearly out of the table even if the calculations are independent and can be carried out in parallel.

6.1 Low-Complexity Uncertainty Evaluation

In the sequel, we discuss two low-complexity constructions of $Cov_{\mathbf{x}}$ the covariance matrix of the control parameters from $Cov_{\mathbf{u}^*}$ the covariance matrix of the data. We want these constructions to have a cost comparable to a deterministic inversion and, again, we want to avoid any sampling of a large dimension space.

7 a-Quantile

Consider a random variable v with its PDF known (either analytic or tabulated). The tail of the PDF can be characterized defining for a given probability level $(0 < a < 1)$ the following threshold value:

$$\mathrm{VaR}_a = \inf\{l \in \mathbb{R} : P(v > l) \leq 1 - a\}.$$

Different a-quantile are available. One very well known is the value at risk which has been widely used in financial engineering as a measure of risk of loss on a given asset [27]. We do not need the time dependency issue here but it is interesting as it permits to account for possible improvement of measurement accuracy as discussed in [34].

7.1 Bounding the Uncertainty Domain

We would like to use the concept of a-quantile (we call in the sequel VaR) to define a closed domain of variation for the uncertain data [34]. Given a threshold $0 \le a < 1$, a data $\mathbf{u}_i^*, i = 1, \ldots, p$ belongs to the interval $[\mu_i + \mathrm{VaR}_a^-, \mu_i + \mathrm{VaR}_a^+]$ with $\mathrm{VaR}_a^- \le 0 \le \mathrm{VaR}_a^+$ with probability a. As we consider Gaussian PDFs we have $\mathrm{VaR}_a^- = -\mathrm{VaR}_a^+$ and the values at risk are explicitly known:

$$\mathrm{VaR}_{0.99}(N(0,1)) = 2.33, \quad \text{and} \quad \mathrm{VaR}_{0.95}(N(0,1)) = 1.65,$$

and $\mathrm{VaR}_a(N(0,\sigma^2)) = \sigma^2 \mathrm{VaR}_a(N(0,1))$. We have therefore, with probability a, an uncertainty domain for the data given by:

$$B_a(\mu) = \Pi_{i=1}^p [\mu_i - 1.65\sigma_i^2, \mu_i + 1.65\sigma_i^2] \subset \mathbb{R}^p$$

This is a large domain and we do not want to proceed with any sampling.

7.2 Directional Extreme Scenarios

However, using the sensitivity of the functional with respect to the data we can identify two directional extreme sets of data corresponding to the intersection of $B_a(\mu)$ and $d = \mu + t\ \partial j / \partial u^*, \quad t \in \mathbb{R}$. Let us call these two data sets $(\mathbf{u}^*)^{\pm}$ defined by:

$$(\mathbf{u}^*)^{\pm} = \mu \pm 1.65\ \sigma_i^2 \left(\frac{\partial j / \partial u^*}{\|\partial j / \partial \mathbf{u}^*\|} \right)_i . \tag{8}$$

To measure the impact of this variability on the result of the inversion, we proceed with two minimizations with the target data given by $(\mathbf{u}^*)^{\pm}$ starting from $\mathbf{x}^* = \mathbf{x}_{opt}(\mathbf{u}^* = \mu)$. Let us call $(\mathbf{x}^*)^{\pm}$ the results of these inversions.

We assume monotonic behavior of the outcome of the inversion with respect to the data:

$$\|\mathbf{x}^*(\mu) - \mathbf{x}^*(\nu)\| \nearrow \quad \text{if} \quad \|\mu - \nu\| \nearrow . \tag{9}$$

This assumption is reasonable and means that larger deviations in data sets bring larger variations in the outcome of the optimization. This also suggests that the maximum deviation for the results of the inversion due to the uncertainty on the data can be estimated through: $X^{\pm} = (\mathbf{x}^{*})^{+} - (\mathbf{x}^{*})^{-}$. Hence, we introduce a first approximation to the covariance matrix $Cov_{\mathbf{x}^{\pm}}$ [56] for $\mathbf{x}$:

$$Cov_{\mathbf{x}^{\pm}} = \mathbb{E}((X^{\pm})(X^{\pm})^{t}) - \mathbb{E}(X^{\pm})\mathbb{E}(X^{\pm})^{t} \sim (X^{\pm})(X^{\pm})^{t} - (\overline{X^{\pm}})(\overline{X^{\pm}})^{t}, \tag{10}$$

with $\overline{X^{\pm}} = ((\mathbf{x}^{*})^{+} + (\mathbf{x}^{*})^{-} - 2\mathbf{x}^{*})/2$.

The monotonicity hypothesis can be a posteriori checked, at least partially, measuring the distance between $(\mathbf{x}^{*})^{\pm}$ and $\mathbf{x}^{*} \pm \rho\nabla_{\mathbf{x}} j(\mathbf{x}^{*}, \mathbf{u}^{*}), \rho > 0$. This expression permits to identify two bounds $\rho^{\pm}$ and two intervals $[0, \rho^{\pm}]$ on which the monotonicity is verified. Larger values of parameters $\rho^{\pm}$ a posteriori enforce the hypothesis.

If one looks at optimization from the view point of controllability for dynamical systems [40, 46], quantiles can be introduced in optimization algorithms [34]. The notion of over-solving appears then naturally where it becomes useless to solve accurately near an optimum when the variations in control parameters between two iterations of the optimizer fall below the uncertainties defined through a local uncertainty ball: all the points inside this ball being indeed equivalent in term of the confidence one can have on their performance.

We have presented the concept of DES in [32, 33] with applications to robust shape optimization in aeronautics, atmospheric dispersion and also to quantify the sensitivity of littoral erosion to uncertainties in bottom sand characteristics [38]. DES can be defined for $\mathbf{x}$ as well, considering the components of the design vector as random variables. It is indeed interesting to account for uncertainties in large dimensional spaces. We have also extended the DES considering ensemble-based simulations after casting the multi-point optimization algorithm into the EnKF formalism (see [35] for the details). The joint application of the EnKF and DES leads to the concept of EDES which provides more exhaustive possible extreme scenarios knowing the PDF of our optimization parameters. A sketch of these constructions is shown in Fig. 3.

8 From the Adjoint to the Covariance Matrix of the Optimization Parameters

Another construction of $Cov_{\mathbf{x}}$ takes advantage of our adjoint calculation leading to $\nabla_{\mathbf{x}} j$ the gradient of the functional with respect to the optimization parameters [37].

Let us recall the adjoint formulation for a generic state equation $F(\mathbf{u}(\mathbf{x}, \alpha)) = 0$. The gradient of j with respect to $\mathbf{x}$ writes

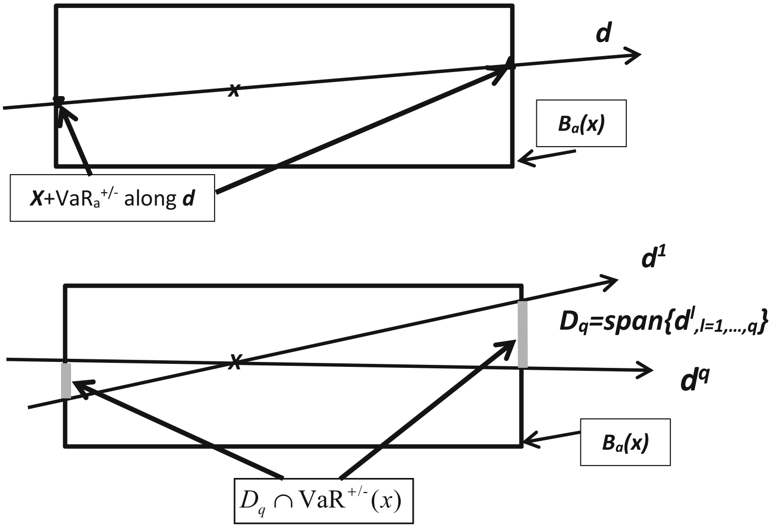


Fig. 3 Sketch of directional extreme scenarios (DES) given by $\mathbf{x}^{\pm} = d \cap \partial \mathbf{B}_a(\mathbf{x})$ and ensemble directional extreme scenarios (EDES) $D_q \cap \partial \mathbf{B}_a(\overline{\mathbf{x}})$ for an ensemble of size q ($\overline{\mathbf{x}}$ being the ensemble mean). The *grey zone* is not necessary connected

$$\nabla_{\mathbf{x}} j = \frac{\partial j}{\partial \mathbf{x}} + \left((\frac{\partial j}{\partial \mathbf{u}})^t \ \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \right)^t = \frac{\partial j}{\partial \mathbf{x}} + \left((\frac{\partial j}{\partial \mathbf{u}})^t \ (\frac{\partial F}{\partial \mathbf{u}})^{-1} \ \frac{\partial F}{\partial \mathbf{x}} \right)^t = \frac{\partial j}{\partial \mathbf{x}} + \left(\mathbf{v}^t \ \frac{\partial F}{\partial \mathbf{x}} \right)^t,$$

where we have introduced the adjoint variable $\mathbf{v}$ solution of:

$$\mathbf{v}^t \frac{\partial F}{\partial u} = (\frac{\partial j}{\partial u})^t, \tag{11}$$

and used in algorithm of Sect. 4. In cases the governing equations are self-adjoint (i.e., $\frac{\partial F}{\partial \mathbf{u}} = (\frac{\partial F}{\partial \mathbf{u}})^t$), one can use the corresponding solver with $\frac{\partial j}{\partial \mathbf{u}}$ as the right-hand side and simply solve:

$$\frac{\partial F}{\partial \mathbf{u}} v = \frac{\partial j}{\partial \mathbf{u}}.$$

Also, if F is linear, $\frac{\partial F}{\partial \mathbf{u}}$ is a constant operator independent of $\mathbf{u}$. The interest of the adjoint formulation is that the cost of getting $\nabla_{\mathbf{x}} j$ becomes independent of the size of $\mathbf{x}$. But, the problem with the adjoint approach is that, except for the two situations we mentioned (linear or self-adjoint state equations), it needs the development (and maintenance) of a new code. This is why we use automatic differentiation when possible.

In multi-criteria problems like the one shown in Sect. 9, where the functional j is minimized under equality or inequality constraints $c_{i=1,\dots,q}$, we need to solve an adjoint problem for the functional and each of the active constraints (needed to express the first order KKT conditions). This can be seen as a block diagonal matrix inversion with all blocks similar and the right-hand side given by $(\partial_{\mathbf{u}}j, \partial_{\mathbf{u}}c_1, \ldots, \partial_{\mathbf{u}}c_q)^t$ if we have q active constraints. Automatic differentiation in reverse mode with multiple right-hand side capacity can be used to address this problem. Otherwise, deflation techniques for linear systems with multiple right-hand sides can be applied [9, 11] taking advantage of the fact that the blocks being the same the Krylov decomposition needs to be conducted only once.

j involves the least square deviation at data location between model and data. $\partial_{\mathbf{u}}j$ in the right-hand side of (11) can be obtained writing:

$$j(\mathbf{x},\mathbf{u}^*) = \tilde{j} + \frac{1}{2}\|\Pi\mathbf{u}-\mathbf{u}^*\|^2 = \tilde{j} + \frac{1}{2} < \Pi\mathbf{u}-\mathbf{u}^*, \Pi\mathbf{u}-\mathbf{u}^* >$$

$$= \tilde{j} + \frac{1}{2} < \Pi^t\Pi\mathbf{u},\mathbf{u} > - < \Pi^t\mathbf{u}^*,\mathbf{u} > + \frac{1}{2} < \mathbf{u}^*,\mathbf{u}^* >,$$

and we have $\partial_{\mathbf{u}}j = \partial_{\mathbf{u}}\tilde{j} + \Pi^t\Pi\mathbf{u} - \Pi^t\mathbf{u}^*$. On the other hand, the sensitivity of j with respect to the data $\partial_{\mathbf{u}^*}j$ needed in (8) is given by $\partial j/\partial\mathbf{u}^* = -(\Pi\mathbf{u}-\mathbf{u}^*)$.

With $\nabla_{\mathbf{x}}j$ in hand, let us establish another expression for the covariance matrix of $\mathbf{x}$ considered as a vector of zero-mean random variables. Denote, for simplicity, by $\mathbf{u}$ the model solution (zero-mean valued: $\mathbf{u} \leftarrow \mathbf{u}-\mu$) at data location and suppose it is linked to the parameters through a linear model: $\mathbf{u} = L\mathbf{x}$. The covariance matrix for $\mathbf{u}$ is therefore:

$$Cov_{\mathbf{u}} = \mathbb{E}(\mathbf{u}\mathbf{u}^t) = \mathbb{E}(L\ \mathbf{x}\mathbf{x}^t\ L^t) = L\ \mathbb{E}(\mathbf{x}\mathbf{x}^t)\ L^t = L\ Cov_{\mathbf{x}}\ L^t.$$

If the dependency of $\mathbf{u}$ with respect to the parameter $\mathbf{x}$ is nonlinear the analysis holds for the linearized model. Introducing $\mathscr{J} = \partial\mathbf{u}/\partial\mathbf{x}$ we have

$$Cov_{\mathbf{u}} = \mathscr{J}\ Cov_{\mathbf{x}}\ \mathscr{J}^t.$$

To get $Cov_{\mathbf{x}}$ we need therefore to invert this expression and because the amount of data can be large and probably impossible to exactly fit, we proceed with a least square formulation looking for $Cov_{\mathbf{x}}$ minimizing:

$$\frac{1}{2} < \mathscr{J}\ Cov_{\mathbf{x}}\ \mathscr{J}^t, \mathscr{J}\ Cov_{\mathbf{x}}\ \mathscr{J}^t > - < Cov_{\mathbf{u}}, \mathscr{J}\ Cov_{\mathbf{x}}\ \mathscr{J}^t > .$$

First order optimality condition with respect to $Cov_{\mathbf{x}}$ gives

$$\mathscr{J}^t\mathscr{J}\ Cov_{\mathbf{x}}\ \mathscr{J}^t\mathscr{J} - \mathscr{J}^t\ Cov_{\mathbf{u}}\ \mathscr{J} = 0,$$

which leads to

$$Cov_{\mathbf{x}} = (\mathscr{J}^t\mathscr{J})^{-1}\ \mathscr{J}^t\ Cov_{\mathbf{u}}\ \mathscr{J}\ (\mathscr{J}^t\mathscr{J})^{-1},$$

and eventually, to

$$Cov_{\mathbf{x}} = \mathscr{J}^{-1}\ Cov_{\mathbf{u}}\ \mathscr{J}^{-t} = \left(\mathscr{J}^{t}\ Cov_{\mathbf{u}}^{-1}\ \mathscr{J}\right)^{-1}. \tag{12}$$

To get $Cov_{\mathbf{x}}$ and knowing $Cov_{\mathbf{u}}$, it is therefore sufficient to evaluate $\mathscr{J} = \partial\mathbf{u}/\partial\mathbf{x}$. The second expression in (12) is interesting as it involves the inversion of a square matrix and gives a least square sense to the inversion of rectangular matrices. Also, if the optimization is successful and model u and data u^* close, we can use the fact that data are usually independent and use the covariance matrix of the observation instead of $Cov_{\mathbf{u}}$:

$$Cov_{\mathbf{u}} \sim Cov_{\mathbf{u}^*},$$

which is then diagonal and its inversion straightforward.

The question is, therefore, how to efficiently evaluate $\mathscr{J} = \partial\mathbf{u}/\partial\mathbf{x}$. The model at data locations $\Pi\mathbf{u}$ is obtained applying, for instance, a linear interpolation operator Π to the model solution $\mathbf{u}$ on the mesh. Therefore, we have

$$\mathscr{J} = \Pi\frac{\partial\mathbf{u}}{\partial\mathbf{x}}.$$

Now recall that $\nabla_{\mathbf{x}}j$ is available and has been computed with an adjoint approach. We now use it to access $\partial\mathbf{u}/\partial\mathbf{x}$ without extra calculation:

$$\nabla_{\mathbf{x}}j = \frac{\partial j}{\partial\mathbf{x}} + \left((\frac{\partial j}{\partial\mathbf{u}})^t\ \frac{\partial\mathbf{u}}{\partial\mathbf{x}}\right)^t = \frac{\partial j}{\partial\mathbf{x}} + \left((\frac{\partial j}{\partial\mathbf{u}})^t\ \Pi^{-1}\mathscr{J}\right)^t,$$

The first terms in the right-hand side is zero if there is no direct dependency on $\mathbf{x}$ in j. It is non-zero, for instance, if a Tikhonov regularization term is introduced in the functional [52]. This leads to:

$$(\frac{\partial j}{\partial\mathbf{u}})^t\ \Pi^{-1}\mathscr{J} = (\nabla_{\mathbf{x}}j - \frac{\partial j}{\partial\mathbf{x}})^t,$$

and eventually to,

$$\mathscr{J} = \Pi\ (\frac{\partial j}{\partial\mathbf{u}})^{-t}\ (\nabla_{\mathbf{x}}j - \frac{\partial j}{\partial\mathbf{x}})^t. \tag{13}$$

the components of $(\partial j/\partial\mathbf{u})^{-t}$ which is a line vector are given by the inverse of those of $(\partial j/\partial\mathbf{u})$ and scaled by the inverse of its size in order to have $(\partial j/\partial\mathbf{u})^{-t}.(\partial j/\partial\mathbf{u}) = 1$. Alternatively, to avoid numerical difficulties with small components of $(\partial j/\partial\mathbf{u})$, (13) can again be seen in a least square sense with the inverse of a normal matrix involved:

$$\mathscr{J} = \Pi \ \left((\frac{\partial j}{\partial \mathbf{u}})(\frac{\partial j}{\partial \mathbf{u}})^t \right)^{-1} \ \frac{\partial j}{\partial \mathbf{u}} \ (\nabla_{\mathbf{x}} j - \frac{\partial j}{\partial \mathbf{x}})^t. \tag{14}$$

This expression involves the inverse of the information matrix $((\partial j/\partial \mathbf{u})(\partial j/\partial \mathbf{u})^t)$. One should be aware that the numerical condition of this matrix can be very poor. We do not discuss this issue here but typically the Bunch and Kaufman [2] algorithm should be used in order to account for this possible deficiency. In particular, if rank deficiency is detected the Moore–Penrose inverse should be used based on the eigenvalue decomposition of the information matrix [8].

Under the hypothesis of the validity of the physical model, this analysis gives indications on the level of backward sensitivity of the optimization parameters with respect to the model solution at data locations which is also the sensitivity with respect to the deviation between the model and data at the data locations (as the data are independent of the optimization parameters):

$$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{\partial (\mathbf{u} - \mathbf{u}^*)}{\partial \mathbf{x}}.$$

9 Applications to Robust Aircraft Shape Design

These ingredients have been applied to several aircraft shape designs in cruise conditions [31–33, 37]. Many sources of variability exist in these problems, for instance, due to a change in the weight of the aircraft during the flight because of fuel consumption or due to variability in the flight conditions. Two parameters should be particularly given consideration as our α: the Mach number and the sideslip incidence angle. The sideslip angle is important to account for situations where the aircraft cruises against transverse winds which are very common. A non-zero sideslip angle induces fully 3D effects on the flow around the plane making necessary the consideration of a full aircraft during the design. Usually aircraft are designed for a range of angle of incidence. But, these designs are usually realized with the sideslip angle set to zero. It is therefore necessary to reduce the sensitivity of the design with respect to this parameter. However, because the airplane geometry is symmetric spanwise, it is not necessary to consider a symmetric range for the transverse wind but we need to consider the whole aircraft as there is no spanwise symmetry in the flow for non-zero sideslip angles.

9.1 Single-Point Shape Optimization Platform

We work in the framework of an existing shape optimization platform. We use, in particular, several of its simulation codes for the shape parameterization and deformation, for the mesh deformation, for the flow calculations around the aircraft, and for the shape adjoint sensitivity analysis of aerodynamic coefficients. This is a very standard and generic situation and one shall consider these as black-boxes.

Let us briefly recall our direct dependency chain linking independent variables $(\alpha, \mathbf{x})$ to the dependent variables $(q(\mathbf{x}), U(\alpha, \mathbf{x}))$ describing geometrical entities and state variables and to the cost function (here the drag coefficient C_d) and to the constraints $c_{i=1,\ldots,4}$:

$$(\alpha, \mathbf{x}) \to (q(\mathbf{x}), U(\alpha, q(\mathbf{x}))) \to (C_d, c_{i=1,\ldots,4})(\alpha, \mathbf{x}, q(\mathbf{x}), U(\alpha, q(\mathbf{x}))). \tag{15}$$

It is important to identify all dependencies in order for the sensitivity analysis to be completed, especially when the operating conditions are not anymore single valued. The functional and constraints will be described in Sect. 9.1.3.

9.1.1 Shape Parameterization and Geometrical Entities

In (15) $\mathbf{x}$ denotes a CAD-free parameterization [39, 40] which does not require a priori local regularity assumptions on the shape as it is implicitly the case in computer aided design (CAD) based shape definitions. More precisely, $\mathbf{x}$ represents shape deformations along the normal to the triangular faces of the surface mesh as shown in Fig. 4. For the problem discussed here this search space has a dimension N of several thousands. This parameterization receives different denominations and belongs to the same class as node-based or free-form shape definitions. In all these approaches the regularity of the deformation needs to be monitored [40, 57]. This parameterization is intermediate in term of generality between CAD definitions of shapes and fully free topological optimization choices where both the regularity and topology of the shape are free. Examples of shape deformation produced by our optimization procedure for different regularity requirements are shown in Fig. 5. Need for regularity control comes from the fact that, unlike with a CAD definition, the shape $\partial\Omega$ of an object Ω and a gradient-based deformation of $\partial\Omega$ do not belong to the same function space in terms of regularity and, actually, the second is always less regular [37, 39, 40].

This can be illustrated on a simple example with $J(\mathbf{x}) = \|A\mathbf{x} - b\|^2$ taking $\mathbf{x} \in H^1(\partial\Omega)$, $A\mathbf{x}$ and b in $L^2(\partial\Omega)$. The gradient $J'_{\mathbf{x}} = 2A^T(A\mathbf{x}-b)$ belongs to $H^{-1}(\partial\Omega)$. Therefore, any variation along $J'_{\mathbf{x}}$ will have less regularity than $\mathbf{x}$: $\delta\mathbf{x} = -\rho J'_{\mathbf{x}} = -\rho(2(A\mathbf{x} - b)A) \in H^{-1}(\partial\Omega)$. We therefore need to project (or filter or smooth) into $H^1(\partial\Omega)$. Now, suppose the shape is described in a finite dimensional parameter space, as, for instance, with a polynomial definition of a surface (this is like a CAD parameterization). When we consider as control parameters the coefficients of the polynomial, changes in those do not change the regularity as the new shape will always belong to the same polynomial space. Sobolev inclusions give the key for the choice of the regularity operator with the CAD-free parameterization [40]. In our case, because we are using a piecewise linear discretization, a second order elliptic system with a local definition of the viscosity is sufficient.

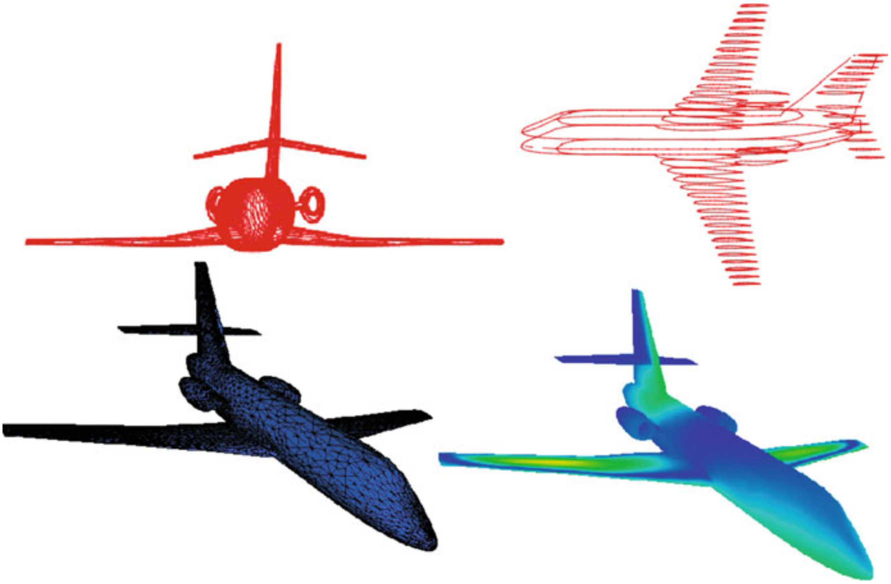


Fig. 4 CAD-free shape parameterization (*lower-left*) and by-section definitions (*upper*) of the shape for geometric constraints enforcement. *Lower-right* is one single $\nabla_{\mathbf{x}} C_d - < \nabla_{\mathbf{x}} C_d, \pi > \pi$ described in Sect. 9.1.3 for this CAD-free parameterization

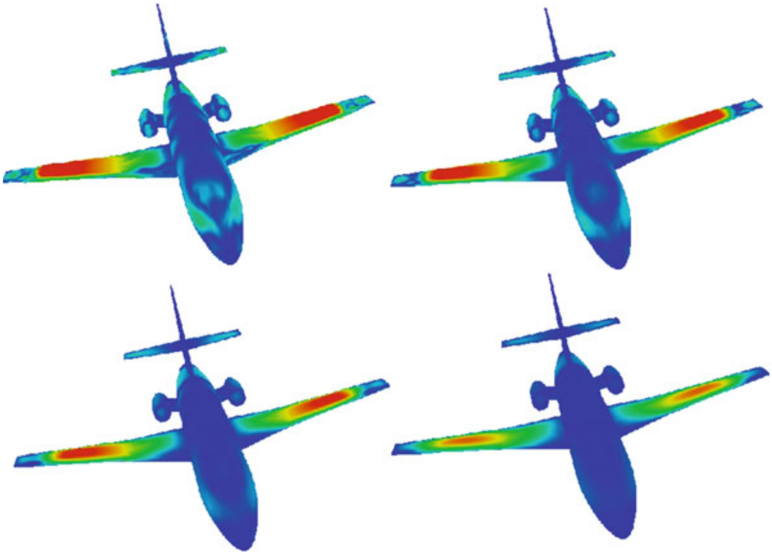


Fig. 5 Regularity control in CAD-free shape parameterization: examples of shape deformation produced by our optimization procedure for different regularity requirements

This capacity to monitor the regularity of the shape is also interesting as often the optimal solution is not reachable by the current CAD parameterization of the shape. Hence, after an optimization with the CAD-free parameterization and using different level admissible regularity for the shape, one can decide which realization is more suitable and also whether it is interesting or not to enrich the current CAD definition of the shape.

$q(\mathbf{x})$ denotes the auxiliary unstructured mesh related geometrical quantities (surfaces, volumes, normals, etc.). When the shape is modified, this change must be propagated through the mesh keeping it admissible and we need to recalculate all related geometrical quantities. Admissible and positive mesh deformation is achieved by a 3D torsional spring analogy method [14].

9.1.2 Flow Solver

In (15) $U(\alpha, q(\mathbf{x})) = (\rho, \rho\mathbf{u}, \rho E)^t$ denotes the flow variables in conservation form solution of the Euler equations where, T being the temperature, the total energy is given by $E = C_v T + \|\mathbf{u}\|^2/2$ and the pressure by the state law $p = \rho RT$ with R the perfect gas constant.

The details of the implementation of the flow solver are available in [40]. It is based on a finite volume Galerkin method on unstructured tetrahedral meshes [10]. Of course, other choices are possible for the flow solver and the literature on numerical methods for compressible flows is huge. This is not central to our discussion. We target steady solutions and use time marching with local time steps to reach these. The time integration procedure is explicit and is based on a low-storage Runge–Kutta scheme. To improve computational efficiency we only use partial convergence for the state equations. In particular, the sufficient level of convergence retained is when the flow solver iterations only modify the third digits in the aerodynamic coefficients. This is achieved with about 100 RK iterations for this inviscid configurations starting from a uniform solution distribution. During optimization a new calculation for a new shape is always started from the previously available solution making us to proceed with typically only 20 RK new iterations [32, 33, 37].

Often in practice the mesh used for such optimization problems is insufficiently fine. It is, however, important that the approach uses the ingredients of a generic high-fidelity platform and does not remove or simplify any of its ingredients as often it is the case in uncertainty quantification procedures using reduced order models. We should rather consider that in practice our modeling capability and our computational resources will always be limited. The backward uncertainty propagation procedures presented in Sect. 6 permit to quantify the impact of this lack of resolution on the design as shown in Fig. 6.

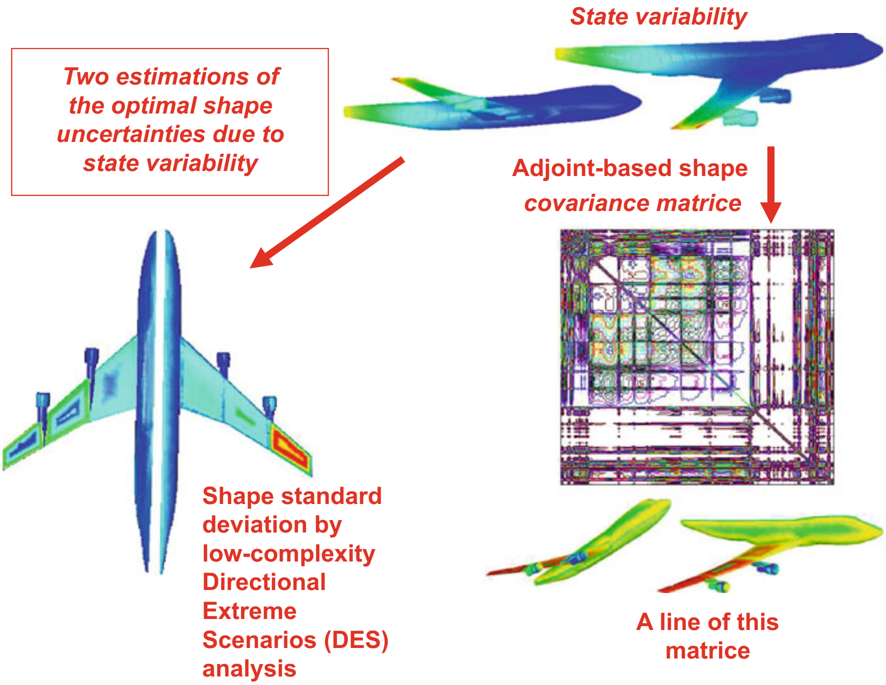

State variability
Two estimations of the optimal shape uncertainties due to state variability
Adjoint-based shape covariance matrice
Shape standard deviation by low-complexity Directional Extreme Scenarios (DES) analysis
A line of this matrice

Fig. 6 Two approaches to build $diag(Cov_{\mathbf{x}})$ from $diag(Cov_{\mathbf{u}^*})$ indicating the variability of $\mathbf{u}^*$ over the shape. The covariance distribution over the shape of the aircraft shows that for the design to be robust in variable flight conditions the engines pylon, fairing, and air intakes should have different shapes following their position on the wing

9.1.3 Optimization Problem

We consider a classical aerodynamic problem where two main quantities of interest are the drag C_d and lift C_l coefficients:

$$C_d(\mathbf{x}) = \frac{1}{2\rho_\infty \|\mathbf{u}_\infty\|^2} \int_{shape(\mathbf{x})} p(q(\mathbf{x}))(\mathbf{u}_\infty.\mathbf{n}(q(\mathbf{x}))d\gamma, \tag{16}$$

where superscript ∞ indicates inflow conditions. The lift coefficient is evaluated with formula (16) where u_∞ is replaced by $\mathbf{u}_\infty^\perp$ in the boundary integral. Aircraft performance analysis concerns its payload and range. These are directly linked to the aerodynamic coefficients of the aircraft called the lift (conditioning the payload) and drag (conditioning the fuel consumption) coefficients. The lift coefficient often appears through an inequality $C_l - C_l^{target} \geq 0$ or equality constraint $c_1 = |C_l^{target} - C_l(p(q(\mathbf{x}))|$ with C_l^{target} a target performance.

Structural efficiency and necessity of useful free volume also implies the consideration of geometric criteria such as a constraint on the volume V of the aircraft or its by-section definition. As for the lift coefficient, this gives a constraint of the form $c_2 = |V^{target} - V(q(\mathbf{x}))|$. The volume of an object Ω (here the aircraft) is expressed through the boundary integral formula: $V = \int_\Omega 1 = \int_\Omega \frac{1}{3}\nabla.(\mathbf{X}) = \int_{\partial\Omega} \mathbf{X}.\mathbf{n}$, where $\mathbf{X} = (x_1, x_2, x_3)^t$ is the local coordinate over the shape.

The last geometric constraint concerns the local wing by-section thickness which is prescribed. We define by-section definitions of the shape where the number of sections n_s is free and can be adapted to account for the complexity of the geometry. Each node in the parameterization is associated with a section Σ_i, and for each section, we define the maximum thickness Δ_i. This last operation requires the projection of the upper-surface nodes over the lower surface for each section. This constraint is expressed as: $c_3 = \sum_{i=1}^{n_s} |\Delta_i(q(\mathbf{x})) - \Delta_i^{target}|$.

An alternative solution which is much simpler to implement is to only enforce a local volume constraint in each section Σ_i using the volume formula above: $V(\Sigma) = \int_{\Sigma_i} 1 = \int_\Omega 1\ \chi_{\Sigma_i} = \int_\Omega \frac{1}{3}\nabla.(\mathbf{X})\ \chi_{\Sigma_i} = \int_{\partial\Omega} \mathbf{X}.\mathbf{n}\ \chi_{\partial\Sigma_i}$, where χ is an indicator function ($\chi = 1$ if the point is in section Σ_i and $\chi = 0$ otherwise). Testing if a point is in Σ_i is easy and only requires an interval-based coordinate check, spanwise in this situation.

Finally, a fourth term concerns the data assimilation criteria for the pressure over the shape as introduced in Sect. 6: $c_4 = \frac{1}{2}\|\Pi p(\mathbf{x}) - p^*\|^2$, p^* is a vector of random variables for the pressure values on the shape and can be used to account for the impact on the design of the uncertainty on the pressure estimation by the Euler model.

During optimization, the constraints can be accounted for by introducing a penalty term in the cost function: $j = C_d + \sum_{i=1,4} a_i c_i,\quad a_i \in \mathbb{R}^+$. But this should be avoided when possible. We use it, however, for the definition of the DES [37].

One classical technique to recover the lift during optimization is to change the flow incidence taking advantage of the linear relationship between the incidence and lift away from stall conditions. Suppose, however, that we do not want to use either penalty or such approximations. An alternative would be to follow what presented in Sect. 3 and consider a locally admissible gradient orthogonal to $\mathbf{S} = Span(\nabla_\mathbf{x} c_i, i = 1, \ldots, 4)$ with $\dim(\mathbf{S}) \leq 4$. Let us denote by π an orthonormal basis of this subspace obtained by the Gram–Schmidt procedure applied to the gradients of the constraints. The admissible gradient is given by:

$$\delta_k = \delta(\mathbf{x}, \alpha_k) = \nabla_\mathbf{x} C_d - < \nabla_\mathbf{x} C_d, \pi > \pi, \tag{17}$$

where $<, >$ indicates the scalar product over subspace $\mathbf{S}$. This is therefore similar to the construction given in (3) where $\pi = \{\mathbf{q}_{i=1,\ldots,3}\}$ and with the constraints C_i replaced by c_i. In the presence of inequality constraints $c_i \leq 0$ instead of equality we build the admissible direction based on the KKT optimality conditions following (5). Once δ_k obtained, the developments of Sect. 3 are followed with the gradient $\nabla_\mathbf{x} j$ replaced by the search direction δ_k.

To complete the picture we need to provide $\nabla_\mathbf{x} C_d$, $\nabla_\mathbf{x} C_l$, $\nabla_\mathbf{x}\|\Pi p(\mathbf{x}) - p^*\|^2$, $\nabla_\mathbf{x} V$, and $\nabla_\mathbf{x} \Delta_i$. The three former require the adjoint of the state equation and we take advantage of the capability for multi-right-hand side adjoint calculation of tapenade in reverse mode to access these gradients without the solution of three separate adjoint problems. Our direct Euler code uses time marching to the steady solution with local time steps. An optimization of the reverse mode of AD comes from the fact that, our situations of interest being stationary in time, there is no interest in storing the forward states for backward integration [5, 36, 40].

10 Concluding Remarks

In order to be easily integrated in engineering environments to quantify our confidence on optimal solutions without intensive sampling of large dimensional parameter spaces a cascade of geometric uncertainty quantification concepts has been presented. The cascade is based on the application of data analysis concepts together with existing deterministic simulation platforms.

The analysis starts with the geometric characterization of global sensitivity spaces through their dimensions and relative positions by the principal angles between global search subspaces. Then, joining a multi-point descent direction and extreme values information from the probability density functions of design variables the concept of DES has been introduced.

The construction goes beyond DES with EnKF after the multi-point optimization algorithm is cast into an ensemble simulation environment. This permits to account for the variability on the functioning parameters through the multi-point formulation and for the variability on the optimization parameters and observation data through the EnKF formulation.

The joint application of the EnKF and DES leads to the concept of EDES which provides exhaustive possible extreme scenarios knowing the PDF of the optimization parameters and this without a sampling of the admissible space.

The UQ cascade ends with low-complexity solutions for reverse propagation of aleatory uncertain target data in inverse design with two approximations of the covariance matrix of the optimization parameters. These provide uncertainty quantification analysis for the inversion solution with confidence margins on the design parameters in very large design spaces. The constructions also permit to account for epistemic uncertainties considering a model or solution procedure as always imperfect. Hence, seeing the associated error as uncertainty these reverse propagation constructions provide a quantification of the impact of these weaknesses on the design.

Acknowledgements The author would like to thank M. Meaux and F. Gallard from Airbus, M. Montagnac from CERFACS, and G. Rogé from Dassault Aviation for their feedback. The different multiple right-hand side adjoint solvers used for the definition of descent directions in multi-point and constrained optimization problems have been obtained using Tapenade AD tool developed at INRIA-Sophia Antipolis by L. Hascoet and his team [21].

References

1. Anderson, B., Moore, J.: Optimal Filtering. Prentice-Hall, New York (1979)
2. Bunch, J.R., Kaufman, L.: Some stable methods for calculating inertia and solving symmetric linear systems. Math. Comput. **31**(137), 163–179 (1997)
3. Bungartz, H.-J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)
4. Casella, G., Berger, R.: Statistical Inference, 2nd edn. Duxbury Press, London (2001)
5. Christianson, B.: Reverse accumulation and implicit functions. Optim. Methods Softw. **9**(4), 307–322 (1998)
6. Cinnella, P., Hercus, S.J.: Robust optimization of dense gas flows under uncertain operating conditions. Comput. Fluids **39**, 1893–1908 (2010)
7. Correa, C., Congedoa, P.M., Martinez, J.-M.: Shape optimization of an airfoil in a BZT flow with multiple-source uncertainties. Comput. Methods Appl. Mech. Eng. **200**, 216–232 (2011)
8. Courrieu, P.: Fast computation of Moore-Penrose inverse matrices. CoRR, abs/0804.4809 (2008)
9. Curioni, A., Kalantzis, V., Bekas, C., Gallopoulos, E.: Accelerating data uncertainty quantification by solving linear systems with multiple right-hand sides. Numer. Algorithms **62**(2), 637–653 (2014)
10. Dervieux, A.: Steady Euler simulations using unstructured meshes. In: Proceedings of the VKI Lecture Series, 1985/04, pp. 23–64. World Scientific, Singapore (1985). Revised version published in Partial Differential Equations of Hyperbolique Type and Applications
11. Dywayne, A., Nicely, A., Abdou, M., Morgan, R.B., Wilcox, W.: Deflated and restarted symmetric Lanczos methods for eigenvalues and linear equations with multiple right-hand sides. SIAM J. Sci. Comput. **32**(1), 129–149 (2010)
12. Evensen, G.: Advanced data assimilation for strongly nonlinear dynamics. Mon. Weather Rev. **125**, 1342–1354 (1997)
13. Evensen, G.: Sequential data assimilation for nonlinear dynamics: the ensemble Kalman filter. In: Ocean Forecasting: Conceptual Basis and Applications. Springer, Heidelberg (2002)
14. Farhat, C., Degand, C.: A three-dimensional torsional spring analogy method for unstructured dynamic meshes. Comput. Struct. **80**(3), 305–316 (2002)
15. Gallard, F., Mohammadi, B., Montagnac, M., Meaux, M.: An adaptive multipoint formulation for robust parametric optimization. J. Optim. Theory Appl. **165**(1) (2014). doi:10.1007/s10957-014-0595-6
16. Gelb, A.: Stochastic Processes and Filtering Theory. Academic Press, New York (1970)
17. Gelb, A.: Applied Optimal Estimation. MIT Press, Boston (1974)
18. Ghanem, R., Doostan, A.: On the construction and analysis of stochastic models: characterization and propagation of the errors associated with limited data. J. Comput. Phys. **217**, 63–81 (2006)
19. Ghanem, R., Spanos, P.: Stochastic Finite Elements: A Spectral Approach. Springer, New York (1991)
20. Ghil, M., Ide, K., Courtier, P., Lorenc, A.: Unified notation for data assimilation: operational, sequential and variational. J. Meteorol. Soc. Jpn. **75**(1B), 181–189 (1997)
21. Hascoet, L., Pascual, V.: Tapenade user's guide. INRIA Technical Report, INRIA, pp. 1–31 (2004)
22. Hoel, P.G.: Introduction to Mathematical Statistics. Wiley, London (1971)
23. Iaccarino, G.: Quantification of Uncertainty in Flow Simulations Using Probabilistic Methods. VKI Lecture Series (2008)
24. Jahn, J.: Vector Optimization: Theory, Applications and Extensions. Springer, Berlin (2004)
25. Jiang, S.: Angles between Euclidean subspaces. Geom. Dedicata **36**(2), 113–121 (1996)
26. Jordan, C.: Essay on geometry in n dimensions. Bull. Soc. Math. Fr. **3**, 103–174 (1875)
27. Jorion, Ph.: Value at Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York (2006)

28. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82**, 35–45 (1960)
29. Lindman, H.R.: Analysis of Variance in Complex Experimental Designs. Freeman, New York (1974)
30. Melchers, R.E.: Structural Reliability Analysis and Prediction. Wiley, Chichester (1999)
31. Mohammadi, B.: Reduced sampling and incomplete sensitivity for low-complexity robust parametric optimization. Int. J. Numer. Methods Fluids **73**(4), 307–323 (2013)
32. Mohammadi, B.: Principal angles between subspaces and reduced order modeling accuracy in optimization. Struct. Multidiscip. Optim. **50**(2), 237–252 (2014)
33. Mohammadi, B.: Uncertainty quantification by geometric characterization of sensitivity spaces. Comput. Methods Appl. Mech. Eng. **280**, 197–221 (2014)
34. Mohammadi, B.: Value at risk for confidence level quantifications in robust engineering optimization. Optimal Control Appl. Methods **35**(2), 179–190 (2014)
35. Mohammadi, B.: Ensemble Kalman filters (EnKF) and geometric characterization of sensitivity spaces for uncertainty quantification in optimization. Comput. Methods Appl. Mech. Eng. **290**, 228–249 (2015)
36. Mohammadi, B.: Parallel reverse time integration and reduced order models. J. Comput. Math. **2**, 17–33 (2015)
37. Mohammadi, B.: Backward uncertainty propagation in shape optimization. Int. J. Numer. Methods Fluids **103**(4), 307–323 (2016). doi:10.1002/fld.4077
38. Mohammadi, B., Bouchette, F.: Extreme scenarios for the evolution of a soft bed interacting with a fluid using the value at risk of the bed characteristics. Comput. Fluids **89**, 22–46 (2014)
39. Mohammadi, B., Pironneau, O.: Shape optimization in fluid mechanics. Annu. Rev. Fluid Mech. **36**(1), 255–279 (2004)
40. Mohammadi, B., Pironneau, O.: Applied Shape Optimization for Fluids, 2nd edn. Oxford University Press, Oxford (2009)
41. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (2006)
42. Obinata, G., Anderson, B.: Model Reduction for Control System Design. Springer, Berlin (2000)
43. Onez, R.O., Spooner, J.T., Maggiore, M., Passino, K.M.: Stable Adaptive Control and Estimation for Nonlinear Systems: Neural and Fuzzy Approximator Techniques. Wiley, New York (2002)
44. Peyret, M., Chery, J., Mohammadi, B., Joulain, C.: Plate rigidity inversion in Southern California using interseismic GPS velocity field. Geophys. J. Int. **187**(2), 783–796 (2011)
45. Qu, Z.: Model Order Reduction Techniques with Applications in Finite Element Analysis. Springer, Berlin (2004)
46. Redont, P., Mohammadi, B.: Improving the identification of general Pareto fronts by global optimization. C. R. Acad. Sci. Paris **347**, 327–331 (2009)
47. Scheinberg, K., Conn, A., Vicente, L.: Introduction to Derivative-Free Optimization. SIAM, New York (2002)
48. Schilders, W., Van der Vorst, H., Rommes, J.: Model Order Reduction: Theory, Research Aspects and Applications. In: Mathematics in Industry, vol. 13. Springer, Berlin (2008)
49. Shonkwiler, C.: Poincare duality angles for Riemannian manifolds with boundary. Ph.D. Thesis, University of Pennsylvania (2009)
50. Smolyak, S.A.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dokl. Akad. Nauk SSSR **148**, 1042–1043 (1963). Russian; Engl. Transl.: Soviet Math. Dokl. **4**, 240–243 (1963)
51. Tang, Z., Periaux, J.: Uncertainty based robust optimization method for drag minimization problems in aerodynamics. Comput. Methods Appl. Mech. Eng. **12**(24), 217–220 (2012)
52. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, New York (1987)
53. Veroy, K., Patera, A.: Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations: rigorous reduced-basis a posteriori error bounds. Int. J. Numer. Methods Fluids **47**(8), 773–788 (2005)

54. Wan, X., Karniadakis, G.E.: Multi-element generalized polynomial chaos for arbitrary probability measures. SIAM J. Sci. Comput. **28**(3), 901–928 (2006)
55. Warner, F., Gluck, H.: Great circle fibrations of the three-sphere. Duke Math. J. **50**, 107–132 (1983)
56. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer, New York (2004). ISBN:0-387-40272-1
57. Wuchner, R., Firl, M., Bletzinger, K.: Regularization of shape optimization problems using fe-based parametrization. Struct. Multidiscip. Optim. **47**(4), 507–521 (2013)
58. Xiu, D.: Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, Princeton (2010)

Aerodynamic Design of 'Box Blade' and 'Non-planar' Wind Turbines

Luigi Molea, Emanuele Di Vitantonio, and Aldo Frediani

Abstract In this paper the aerodynamic efficiency of wind turbines with horizontal axis is discussed and the so-called box blade concept, inspired by the Prandtl's 'Best Wing System' is analysed; this wind turbine configuration is proved to be efficient than a conventional blade. Moreover, other non-planar blades, such as the winglet and C extension are analysed via vortex theory with the numerical method of Ribner and Foster for the optimum circulation and the recent model of Okulov and Sørensen for the performance evaluation; a generalization of the above mentioned models is also presented in this work. Finally, the box blades are verified by means of a commercial CFD software.

1 Introduction

The use of power from the wind to spin a wheel can be dated back to 200 BC when the first wind mills were built in Persia. During the following centuries wind energy was used only for agricultural purposes; the Dutch were the first to employ wind energy to move a pump for water draining.

Only after the invention of the dynamo it was possible to built the first wind generator for electricity (Duc de la Peltrie, 1887). The first wind turbines were drag based machines and, thus, with a low efficiency. After the studies on aerodynamics on thin lifting bodies it was clear how lift-driven machines could be much more performing. However, wind generators were not considered industrially appealing until the oil crisis of the 1970s. This crisis caused a boost in the wind energy growth, in which countries like Denmark played a main role, since they first extensively employed this technology. Nowadays, the state of the art on the design of wind turbines blades has reached a full maturity, and further increases in performances are achieved, for example, by developing suited airfoils for wind turbines, optimizing the wind farm arrays, or reducing the mechanical losses of the transmission. A more efficient design could provide more power extracted for a given wind speed and a given blade radius and, thus, it reduces the costs of the energy produced, provided

L. Molea • E. Di Vitantonio (✉) • A. Frediani
University of Pisa, Pisa, Italy
e-mail: luigi.molea@gmail.com; e.divitantonio@gmail.com; a.frediani@dia.unipi.it

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_13

that more expensive structures are not necessary. But the aerodynamic performances of conventional wind turbines can be hardly improved after having reached the full maturity.

A similar situation occurs in the case of civil aircrafts: indeed, significant aerodynamic improvement of efficiency will be obtained only by introducing new architectural configurations. This is the direction the PrandtlPlane concept is moving towards: it consists in a box wing in the front view, with a proper aerodynamic design of the wings, and allows to minimize the induced drag. The same basic idea is studied in this paper applied to a wind turbine.

The aerodynamics of conventional blades uses the models developed during the second decade of the twentieth century, with the fundamental contributions of Prandtl, Pistolesi, Goldstein, and others.

The main task of this work is to demonstrate the applicability of this concept to the design of a wind turbine for the first time. In particular a theoretical model is derived to study any non-planar blade configuration and a PrandtlPlane blade is designed. Then the power extracted from the wind by this new blade concept is compared with a conventional blade designed by the authors according to the optimum rotor model. It is shown that the PrandtlPlane blade increases the power coefficient in a percentage ranging from 2.5 to 4.5 % in absolute terms, that correspond to an increase of 5–9 % in relative terms; CFD analysis confirmed these results. This is a significant increase of performance but, during this work, only aerodynamic aspects are analysed and no other issues such as structural design or aeroelastic phenomena are taken into account.

2 Elementary Theories

In the design of a wind turbine the fundamental problem is to maximize the power extracted by the wind.

We define the power coefficient as:

$$C_P = \frac{P}{\frac{1}{2}\rho\pi R^2 V^3}\ , \tag{1}$$

where P is the power extracted, R is the radius of the rotor, and V is the wind speed.

According to the *simple momentum theory* [8, 14], the maximum value achievable is $C_P = 0.593$, known as the 'Betz limit'.

An improvement of the simple momentum theory is given by the *general momentum theory* [10], which considers the wake rotation downstream the turbine disk; the main result is that C_P depends on the *tip speed ratio*, a fundamental dimensionless parameter, defined as:

$$\lambda = \frac{\Omega R}{V}\ , \tag{2}$$

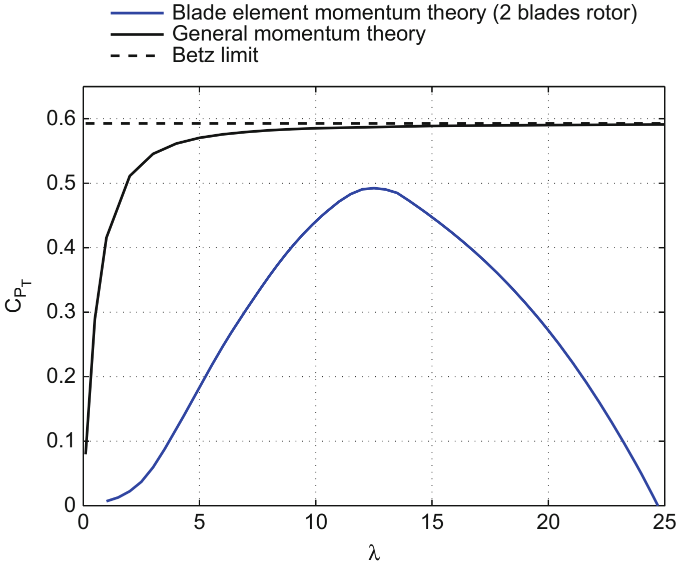


Fig. 1 Comparison among the $C_P - \lambda$ curves of the elementary theories

where Ω is the rotational speed of the turbine.

Another basic theory for the performance analysis of turbines is the *blade element momentum theory* which, contrary to the previous models, takes into account the finite number of blades and friction forces; the finite number of blades causes an increase of the wake losses and the friction forces limit the maximum value of C_P, depending on the airfoil selected.

The results for the basic theories mentioned before are summarized in Fig. 1.

3 The Model of Optimum Rotor

The condition of optimum rotor is derived by writing the functional of the power extracted from the wind in the Trefftz plane, by means of the vortex theory, which considers an irrotational and incompressible flow. The blade is assumed as a rotating lifting line from which leaves a helicoidal vortex sheet in the stream; this helicoidal vortices induce a velocity field on the blade according to the Biot–Savart law. The maximization of the functional with respect to blade circulation leads to the *Betz condition* [22], which states that the helicoidal sheet moves rigidly along its axis.

Now we define the velocity triangle in the Trefftz plane, shown in Fig. 2, where **u** represents the total velocity field and **v** is the induced velocity one. Hence, v_ϑ and v_z are the tangential and axial induced velocities, respectively, and Φ is the helix angle. As shown in Fig. 2, a similar triangle is defined by the displacement velocity

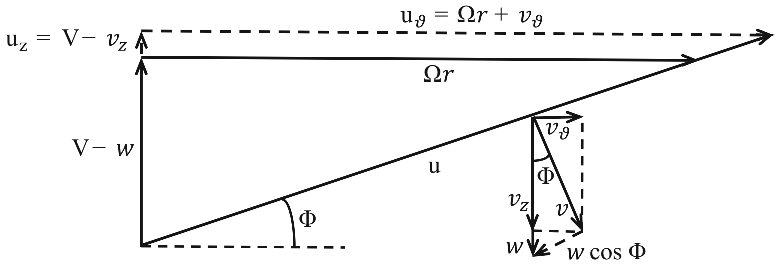


Fig. 2 Velocity triangle in the Trefftz plane

w of the wake relative to the wind, related to the prior defined quantities as follows:

$$w = v_z + v_\vartheta \tan \Phi \ . \tag{3}$$

The Betz condition is strictly true only for lightly loaded rotors, where $\bar{w} := w/V \ll 1$. In fact, since the pitch is defined as follows:

$$p = \frac{2\pi\ (V - w)}{\Omega R} \ , \tag{4}$$

if $w \ll V$, then the helix has a constant pitch. This condition could also be verified if we had a constant $\bar{w}$ but if the mean wake advance ratio is comparable to the wind speed it can be shown that it cannot be constant.

In this paper the Betz condition is used also for heavily loaded planar and non-planar rotors since, as shown later, it is a good approximation of the real wake behaviour.

4 Theoretical Model to Evaluate the Optimum Circulation in the Trefftz Plane

In the present section we introduce the analytical formulation of the problem and then, we propose a generalization of the Ribner and Foster numerical model [6] of a generic rotating lifting line, shown in Fig. 3. Also an example of the wake in the Trefftz plane shed from a planar blade is shown in Fig. 4. It can be observed how the screw surface extends infinitely both upstream and downstream.

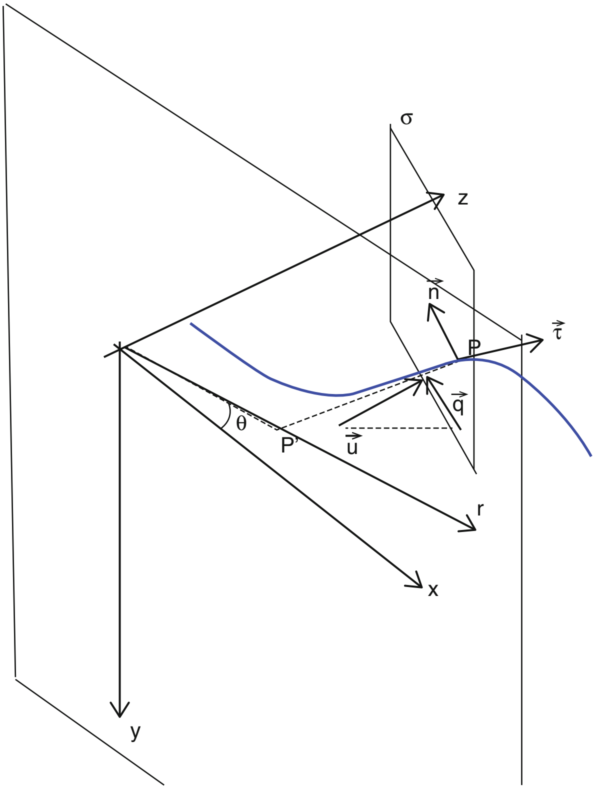


Fig. 3 Generic lifting line (*blue*) rotating about the z axis

4.1 Analytical Model

We define a Cartesian reference frame $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$ and a cylindrical one $(\mathbf{e}_r, \mathbf{e}_\vartheta, \mathbf{e}_z)$, both with the z axis lying on the rotor axis of rotation. The generic lifting line is described by a parametric curve given by the vector $[X = X(s), Y = Y(s), Z = Z(s)]$, where $s \in [s_1, s_2]$. Along the lifting line, a bound circulation $\mathbf{\Gamma}(s) = \boldsymbol{\tau}(s)\Gamma(s)$ is defined, where $\boldsymbol{\tau}$ is the local unit vector tangent to the curve. A helicoidal vortex surface is shed from the lifting line and is so parametrized in the Cartesian reference frame:

$$f(s, \bar{\Psi}) = \begin{bmatrix} r(s)\cos(\bar{\Psi} - \alpha(s)) \\ -r(s)\sin(\bar{\Psi} - \alpha(s)) \\ Z(s) + \frac{p(\bar{\Psi} - k\Theta_B)}{2\pi} \end{bmatrix} = \begin{bmatrix} X(s)\cos\bar{\Psi} + Y(s)\sin\bar{\Psi} \\ -X(s)\sin\bar{\Psi} + Y(s)\cos\bar{\Psi} \\ Z(s) + \frac{p(\bar{\Psi} - k\Theta_B)}{2\pi} \end{bmatrix}, \tag{5}$$

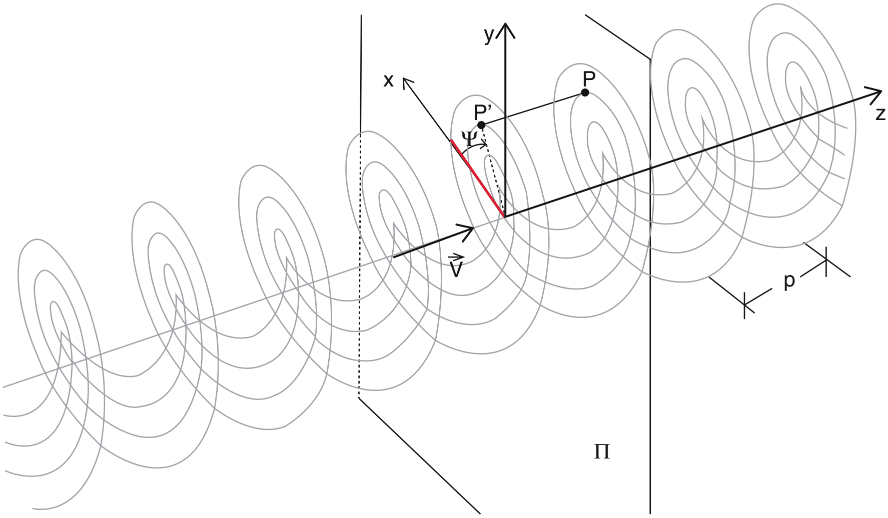


Fig. 4 Wake in the Trefftz plane of a planar blade

where $r(s) = \sqrt{\left(X(s)\right)^2 + \left(Y(s)\right)^2}$ is the distance between the generic point of the surface and the origin point of the reference frame; $\alpha(s) = \arctan(Y(s)/X(s))$ is the angle between the position vector of the generic point on the surface and the x–z plane, so that the generic point corresponding to $\bar{\Psi} = 0$ belongs to the lifting line; $\bar{\Psi}$ is the angle which parametrizes the generic helicoidal filament. The parameter $\bar{\Psi}$ of the kth helicoidal surface that is shed from the kth lifting line, which models the kth blade of a wind turbine rotor with B blades, is related to the azimuthal coordinate Ψ, which describes the blade selected as the first one, by the following:

$$\bar{\Psi} = \Psi + k\Theta_B\,, \qquad k = 1 \dots B-1\,, \tag{6}$$

where B is the number of blades and $\Theta_B = \frac{2\pi}{B}(k-1),\ k = 1\dots B$ is the angle between two corresponding points (i.e. they have the same abscissa s) lying on different blades.

To assign the proper boundary conditions to the problem we need to define the unit vector $\mathbf{n}$ normal to the screw surface. As known from the calculus, we have

$$\mathbf{n}(s,\bar{\Psi}) = \frac{f_{,s} \times f_{,\bar{\Psi}}}{|f_{,s} \times f_{,\bar{\Psi}}|} = \frac{1}{|f_{,s} \times f_{,\bar{\Psi}}|}\begin{vmatrix} \mathbf{e}_x & \mathbf{e}_y & \mathbf{e}_z \\ X'\cos\bar{\Psi} + Y'\sin\bar{\Psi} & -X'\sin\bar{\Psi} + Y'\cos\bar{\Psi} & Z' \\ -X\sin\bar{\Psi} + Y\cos\bar{\Psi} & -X\cos\bar{\Psi} - Y\sin\bar{\Psi} & \frac{p}{2\pi} \end{vmatrix} =$$

$$= \frac{1}{|f_{,s} \times f_{,\bar{\Psi}}|}\begin{bmatrix} -\frac{p}{2\pi}Y'\sin\bar{\Psi} + \frac{p}{2\pi}Y'\cos\bar{\Psi} + Z'X\cos\bar{\Psi} + Z'Y\sin\bar{\Psi} \\ -\frac{p}{2\pi}X'\cos\bar{\Psi} - \frac{p}{2\pi}Y'\sin\bar{\Psi} - Z'X\sin\bar{\Psi} + Z'Y\cos\bar{\Psi} \\ -X'X - Y'Y \end{bmatrix}\,, \tag{7}$$

thus the unit vector **n** evaluated at the generic induced point belonging to the lifting line ($s = \bar{s}, \bar{\Psi} = 0$) is given by:

$$\mathbf{n}(\bar{s}, \bar{\Psi} = 0) = \frac{\bar{\mathbf{n}}}{|\bar{\mathbf{n}}|}\ , \qquad \text{where} \qquad \bar{\mathbf{n}} = \begin{bmatrix} y'(\bar{s})r(\bar{s}) \tan\Phi + x(\bar{s})z'(\bar{s}) \\ -x'(\bar{s})r(\bar{s}) \tan\Phi + y(\bar{s})z'(\bar{s}) \\ -x(\bar{s})x'(\bar{s}) - y(\bar{s})y'(\bar{s}) \end{bmatrix} \tag{8}$$

and the superscript ‘ $'$ ’ represents the derivative with respect to s.

The following notation is assumed:

- the subscript ‘ 0 ’ represents the velocities in the rotor plane;
- the symbol ‘ ˆ ’ represents the dimensionless velocity with respect to w:

$$\hat{v}_{\alpha_0} = \frac{v_{\alpha_0}}{w}\ ;$$

- the symbol ‘ ¯ ’ represents the dimensionless velocity with respect to V:

$$\bar{v}_{\alpha_0} = \frac{v_{\alpha_0}}{V}\ ,$$

where α represents the αth component.

So, we can write

$$\bar{v}_{\alpha_0} = \bar{w}\hat{v}_{\alpha_0}\ . \tag{9}$$

Considering an irrotational and incompressible flow, the induced velocity field derives from a potential φ and resolves the Laplace equation $\nabla^2\varphi = 0$ with the following boundary conditions:

$$\begin{cases} \dfrac{\partial\varphi}{\partial n} = wn_z & \text{on the wake,} \\ \varphi|_{S_\infty} = 0\ , \end{cases} \tag{10}$$

where the Neumann boundary condition is applied on the wake surface and states that the projection of the induced velocity along the unit vector normal to the screw surface is equal displacement velocity in the wake along its normal direction; the second boundary condition simply states that the potential goes to zero very far from the wake axis. A numerical method to solve this problem is explained in the following. There is an analytical solution for this problem due to Goldstein [11], applicable only for lightly loaded planar rotors; then Theodorsen was the first to try to generalize this model in the more general case of heavily loaded planar rotors.

4.2 Numerical Model

In the present section a generalization of the original Ribner and Foster method is presented in order to extend this model to non-planar lifting lines.

The core of the model is that it enables to calculate the dimensionless strength of the vortex filaments by solving a system of linear equations. A finite number of N_v inducing points is defined along the lifting line, from which screw vortex filaments are shed; each filament is discretized by a finite number of segments that induce velocities at the $N_v - 1$ control points located in the middle of two adjacent inducing points. The dimensionless bound circulation, K, and the dimensionless vortex filaments strength, $\bar{\gamma}$, are defined by the equations below:

$$K_j = \frac{B}{p\,w}\Gamma_j \ , \qquad \bar{\gamma}_j = \frac{B}{p\,w}\gamma_j \tag{11}$$

and are related to each other by the following:

$$K_j = \bar{\gamma}_j - \bar{\gamma}_{j-1} \ , \tag{12}$$

where Γ_j is the bound circulation relevant to the jth inducing point.

The αth component of the induced velocity v_i^α in the ith control point is seen as the sum of each elementary contribution α-th, v_{ij}^α given by the jth helicoidal vortex, and so we have

$$v_i^\alpha = \sum_j v_{ij}^\alpha \ , \tag{13}$$

where:

$$v_{ij}^\alpha = c_{ij}^\alpha \bar{\gamma}_j \tag{14}$$

and c_{ij}^α is the αth component of the induced velocity at the ith collocation point by the jth unit vortex, N_i^α is the αth component of the unit vector normal to the screw surface at collocation point i.

The solving system derives from the discrete form of the Neumann's boundary condition:

$$F_{ij} = \sum_\alpha c_{ij}^\alpha N_i^\alpha \ , \tag{15}$$

$$\sum_j F_{ij}\bar{\gamma}_j = n_{z_i} \quad \text{for} \quad i = 1 \ldots N_v - 1 \ , \tag{16}$$

where n_{z_i} is the z component of the unit vector normal to the screw surface at the point i.

The dimensionless coefficients of influence are given by combining Eq. (14) and the Biot–Savart law in discrete form; called $\Delta\bar{\Psi}_v$ the angular integration pitch along the generical vortex filament and **dl** the elementary displacement along the helicoidal vortex, we have

$$c_{ij}^{\alpha} = \sum_{k=1}^{B} \sum_{v=-2NR\pi}^{2NR\pi} \frac{L}{4\pi} \left(\frac{\mathbf{dl} \times (\mathbf{r_i} - \mathbf{r_j})}{|\mathbf{r_i} - \mathbf{r_j}|} \right) \Bigg|_{\alpha} \Delta\bar{\Psi}_v \ , \tag{17}$$

where $L = p/B$; the vector $\mathbf{r_i}$ identifies the ith control point in the system of reference and is given by fixing the generic abscissa $\bar{s}$ and $\bar{\Psi} = 0$ in Eq. (5) and, similarly, $\mathbf{r_j}$ is the position vector of the points of the surface, given by substituting $s = \bar{s}$ and $\bar{\Psi} = \bar{\Psi}_v$ in Eq. (5); NR is the number of the helix turns for the numerical integration. The parameter NR must be chosen such that the integration ends far away from the disk, since further contributions of the Biot–Savart induction formula do not influence significantly the result.

As clear from Eq. (16), there are $N_v - 1$ equations and N_v unknowns which are gathered in the vector $\bar{\boldsymbol{\gamma}}$. The last relationship is given by the total vorticity conservation:

$$\sum_j \bar{\gamma}_j = 0 \ . \tag{18}$$

To include this last relationship we need to add a row of 1s to the matrix F and also a 0 for the last term of the vector $\boldsymbol{\Phi}$ of the known terms:

$$\mathsf{F}(N_v, 1 \ldots N_v) = [1 \ldots 1] \ , \qquad \cos\Phi(N_v) = 0 \ . \tag{19}$$

If we consider a lifting line given by a straight segment lying on the x axis, extending from $r = 0$ to $r = R$ (the rotor radius) we are find the classical Goldstein function which gives the optimum circulation distribution for a planar blade. In Fig. 5 a comparison is reported between the numerical method and the analytical solution. The agreement between the two models is high. This solution was calculated by numerical integration of the Biot–Savart law along each vortex filament, which is quite computationally heavy.

However, recently, Okulov derived an analytical formula for the calculation of the induced velocity field of an infinite helicoidal vortex [13, 18]; according to Okulov, the coefficients c_{ij}^{α}, in cylindrical coordinates, are so expressed:

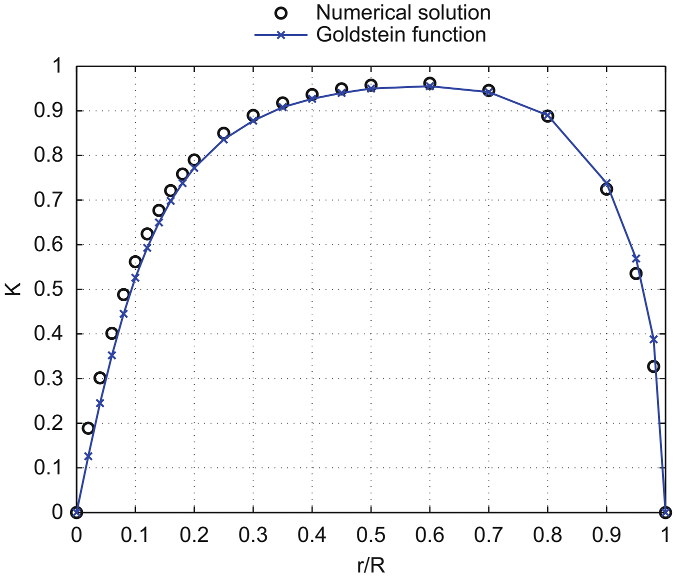


Fig. 5 Comparison between Goldstein and the present numerical solution

$$c_{ij}^{r} \cong -\frac{L}{2\pi r_i}\frac{\sqrt[4]{\left(l^2+r_i^2\right)\left(l^2+r_j^2\right)}}{l} Im\Bigg[\frac{e^{i\chi}}{e^{\mp\xi}-e^{i\chi}}+ \\ \pm\frac{l}{24}\left(\frac{2l^2+9r_j^2}{\left(l^2+r_j^2\right)^{\frac{3}{2}}}-\frac{2l^2+9r_i^2}{\left(l^2+r_i^2\right)^{\frac{3}{2}}}\right)\ln\left(1-e^{\pm\xi+i\chi}\right)\Bigg], \tag{20}$$

$$c_{ij}^{z} \cong \frac{L}{2\pi l}\begin{Bmatrix}1\\0\end{Bmatrix}+\frac{L}{2\pi l}\frac{\sqrt[4]{l^2+r_i 2}}{\sqrt[4]{l^2+r_i^2}}Re\Bigg[\frac{\pm e^{i\chi}}{e^{\mp\xi}-e^{i\chi}}+ \\ -\frac{l}{24}\left(\frac{3r_i^2-2l^2}{\left(l^2+r_i^2\right)^{\frac{3}{2}}}-\frac{9r_j^2+2l^2}{\left(l^2+r_j^2\right)^{\frac{3}{2}}}\right)\ln\left(1-e^{\pm\xi+i\chi}\right)\Bigg], \tag{21}$$

$$c_{ij}^{\vartheta}=(c_0-c_{ij}^{z})\,\frac{l}{r_i}\ , \tag{22}$$

where

$$l = \frac{p}{2\pi}\,, \quad c_0 = \frac{L}{2\pi l}\,, \quad \chi = \vartheta - \frac{z}{l}\,, \quad e^{\xi} = \frac{r_i\left(1+\sqrt{1+\frac{r_j^2}{l^2}}\right)e^{\sqrt{1+\frac{r_i^2}{l^2}}}}{r_j\left(1+\sqrt{1+\frac{r_i^2}{l^2}}\right)e^{\sqrt{1+\frac{r_j^2}{l^2}}}}. \tag{23}$$

‘$\pm$’ and ‘$\mp$’ mean that the upper and the lower signs are employed when $r_i < r_j$ and $r_i > r_j$, respectively.

By employing these formulas, the computation time decreases by more than 1000 times with respect to numerical integration, thus allowing to simulate a huge number of vortices.

The problem of a heavily loaded rotor is non-linear since the wake pitch depends on the axial displacement velocity w which is not known a priori. Hence the value of the wake non-dimensional pitch λ_T is fixed in advance. The rotor tip speed ratio is then derived as follows:

$$\lambda = \frac{\Omega_0 R}{V} = \left(1 - \frac{\bar{w}}{2}\right)\lambda_T\,. \tag{24}$$

The calculation procedure for the dimensionless circulation K is summarized in Fig. 6.

The theoretical model developed during this work has general validity; however, the numerical code in the Trefftz plane has been written in order to solve geometries composed by polygonal chain. In the subsequent sections the results for different configuration and a comparison among them are shown and the influence of geometric parameters is analysed.

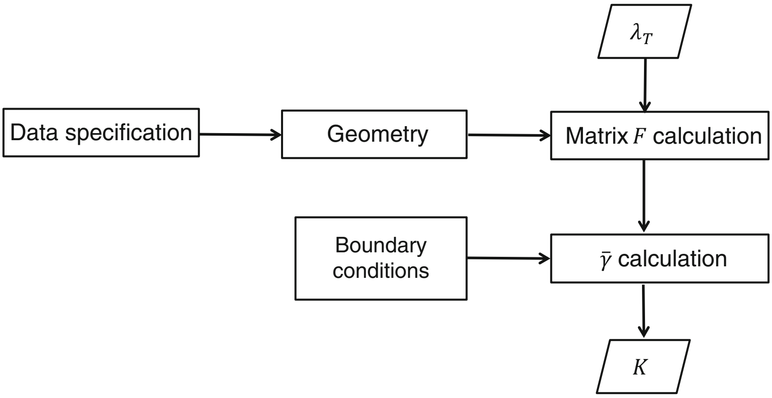


Fig. 6 Calculation procedure for the optimum circulation

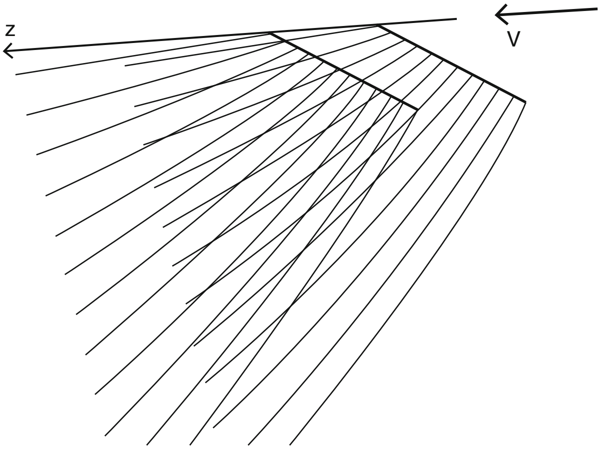


Fig. 7 Wake of a biplane blade

The analysis carried out in the plane far downstream the turbine disk considers infinite helicoidal filaments extended in both the vectors of the wind direction; this procedure does not produce relevant errors if applied to a classical blade: in this case the induced velocities on the rotor plane are half of that ones calculated in the Trefftz plane because of the absence of the wake deformation. Instead this is not also strictly true for non-planar geometries: considering the biplane blade shown in Fig. 7, while they would see the same vortex system in the Trefftz plane, we can observe how in the rotor plane this symmetry disappears: the aft blade is influenced by the vortex filaments more than the fore one and so the lifting force generated by the aft blade is lower. However, this model gives us quickly some reliable and useful pieces of information about the location of the design point of the turbine that represent a good starting point for a more precise calculation of the power coefficient for non-planar configurations and the corresponding geometry. These more accurate analyses are given by the numerical model in the rotor plane pointed out in Sect. 10.

5 Okulov and Sørensen Model for the Performance Assessment

Given the circulation, we need a theoretical model to evaluate the performance of the wind turbine in terms of inviscid C_P. As reported in Appendix, the Okulov and Sørensen model [18–20] is in agreement with the simple and generalized momentum theory, unlike the Theodorsen theory, and thus it is employed. This model neglects the expansion and the roll-up of the wake; applying the blade element momentum theory directly on the rotor disk we obtain, for a classical turbine, the following formula:

$$C_P = 2\bar{w}I_1\left(1-\frac{\bar{w}}{2}\right)\left(1-\frac{\bar{w}}{2}\frac{I_3}{I_1}\right) \,, \tag{25}$$

where:

$$I_1 = 2\int_0^1 K(x)x\,dx \,, \tag{26}$$

$$I_3 = 2\int_0^1 K(x)\frac{x^3}{x^2+l^2}\,dx \,. \tag{27}$$

Deriving the expression of C_P with respect to the dimensionless displacement velocity of the wake we obtain the value of $\bar{w}$ which maximize the power extracted:

$$\bar{w} = \frac{2}{3I_3}\left(I_1+I_3-\sqrt{I_1^2+I_3^2-I_1I_3}\right) \,. \tag{28}$$

This model can be generalized for the non-planar lifting line introduced in the previous section. First of all let us write the expression for the elementary lift and drag along the abscissa s:

$$\mathbf{l}(s) = \rho\,\mathbf{u}(s)\times\boldsymbol{\tau}(s)\,\Gamma(s) \tag{29}$$

$$\mathbf{d}(s) = \rho\,\varepsilon(s)\,\boldsymbol{\tau}(s)\times\big(\mathbf{u}(s)\times\boldsymbol{\tau}(s)\big)\,\Gamma(s) \,, \tag{30}$$

where $\varepsilon = 1/E$ is the inverse of the aerodynamic efficiency. The dimensionless power coefficient C_P and the loss coefficient T_T are evaluated with the following formulas:

$$C_P = \frac{\int_s \rho\Gamma r\mathbf{u}\times\boldsymbol{\tau}\cdot\mathbf{e}_\vartheta\,ds}{\frac{1}{2}\rho\pi R^2V^3} \,, \tag{31}$$

$$T_T = \frac{\int_s \rho\Gamma\varepsilon r\boldsymbol{\tau}\times(\mathbf{u}\times\boldsymbol{\tau})\cdot\mathbf{e}_\vartheta\,ds}{\frac{1}{2}\rho\pi R^2V^3} \,. \tag{32}$$

After some algebraical calculation the final relationships are derived and we have

$$C_P = 4\frac{\Lambda}{R}\bar{w}\left(1-\frac{\bar{w}}{2}\right)\int_0^1\left[\sin\alpha\left(\tau_y\bar{u}_{z_0}-\tau_z\bar{u}_{y_0}\right)+\cos\alpha\left(\tau_x\bar{u}_{z_0}-\tau_z\bar{u}_{x_0}\right)\right]d\bar{s} \,, \tag{33}$$

in which x is the dimensionless radial coordinate r/R and $\Lambda = \int_{s_1}^{s_2} ds$ is the curve length.

Equation (33) can be rewritten in the same form of Eq. (25) by defining the general expressions of the coefficients I_1 and I_3:

$$I_1 = 2\frac{\Lambda}{R}\int_0^1 \left(\tau_y \sin\alpha + \tau_x \cos\alpha\right) Kx\, d\bar{s}\ , \tag{34}$$

$$I_3 = 2\frac{\Lambda}{R}\int_0^1 \left[\left(\hat{v}_{y0}\sin\alpha + \hat{v}_{x0}\cos\alpha\right)\tau_z - \hat{v}_{z0}\left(\tau_y\sin\alpha + \tau_x\cos\alpha\right)\right] Kx\, d\bar{s}\ . \tag{35}$$

The general form of the loss coefficient is also derived:

$$\begin{gathered} T_T = 4\frac{\Lambda}{R}\bar{w}\big(1-\frac{\bar{w}}{2}\big)\int_0^1 \Big\{\big(\tau_x\sin\alpha - \tau_y\cos\alpha\big)\big(1+\bar{v}_{z0}\big)\tau_z + \\ \Big[\big(\bar{v}_{y0}\sin\alpha - \bar{v}_{x0}\cos\alpha\big) - \lambda x\sin 2\alpha\Big]\tau_x\tau_y + \cos\alpha\big(\bar{v}_{y0} - \lambda x\cos\alpha\big)\tau_x^2 + \\ -\sin\alpha\big(\bar{v}_{x0} + \lambda x\sin\alpha\big)\tau_y^2 + \Big[\big(\bar{v}_{y0}\cos\alpha - \bar{v}_{x0}\sin\alpha\big) - \lambda x\Big]\tau_z^2\Big\} Kx\, d\bar{s}\ . \end{gathered} \tag{36}$$

Finally the total power coefficient can be calculated as follows:

$$C_{P_T} = C_P - T_T\ . \tag{37}$$

The numerical calculation of T_T is done through an iterative scheme which is summarized in Fig. 8. First of all we have to choose an airfoil; we consider a

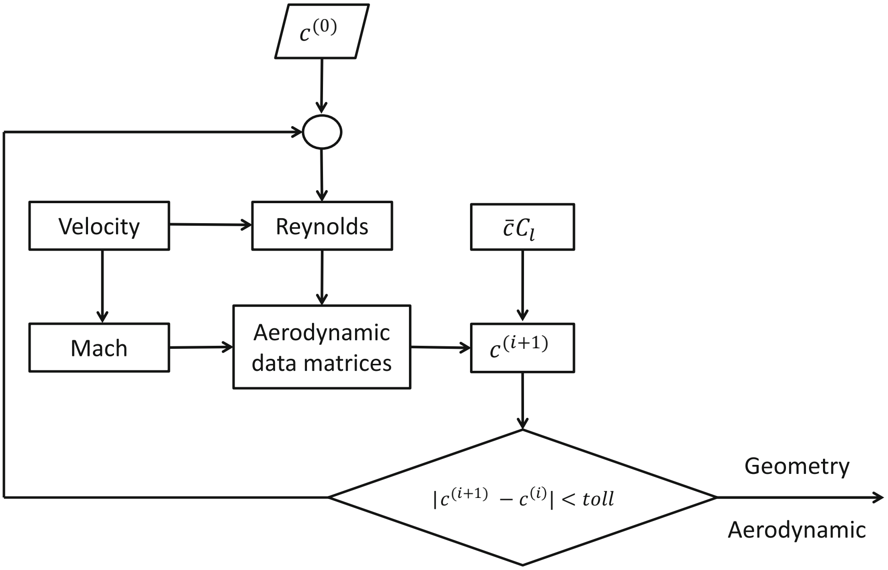


Fig. 8 Flow diagram of friction iterative cycle calculation

NACA 64-418 airfoil. Since the aim of this paper is to compare the aerodynamic performance of wind turbines, the effect of the airfoil is the same effect on all the blade configurations.

From the Kutta–Joukowski theorem it is possible to derive the trend of the product between the dimensionless chord and the local lift coefficient:

$$\bar{c}C_l = \frac{4\pi\bar{w}K}{B\lambda_T\bar{u}} \; . \tag{38}$$

From the latter it is clear that infinite solutions exist for the chord and twist distributions to produce the optimum circulation distribution. So an additional constraint must be added, and in this case is that all the airfoils work in maximum aerodynamic efficiency condition in order to minimize the friction losses.

The aerodynamic of the airfoils is evaluated with the software XFOIL, and a set of data for a large range of Reynolds and Mach numbers is produced. In particular matrices are defined in which are stored the values of the incidence angle α, the efficiency E, and the lift and drag coefficients C_l and C_d in condition of maximum efficiency. For the iterative procedure for each blade section an initial value of the chord is set, then from the solution the velocity and thus Re and M are known. From the product $\bar{c}C_l$ the value of the chord is updated and the procedure is repeated until convergence. Finally the twist distribution is derived with the following:

$$\vartheta = \Phi - \alpha \; . \tag{39}$$

6 The Classical Blade

The first application of the model explained until now is the design of a conventional blade to be used also for the comparison with the non-planar configurations. The first step of the design process is the evaluation of the inviscid power coefficient and the second one is the friction loss coefficient; thus, we construct the design curves of the rotors, consisting into the total power coefficient versus the tip speed ratio. The optimization process is carried out on the inviscid C_P and then the friction losses are evaluated such that they are minimum. Finding the optimum of the total power coefficient makes only a negligible difference; however, it would require an iterative scheme since T_T is not an explicit function of $\bar{w}$ but also depends on the airfoils Reynolds and Mach numbers.

The following specification is used for all the blades:

- number of blades: $B = 2$;
- rotor radius: $R = 20\,\text{m}$;
- hub radius: $R_m = R/20 = 1\,\text{m}$;
- design wind speed: $V = 10\,\text{m/s}$;
- air density: $\rho = 1.225\,\text{kg/m}^3$.

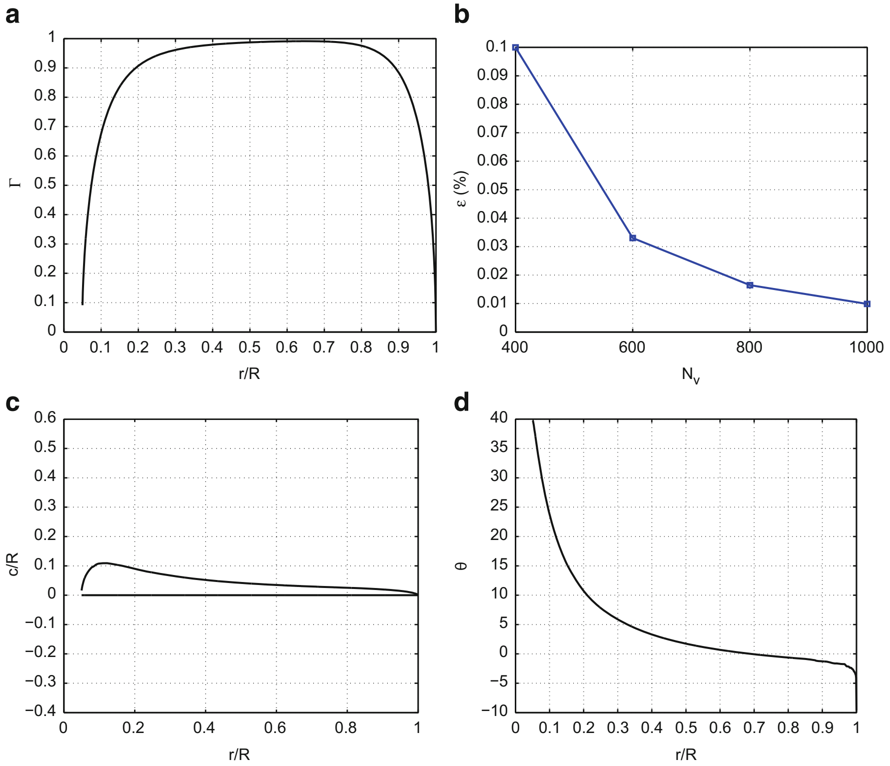


Fig. 9 Results for classical blade; $B = 2$, $\lambda_T = 18$. (**a**) Dimensionless circulation. (**b**) Convergence of C_{P_T}. (**c**) Chord distribution. (**d**) Twist distribution

The present data are only of reference, but the conclusions about the comparison between a conventional blade and a box blade have been validated with other examples. Figure 9 reports the results obtained: in particular, Fig. 9a refers to the optimum circulation around the blade, Fig. 9b illustrates the trend of convergence of the C_{P_T} coefficient; the chord and twist angle distributions along the span are shown in Fig. 9c and d, respectively.

The design curve, i.e., the locus of points of minimum wake and friction losses for a given value of λ, of the classical blade sized above is shown in Fig. 10.

7 The Blade with Winglet

In this section we analyse the performances of wind turbines equipped with winglets, made of a lateral wing at the blade tip. Only upwind winglets in Fig. 11 are considered because, as largely demonstrated in the literature [2, 9, 17], they

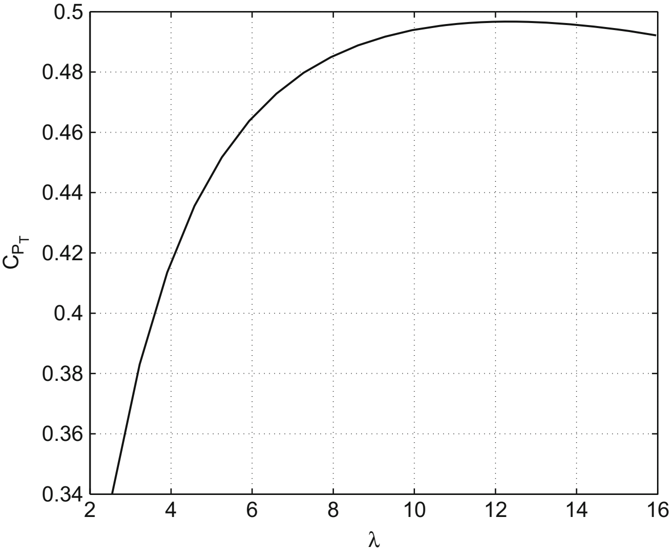
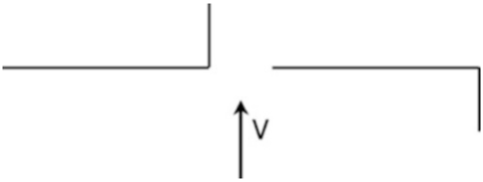


Fig. 10 Design curve of the classical blade; $R_m = 1\,\mathrm{m}$, $R = 20\,\mathrm{m}$, $B = 2$

Fig. 11 *Left*: upwind winglet; *right*: downwind winglet



are more efficient than the downwind ones. In an aircraft the winglet increases the aerodynamic efficiency of the wing by reducing the tip vortex strength; in a rotor it increases the load at the tip of the blade, producing an increase of the developed torque.

To analyse the problem of a blade with a winglet we assume that the Betz condition is valid independently of the blade geometry and, thus, the wake shed from blade still moves rigidly along the rotor axis. We indicate with n the number of vortices lying on the blade and with n_p the ones belonging to the winglet (Fig. 12), hence $N_v = n + n_p - 1$ is the total number of vortices. The inducing points are enumerated following the direction of the curvilinear abscissa. The winglet is discretized by equispaced vortices, in particular 1000 vortices on the blade and 1000 h/R on the winglet, where h/R is the ratio of the winglet height and the blade length and represents the characteristic dimensionless geometric parameter for this configuration. Both the lifting segments are in the condition of maximum aerodynamic efficiency.

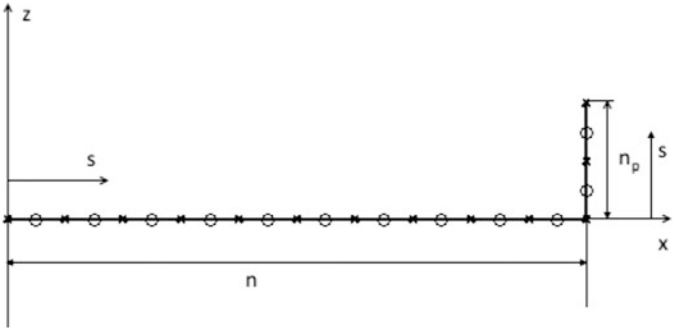


Fig. 12 Discretization of the wingletted blade; *times symbol* = inducing point, *open circle* = induced point

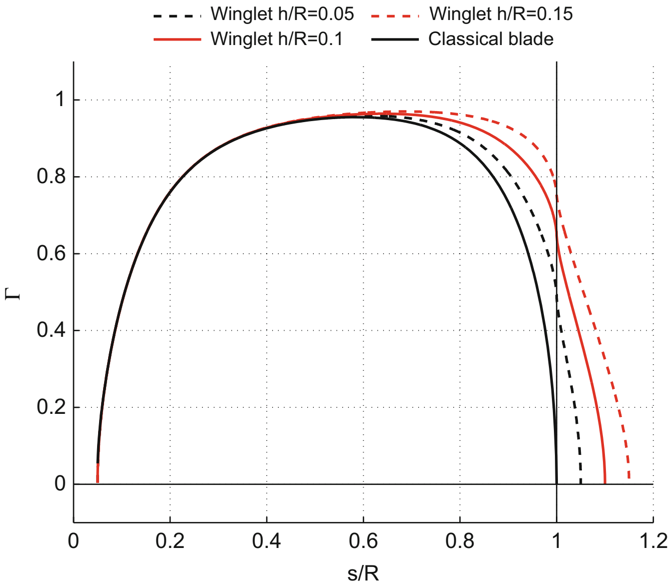


Fig. 13 Optimum circulation over the blade with winglet for various h/R ratios

Figure 13 shows the dimensionless circulation along the blade in the two cases of a classical blade and a blade with winglets, for different values of the ratio h/R; we can see that the tip load increases with h/R and, thus, we obtain a higher C_{P_T}. The maximum values of the design curves with different values of λ are shown in Table 1.

Table 1 Maximum power coefficients and correspondent λ values for various h/R ratios for the wingletted blades

h/R	λ_T	λ	$C_{P_{Tmax}}$
0.05	17	11.29	51.6
0.1	15	9.95	52.61
0.15	13	8.61	53.16
0.2	12	7.94	53.46

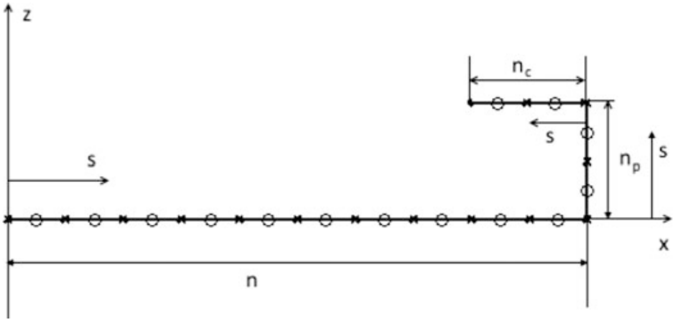


Fig. 14 Discretization of the C-blade; *times symbol* = inducing point, *open circle* = induced point

8 The C-blade

A blade with a C extension is an intermediate case between a blade with a winglet and a box blade, as shown in Fig. 14. There is a lateral wing, as in the case of winglet blade, and an aft blade of length R_c which represents an extension for the lateral blade. If n_c is the number of vortices lying on the C extension, the total number of vortices becomes $N_v = n + n_p + n_c - 2$. In this configuration, the circulation on three segments has equal sign up to $R_c = \bar{R}_c$ following the abscissa s and, thus, we have a down-force on the rear blade (Fig. 15a), while for $R > R_c$ we can observe that the circulation changes its sign on the rear blade (Fig. 15b) or on the lateral blade (Fig. 15c). In the first case the three lift segments are sized by means of maximum aerodynamic efficiency; instead in the others we have a point with zero load, where the chord would be equal to zero if this blade were sized by means of maximum aerodynamic efficiency, meaning physical disconnection for the blade. We can follow these criteria to avoid this problem:

- zero-load point on the lateral blade: the fore blade is sized by means of maximum aerodynamic efficiency and lateral blade characteristics are evaluated with constant chord, equal to the fore blade tip value; when the product $\bar{c}C_l$ is given by Eq. (38) the lift coefficient is known for each airfoil. With the values of chord and velocity, the corresponding values of M and Re can be calculated for each inducing point in order to obtain the $C_l - \alpha$ curve and the angle of attack of the airfoils as seen for the classical blade. Once the lateral blade sizing is over, we shall proceed by means of maximum aerodynamic efficiency design of the aft blade;

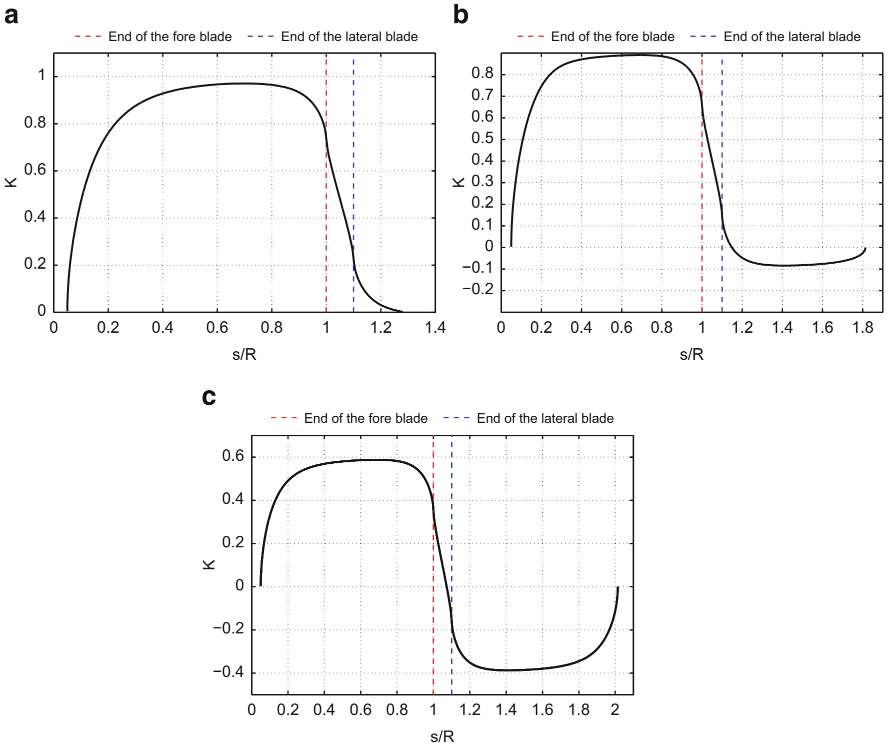


Fig. 15 Dimensionless circulation along the curvilinear abscissa of the C-blade for various values of R_c/R; $B = 2$, $h/R = 0.1$, $\lambda_T = 10$. (**a**) $R_c/R = 0.18$. (**b**) $R_c/R = 0.7$. (**c**) $R_c/R = 0.9$

- zero-load point on the rear blade: the front and lateral blades are sized by means of maximum aerodynamic efficiency while the rear one has constant chord, using the procedure seen before.

9 The Box Blade

The box blade represents a limiting case of the C-blade, in which the rear blade terminates on the hub.

The study of this blade configuration is inspired to the 'Best Wing System' aircraft configuration, based on an idea of Ludwig Prandtl. A considerable induced drag reduction is the remarkable characteristic of the configuration under discussion. According to Prandtl, once the lift is given, the wing system that minimizes the induced drag is the box wing, shown in Fig. 16, in which we have the same lift distribution on the horizontal wings and linear lift distribution with zero resultant

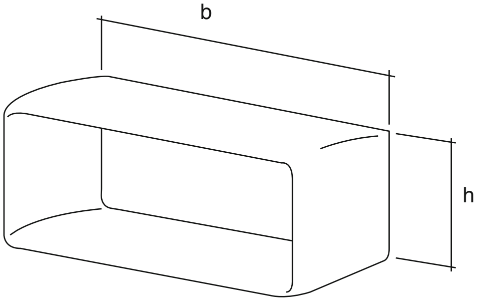


Fig. 16 Example of Prandtl's box wing concept

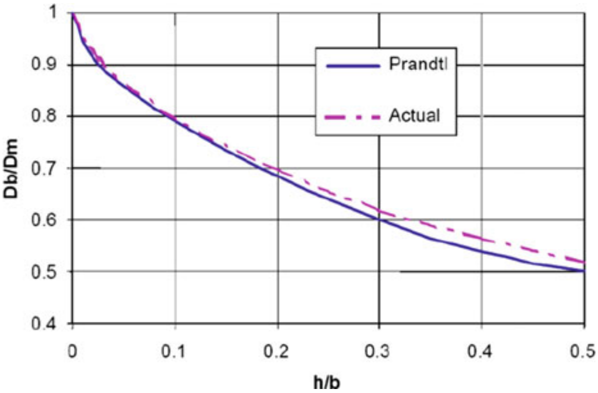


Fig. 17 Ratio of the monoplane induced drag over the box wing induced drag

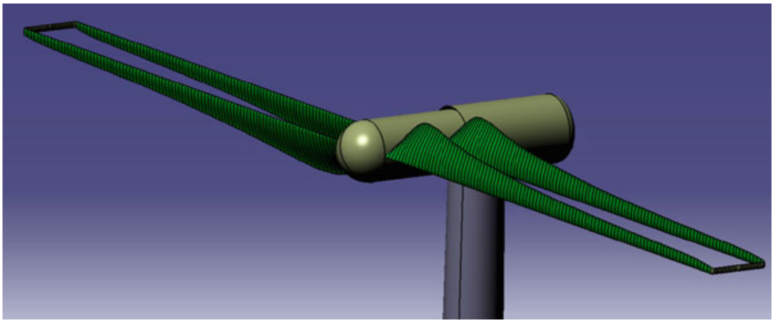


Fig. 18 Example of box blade, inspired to the Prandtl's 'Best Wing System' concept

on the lateral one. The above problem, that Prandtl gave an approximated solution, was solved in a closed form in 1999; the process is reported in [7]. Figure 17, from [4], shows the induced drag reduction of a box wing compared to a monoplane versus the h/b ratio.

In Fig. 18 an example of PrandtPlane blade is shown.

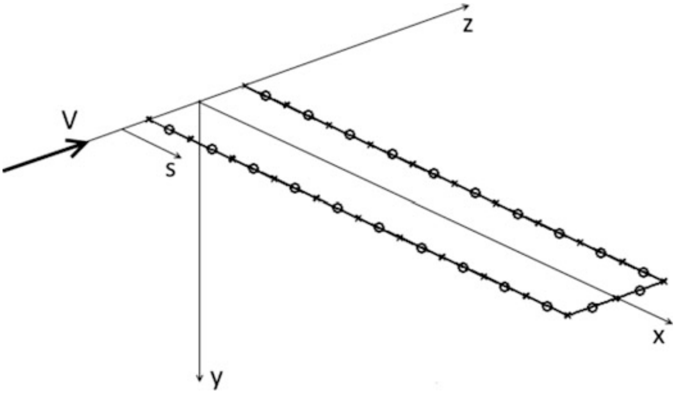


Fig. 19 Discretization of the lifting line describing the box blade; *times symbol* = inducing point, *open circle* = induced point

The configurations analysed in this work help us to evaluate the applicability of this concept to the blades, but nothing is said about the 'Best Blade System'; the problem of the optimum blade geometry that minimizes the wake losses needs an analytical solution not proved until now. In the present work we analyse configurations such that the box is nearly normal to the velocity direction at the tip, following the same technical specification used for the classical blade. All the cases analysed here consist of a rear blade with the same length $R = 20\,\mathrm{m}$ of the front one and a hub with radius $R_m = 1$ m; h represents the lateral blade length. Figure 19 shows the first configuration analysed, in which the whole blade belongs to the x–z plane and is discretized by $N_v = 2n + n_p - 2$ vortices, where n is the number of vortices on the front and rear blade and n_p is the number of vortices of the lateral one; the curvilinear abscissa, starting from the front blade at the hub, follows the lifting line up to the intersection between the rear blade and the hub.

As before said, the analysis is made in the Trefftz plane; we can observe that the wake is antisymmetric with respect to the x–y plane and thus the solution of the circulation along the blade must be antisymmetric with respect to that plane. We can observe that numerically the solution becomes more and more symmetrical with the increasing number of vortices. Therefore, it can be used a modified version of the code to impose the symmetry condition just mentioned: the vorticity conservation is not imposed on the whole blade, but on the two parts of the blade delimited by the intersection between the lifting line and the x–y; thus we have

$$\sum_{j=1}^{\frac{N_v}{2}} \bar{\gamma}_j = 0 \quad , \quad \sum_{j=\frac{N_v}{2}+1}^{N_v} \bar{\gamma}_j = 0 \,. \tag{40}$$

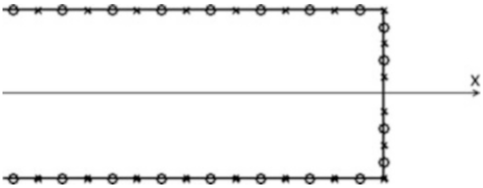


Fig. 20 Discretization of the box blade for the numerical code with symmetry

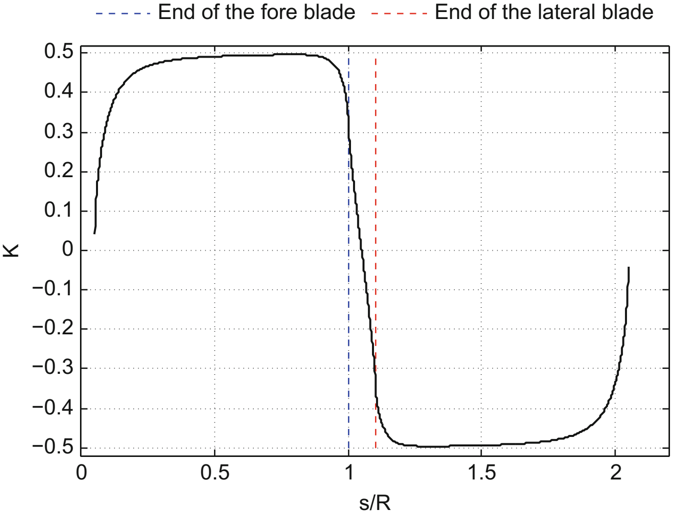


Fig. 21 Dimensionless circulation along the curvilinear abscissa which describes the box blade. $\lambda_T = 18, B = 2$

In a similar way to Eq. (19) the condition described above is imposed in the matrix F and in the constant terms vector, obtaining the following:

$$\begin{cases} \mathsf{F}\,(N_v - 1, 1 \ldots N_v) = [\underbrace{1 \ldots 1}_{\frac{N_v}{2}}, \underbrace{0 \ldots 0}_{\frac{N_v}{2}}] \\ \mathsf{F}\,(N_v, 1 \ldots N_v) = [\underbrace{0 \ldots 0}_{\frac{N_v}{2}}, \underbrace{1 \ldots 1}_{\frac{N_v}{2}}] \end{cases}, \tag{41}$$

$$\cos\Phi\,(N_v - 1) = 0 \qquad \cos\Phi\,(N_v) = 0\,. \tag{42}$$

By imposing the two conservation conditions, in order to have a square F matrix, the number of induced points must be equal to $N_v - 2$ and thus we decide to remove the middle point of the lateral blade (Fig. 20). The symmetry condition here employed enables to evaluate only the terms of induction on half blade, thus halving the time of computation. Figures 21 and 22 show the dimensionless load along the curvilinear abscissa and a comparison between the dimensionless circulation over the front

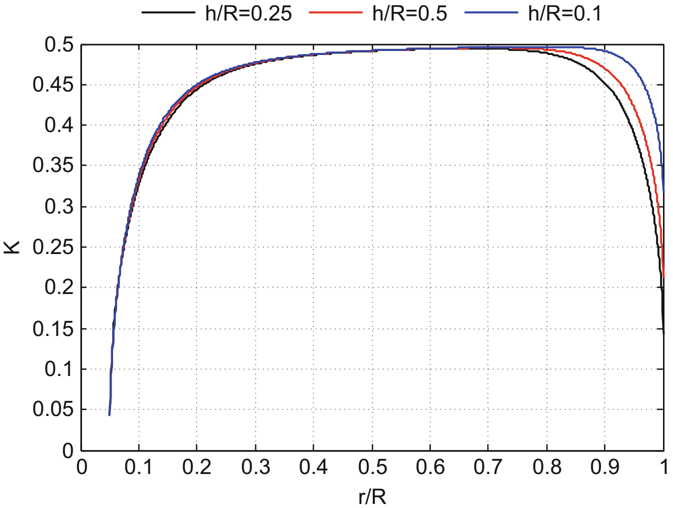


Fig. 22 Comparison between the dimensionless circulation of a radial blade of the box blade for various h/R

blade for different values of the length of the lateral blade; we can observe that the load near the tip increases with the increasing of the ratio h/R.

For the antisymmetric property earlier mentioned, the circulation is zero at the middle point of the lateral blade: hence, similarly to the case of the C-blade, we choose a constant chord sizing for the lateral blade and a maximum aerodynamic efficiency for the others. Furthermore, the load distribution on the lateral blade is multiplied by a coefficient in order to avoid the stall of the airfoils: for high value of $\bar{c}C_l$, if the chords are small the value of C_l is incompatible with the airfoil selected and thus the coefficient just introduced is needed to curtail the value of C_l to 1.1 at the inducing point where this value is the higher. In Fig. 23 the convergence of the C_{P_T} is reported, where the percentage error is so calculated:

$$\varepsilon_{\%} = \left(\frac{C_{P_T}\left(N_v^{(i+1)}\right) - C_{P_T}\left(N_v^{(i)}\right)}{C_{P_T}\left(N_v^{(i)}\right)} \right) \times 100\,\% \,. \tag{43}$$

In Figs. 24 and 25 the design curves are shown, calculated with $N_v = 2500$ vortices as the number of blades and the ratio h/R changes: as we can observe, the values of maximum of the curves rise and move towards left because of the increasing friction losses for smaller values of λ; the maximum values of the design curves are reported in Table 2.

The second box blade configuration analysed in this work has a relative inclination, δ, between the two radial blades around the rotor axis (Fig. 26). The symmetry of the wake in the Trefftz plane allows us to use the numerical code with symmetry even in this case. The design curves are shown in Fig. 27, with varying

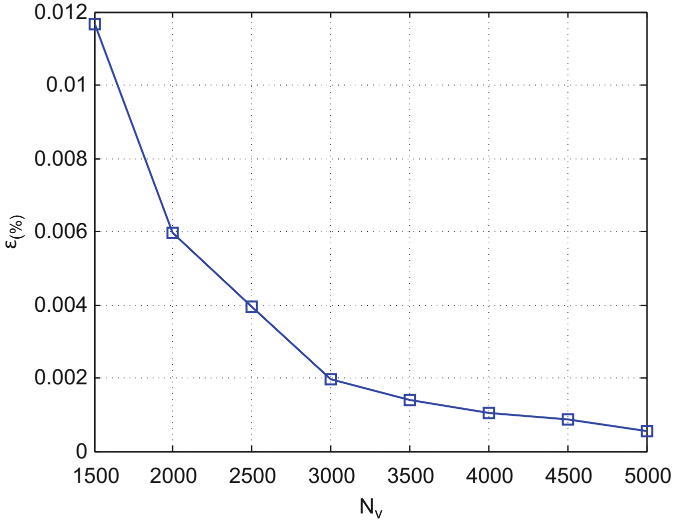


Fig. 23 Convergence of the C_{P_T} for the box blade; $B = 2$, $h/R = 0.05$, $\lambda_T = 15.5$

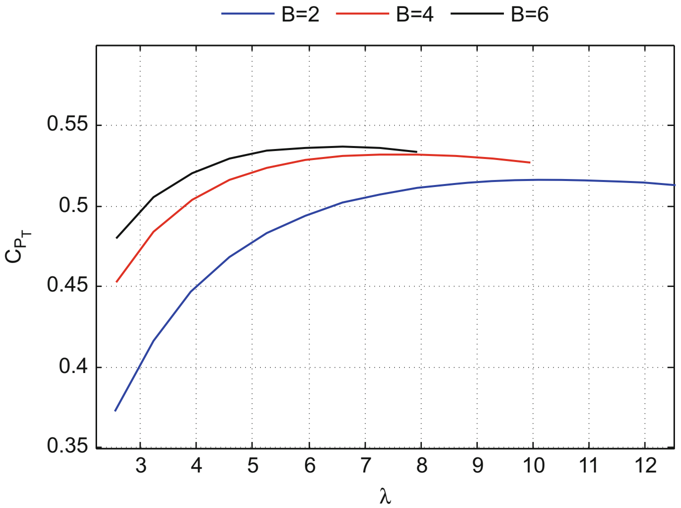


Fig. 24 Design curves of the box blade for $h/R = 0.05$ and various number of blades

the value of δ and keeping constant the distance h along the wind axis between the two planes normal to it and containing both front and rear blades. As we can observe there is a value of δ that we have the absolute maximum value for the C_{P_T}. The increase in the power coefficient due to this configuration is very small, but this analysis proves that the box blade lying on the x–z plane is not the best blade possible.

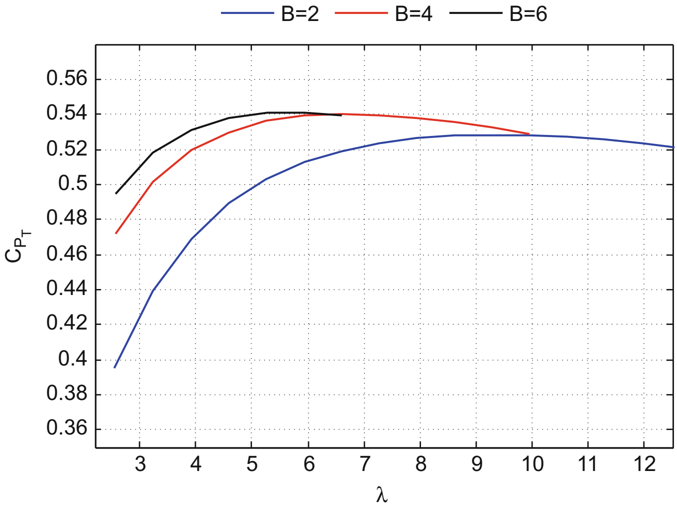
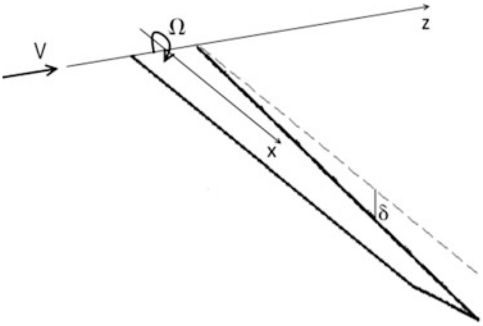


Fig. 25 Design curves of the box blade for $h/R = 0.1$ and various number of blades

Table 2 Total power coefficients and corresponding values of λ and λ_T for the maximum points of the box blades design curves, varying B and h/R.

h/R	0.05			0.1		
B	2	4	6	2	4	6
$C_{P_T max}$	51.57	53.19	53.62	52.73	53.94	54.04
λ_T	15.5	11.5	9.5	14	10	8.5
λ	10.28	7.61	6.27	9.28	6.6	5.6

Fig. 26 Configuration of box blade with relative angle of inclination of δ between the radial blades



Lastly, the behaviour of the inviscid C_P for classical, winglet, and box configurations is shown in Fig. 28. The numerical result is that the C_P tends to the Betz limit for non-planar geometries too; however, this hypothesis needs an analytical proof.

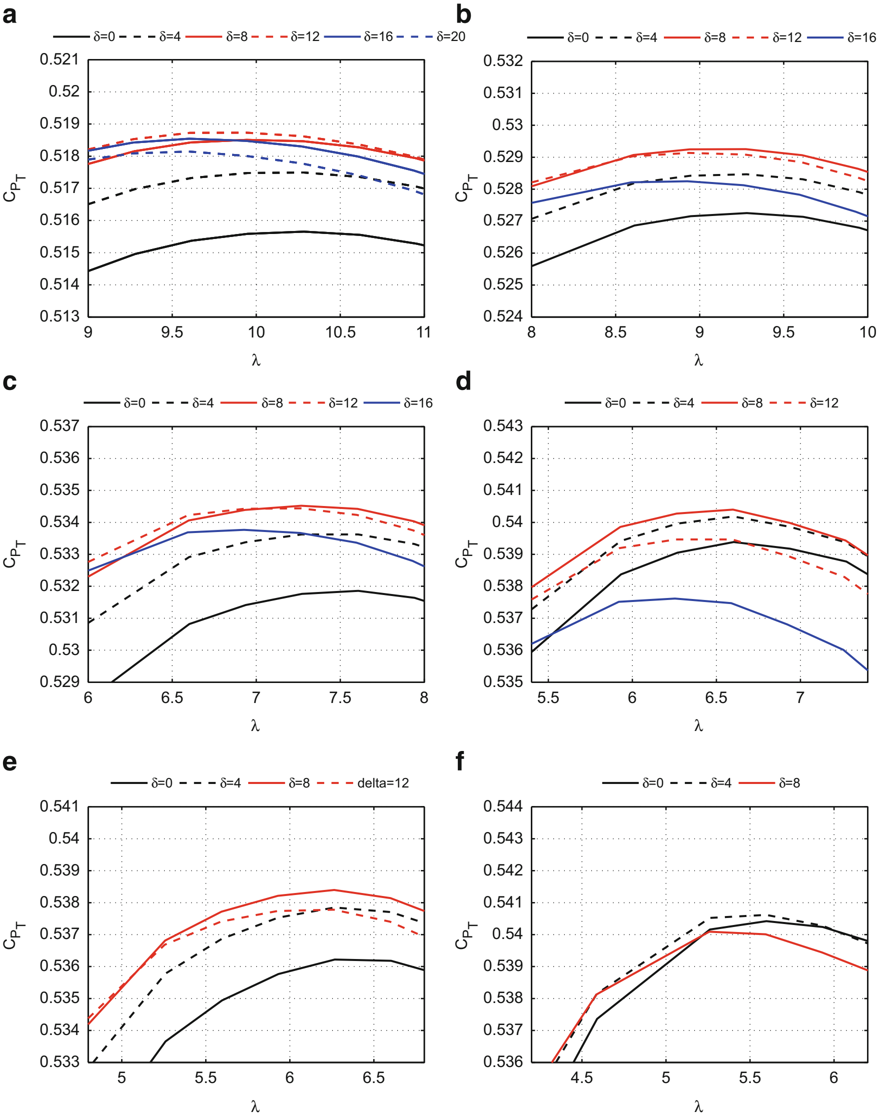


Fig. 27 Results for box blade with inclination. (**a**) $B = 2$, $h/R = 0.05$. (**b**) $B = 2$, $h/R = 0.1$. (**c**) $B = 4$, $h/R = 0.05$. (**d**) $B = 4$, $h/R = 0.1$. (**e**) $B = 6$, $h/R = 0.05$. (**f**) $B = 6$, $h/R = 0.1$

10 Performance Evaluation in the Rotor Plane

In this section the results obtained up to now are verified through a model known in the literature [1, 3, 5, 12] as the *Lagrange multiplier method*. This method is known since the fifties and was employed to design propellers and wings. But its interesting

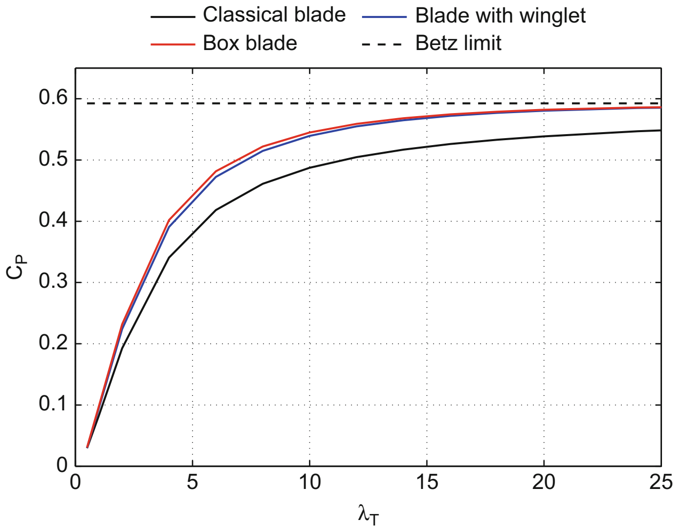


Fig. 28 Comparison among the inviscid C_P calculated for classical, wingletted, and box blade

aspect is that it can be used to analyse non-planar geometries as well. The analysis is conducted directly in the rotor plane and the Betz condition is no longer used; instead an object function related to the rotor torque is defined and it is maximized with respect to the blade circulation. For the case of a wing the object function is the induced drag which is optimized with respect to the wing circulation. Another main advantage of this approach is that the functional to be optimized can include one or more constraints of different nature, such as the wing weight or the root bending moment.

Since the Betz condition is not employed anymore we expect a difference in the optimum circulation distribution for heavily loaded rotors. This model treats the non-linearity due to a high value of w, typical of heavily loaded blades. In this case the wake non-dimensional pitch is not assumed a priori, but an initial value of w is set, and thus λ_T is known, then after calculating the solution the value of w is updated and the procedure is repeated until convergence is achieved. Finally, being the analysis conduced in the rotor plane, the bound circulation contribution must be added to the aerodynamic induction coefficients calculation. In fact, in the Trefftz plane case only the trailing vortices were considered to evaluate the matrix F.

Since the problem must be solved numerically the usual blade discretization is used, with a total of N_v trailing vortices detaching from each blade. Following the notation reported in [17] we can write the expressions for the inviscid thrust and torque in a matrix form as follows:

$$T_i = B\rho \int_s \mathbf{e}_z \cdot (\mathbf{v}_0 \times \boldsymbol{\tau})\, \Gamma\, ds + B\rho \int_s \mathbf{e}_z \cdot (\, \mathbf{Q}_k \times \boldsymbol{\tau}\,)\; \Gamma\, ds\,, \tag{44}$$

$$Q_i = B\rho \int_s \mathbf{e}_z \cdot (\mathbf{r} \times (\mathbf{v}_0 \times \boldsymbol{\tau}))\, \Gamma\, ds + B\rho \int_s \mathbf{e}_z \cdot (\mathbf{r} \times (\mathbf{Q}_k \times \boldsymbol{\tau}))\, \Gamma\, ds\ . \tag{45}$$

The induced velocity field can also be written in a matrix form through the aerodynamic induction matrices:

$$\begin{cases} \{u_x\} = [UIC]\{\Gamma\}\ , \\ \{u_y\} = [VIC]\{\Gamma\}\ , \\ \{u_z\} = [WIC]\{\Gamma\}\ , \end{cases} \tag{46}$$

where Γ is the blade circulation vector. By defining the vector:

$$\mathbf{Q}_0(i,j) = \{UIC(i,j), VIC(i,j), WIC(i,j)\}^T \tag{47}$$

such that:

$$\mathbf{v}_0(i,j) = \mathbf{Q}_0(i,j)\Gamma(j)\ . \tag{48}$$

We can rewrite the expressions for thrust and torque:

$$T_i = \{\Gamma\}^T[TIC1]\{\Gamma\} + \{TIC2\}^T\{\Gamma\}\ , \tag{49}$$

$$Q_i = \{\Gamma\}^T[QIC1]\{\Gamma\} + \{QIC2\}^T\{\Gamma\}\ . \tag{50}$$

The matrices which figure in the latest equations are

$$TIC1(i,j) = B\rho\, [\mathbf{e}_z \cdot (\mathbf{QIC}(i,j) \times \boldsymbol{\tau}(i))]\, \Delta s(i)\, ; \tag{51}$$

$$QIC1(i,j) = B\rho\, [\mathbf{e}_z \cdot (\mathbf{r}(i) \times (\mathbf{QIC}(i,j) \times \boldsymbol{\tau}(i)))]\, \Delta s(i)\, ; \tag{52}$$

$$TIC2(i) = B\rho\, [\mathbf{e}_z \cdot (\mathbf{Q_k}(i) \times \boldsymbol{\tau}(i))]\, \Delta s(i)\ , \tag{53}$$

$$QIC2(i) = B\rho\, [\mathbf{e}_z \cdot (\mathbf{r}(i) \times (\mathbf{Q_k}(i) \times \boldsymbol{\tau}(i)))]\, \Delta s(i)\ . \tag{54}$$

The optimization problem consists on maximizing the torque with a given thrust constraint ($T = T_{ref}$). We now define the object function J:

$$J = Q_i + \lambda(T_i - T_{ref})\ . \tag{55}$$

This function is derived with the respect to the blade circulation vector, and also with respect to the Lagrange multiplier λ to meet the thrust constraint:

$$\begin{cases} \dfrac{\partial J}{\partial\{\Gamma\}} = 0 \\ \dfrac{\partial J}{\partial \lambda} = 0 \end{cases} \quad . \tag{56}$$

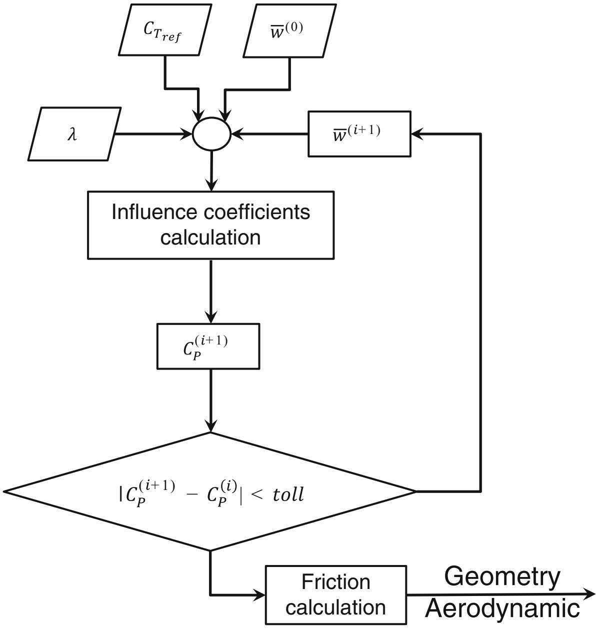


Fig. 29 Iterative cycle of performance evaluation in the rotor plane

This set of equation can be rewritten using the above-defined matrices:

$$\left[\left[\overline{QIC1}\right] + \lambda\left[\overline{TIC1}\right]\right]\{\Gamma\} + \left[\{QIC2\} + \lambda\{TIC2\}\right] = \{0\}\ , \tag{57}$$

$$\{\Gamma\}^T[TIC1]\{\Gamma\} + \{TIC2\}^T\{\Gamma\} = T_{ref}\ , \tag{58}$$

where:

$$[\overline{TIC1}] = [TIC1] + [TIC1]^T\ , \quad [\overline{QIC1}] = [QIC1] + [QIC1]^T\ . \tag{59}$$

The derivative with respect to Γ gives a set of N_v equations which, once resolved, provide the blade circulation. Then Γ must be inserted in the constraint equation to update the value of λ until convergence is reached. The flow diagram shown in Fig. 29 summarizes the numerical procedure used in this method. It must be observed that for non-planar geometries the induced velocity field tends to be singular at the junction, and this is a source of numerical instability. In fact, little variations of the mesh parameters result in quite different circulation distributions. Despite this method is more accurate (as pointed out later with the CFD analysis), this is much more computationally heavy, since the C_T and λ must also be varied to find the global optimum. The Trefftz plane model does not employ any iterative procedure for the optimum circulation and also, thanks to the analytical formula, is much faster.

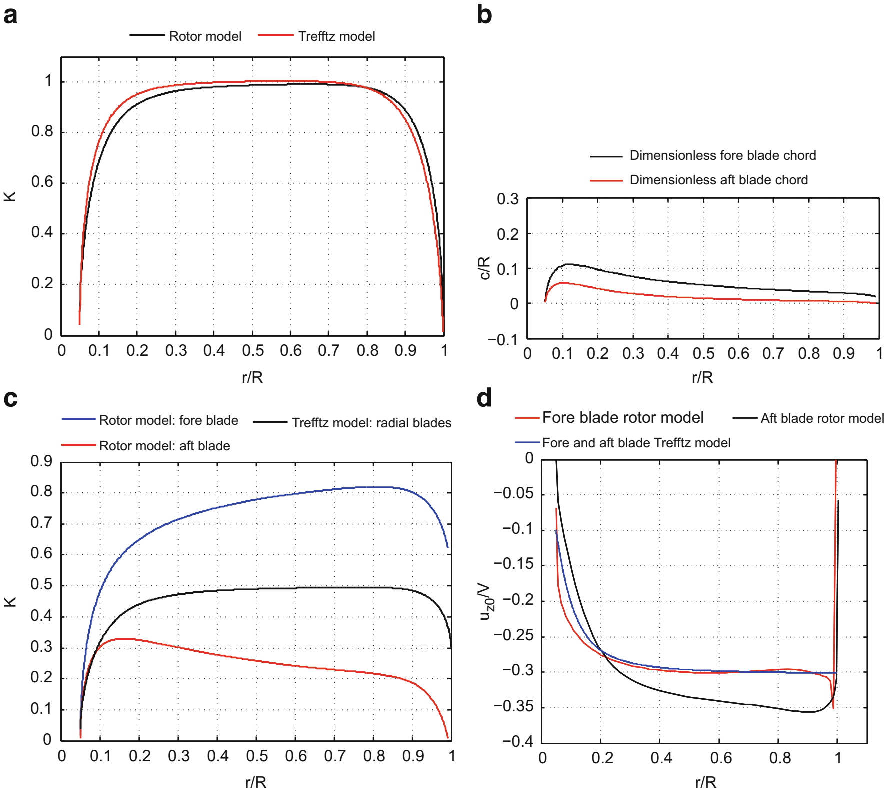


Fig. 30 Results for the rotor model. (**a**) Classical blade. (**b**) Chord distribution. (**c**) Box blade. (**d**) Axial induced velocity

Table 3 Comparison between the numerical models for the classical blade; $B = 2$

Numerical code	C_P	C_{P_T}
Trefftz plane model	54.03	49.67
Rotor plane model	53.26	49.1

A comparison between the two models is presented in Fig. 30, Tables 3 and 4: no significant differences for the C_{P_T} can be observed; however, for the case of box blade turbine the optimum circulation distribution is now quite asymmetrical, since now the induced velocity field is no longer symmetrical.

Table 4 Comparison between the numerical models for the box blade; $B = 2$, $h/R = 0.1$, $\delta = 0$

Numerical code	C_P	C_{P_T}
Trefftz plane model	56.81	52.73
Rotor plane model	57.42	52.83

11 CFD Validation

Once the design process is over, the blade geometry is fully defined by the twist, chord, and position distributions of each section. These vectors are then exported into ASD code to generate the CAD of the blades. ASD is a software developed to quickly draw aeroplane shapes by employing an N.U.R.B.S. representation. In particular, it has features to generate wings and fuselages, and allows also to generate fillets between the parts. In addition, there are specific commands for PrandtlPlane aircraft; in fact, the two main wings can be generated and there is a special feature for the side wings, i.e., bulk option. ASD can be successfully used to generate blades geometries as well.

Figure 31 shows a classical blade, a wingletted blade, and a box blade drawn with ASD, respectively.

After the geometry has been generated it is imported in CATIA® for the fillets finish and the hub generation. Once the CAD is completed it is then imported in ANSA®, a commercial software for surface meshing operations. The volume mesh generation is done with ANSYS Fluent®.

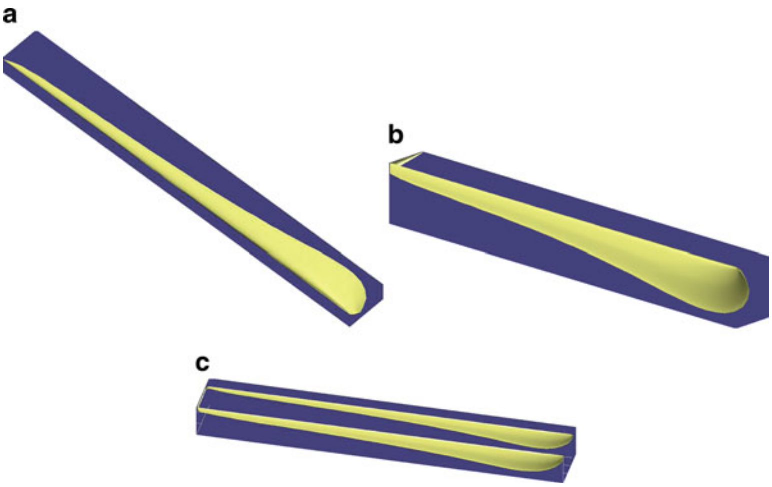


Fig. 31 Blades designed with ASD. (**a**) Classical blade. (**b**) Blade with winglet. (**c**) Box blade

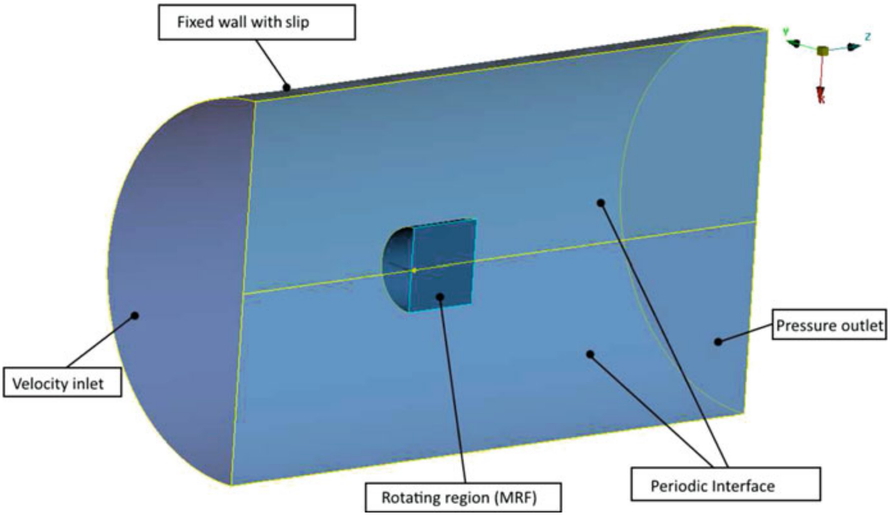


Fig. 32 Computational domain with the assigned boundary conditions

Moreover only the rotor is modeled, and so the problem can be solved as steady state if a suitable non-inertial reference frame is defined which rotates with the turbine. ANSYS Fluent® allows to do so with the option of Moving Reference Frame. For this purpose an internal cylinder which encloses the blade is defined, and being the problem periodically symmetrical only half of the domain is modeled, since we are dealing with two bladed rotors. In Fig. 32 we can see the computational domain with the boundary conditions which were assigned for the CFD analysis. For the sake of comparison we do not need a very fine mesh, so we choose a mesh which is a fair compromise between accuracy and fast computational time; thus, a value of $30 \leq Y^{+} \leq 300$ is chosen and the turbulence models used are the $k - \varepsilon$ and the $k - \omega$ Standard. The blockage factor is set at $\varphi = D_{turb}/D_{ext} = 0.2$ and the domain extends for 6 radius upstream and 12 radius downstream the turbine. The volumetric growth rate in the internal domain is set to 1.2 and the number of chordwise elements is 50. These parameters result in a volume mesh of about 7.2 million of cells for the classical blade and about 14.4 million of cells for the box blade.

Before proceeding with the solution of the model a quality check is performed, in particular the two following criteria have to be met:

- maximum allowable skewness is 0.85;
- number of cells exceeding the skewness of 0.5 must be less than the 15 % of the total cells.

The graph in Fig. 33 shows the mesh quality; it can be seen how the above-mentioned criteria are met.

The model requires about 3000 iterations to reach convergence, as we can see from Fig. 34.

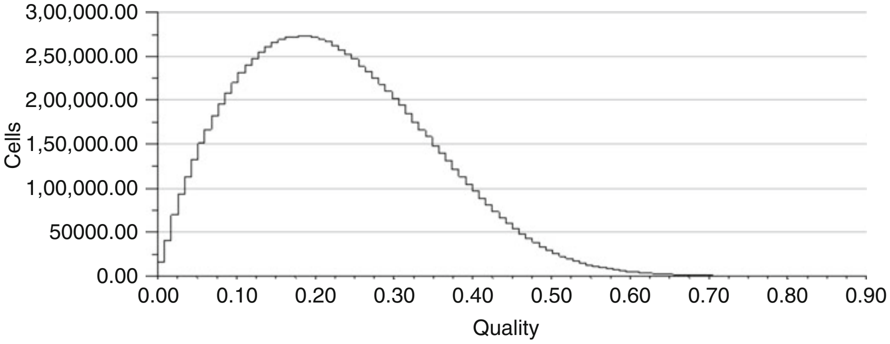


Fig. 33 Skewness cells distribution

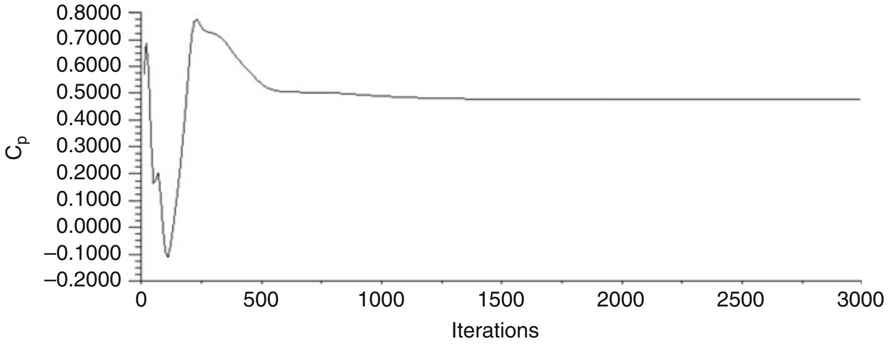


Fig. 34 Convergence of the C_{P_T} with respect to the number of iterations computed by Fluent

Table 5 Comparison of the box blade CFD analysis for different turbulence models

Turbulence model	$C_{P_T}(\%)$ classical	$C_{P_T}(\%)$ box	$\Delta C_{P_T}(\%)(\%)$	$\Delta C_{P_T}/C_{P_T}\text{classical}(\%)$
$k-\epsilon$ Realizable	46.2	49.82	3.62	7.84
$k-\epsilon$ RNG	46.69	49.63	2.94	6.3
$k-\omega$ Standard	46.47	48.82	2.35	5.06

First of all the simulation was conduced in the design point of the numerical codes, and then the tip speed ratio λ is changed to compare the maximums of the $C_{P_T}-\lambda$ curves of the classical and the box blades. Then, the influence of the selected turbulence model is analysed and the results are summarized in Table 5.

In the following tables the rotors power coefficients are reported.

An analysis of the geometry derived by means of the numerical code in the Trefftz plane is carried out for different values of λ and the results are reported in Table 6 and in Fig. 35.

Also the blades designed with the rotor plane model were modeled for the CFD analysis. We can see how this design further increases the power extracted by the wind. But the more interesting aspect is that the lift subdivision measured on Fluent was fully in agreement with the values forecast by the numerical code.

Table 6 Total power coefficients and corresponding values of λ nearby the maximum point of the classical and the box blade power coefficient curves

	Classical blade			Box blade		
λ	11.95	12.95	13.95	9.28	10.28	11.28
$C_{P_T}(\%)$	46.2	46.75	45.7	49.82	49.9	48.15

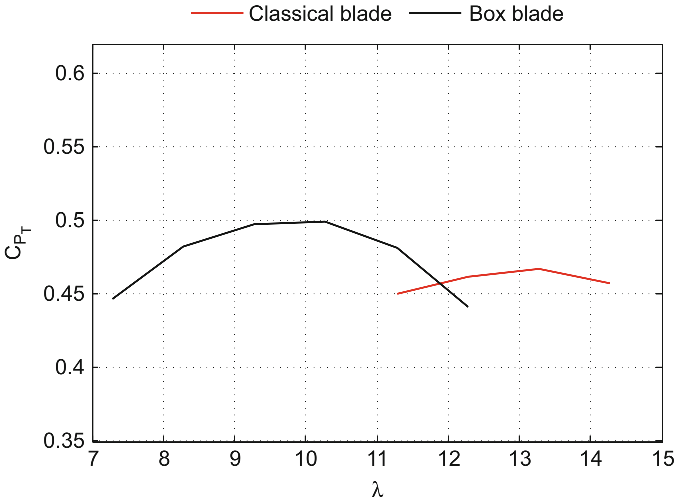


Fig. 35 CFD results of the total power coefficient close to the maximum point of the classical and the box blade power coefficient curves; turbulence model: $k - \epsilon$ Realizable

Table 7 Increments of the box blade performance with respect to the classical blade evaluated with CFD; TURBULENCE model: $k - \epsilon$ Realizable; $\lambda_{\text{classical}} = 12.28$, $\lambda_{\text{box}} = 9.28$ (design point from the vortex model codes)

	$C_{P_T}(\%)$	$\Delta C_{P_T}(\%)$	$\Delta C_{P_T}/C_{P_T}\text{classical}(\%)$
Model in the Trefftz plane (corrected)	50.3	4.1	8.87
Model in the rotor plane	50.82	4.62	10

On the contrary for the Trefftz plane model the fore blade produces more lift (about 3 % more than the aft blade), because of the approximation of equal axial induced velocity over the two radial blades. Actually the downwash is stronger for the aft blade thus leading to a smaller incidence of the airfoils compared to the values of the code. However, the error is quite moderate, thus enabling us to retain the approximation of the Trefftz plane model fairly reasonable. Anyway it can be observed that an increase of 1° of the aft blade pitch angle leads to a higher load acting on this blade and to a higher value of the C_{P_T}, despite to a little decrease of the lift developed by the fore blade, and makes the load symmetrical again.

The results from the CFD of the two cases just mentioned, calculated for the values of λ corresponding to the maximum point of the vortex model codes, are summarized in Table 7.

12 Conclusions

We can state the following conclusions:

- CFD confirmed that the box blade is more aerodynamically efficient;
- Some turbulence models are more optimistic but in general the results are quite satisfactory;
- The design of the box blade performed with the method of the Lagrange multipliers is a little more efficient;
- The approximation used in the Trefftz plane of equal axial induced velocity over the two radial blades is fairly reasonable, leading to a small error in the lift subdivision between the two blades. In fact, the lift measured by means of ANSYS Fluent® on the fore blade is only 3 % higher than the aft blade;
- The symmetry can be restored by simply increasing the aft blade twist angle by 1° for each section, resulting also in a small performance increase;
- The symmetrical design is more desirable from a technological point of view, because, since the blades are usually manufactured with a mold, this design would require the same mold for both the blades.

As said at the beginning of this paper structural issues were not taken into account; however, qualitatively we can list some of the possible advantages of this blade configuration. For example, the structure is overconstrained to the axle and each blade has only the half of the load of a classical blade. The designs performed here minimize the friction losses resulting in chords for the box blade which are about the half of the classical blade; however, the blades could be designed with larger chords at the cost of a light increase in friction forces. This would result in a stiffer structure which is less sensible to deformation under loads. A natural application of this concept could be the large offshore installation, in which, for example, a double diameter of the blade allowable in theory with this technology would make the design of 40 MW rotors possible.

In addition to this possibility another important issue is that for a conventional size box blade if the manufacturing costs of this configuration do not rise consistently the energy provided by this technology could be less expensive, which means less direct operating costs compared with a traditional design.

Another important issue to be considered in the future is the control system strategy for this configuration. A likely solution is to put hinges between the main blades and the bulk in order to allow the two main blades to rotate about their axis, in a similar way of a traditional pitch control system.

Appendix: Comparison Between Theodorsen and Okulov–Sørensen Theories

In this Appendix we show the limits of the theory developed by Theodorsen for the study of heavily loaded rotors when applied to a wind turbine. As pointed out in Sect. 4.1, the Goldstein analytical solution was derived for a lightly loaded rotor. Hence, Theodorsen tried to generalize it to the case of a heavily loaded propeller. When $\bar{w} \ll 1$ it is possible to neglect the slipstream contraction/expansion. The first case occurs when dealing with a propeller while the latter occurs when considering a wind turbine. The wake deformation is another source of non-linearity for the problem since the wake expansion ratio is not known a priori and it depends on the solution itself. Following the Theodorsen approach we need to derive some additional relationships to relate the ratio R_∞/R of the wake radius far downstream over the rotor radius.

First of all the performance coefficients are evaluated in the Trefftz plane, by means of Goldstein circulation function. To this end the subsequent integral balances are carried out on the control volume shown in Fig. 36:

- mass balance;
- z-momentum balance;
- energy balance.

For the detailed calculations we refer to [21]. Here we just report the main results for the power and thrust coefficients:

$$T = \rho w A k \left[V - w \left(\frac{1}{2} + \frac{\varepsilon}{k} \right) \right] \Rightarrow C_T = 2k\bar{w} \left[1 - \bar{w} \left(\frac{1}{2} + \frac{\varepsilon}{k} \right) \right] , \tag{60}$$

$$P = \rho k w A \left(V - w \right) \left(V - w \frac{\varepsilon}{k} \right) \Rightarrow C_P = 2k\bar{w} \left(1 - \bar{w} \right) \left(1 - \bar{w} \frac{\varepsilon}{k} \right) , \tag{61}$$

where ε is the mass coefficient and k is the kinetic energy coefficient, both dependent on the dimensionless circulation and the wake pitch.

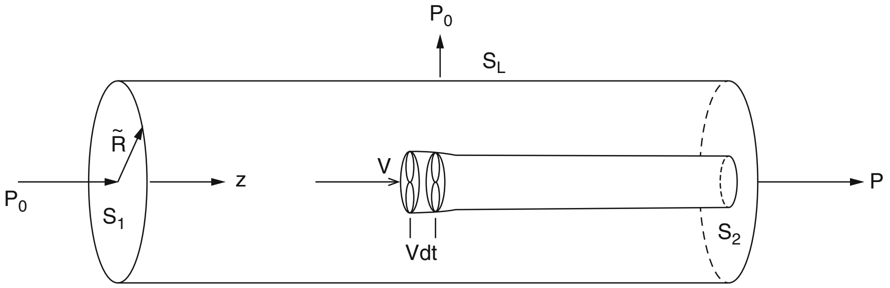


Fig. 36 Control volume for the integral balances of the Theodorsen theoretical model; figure taken from [21]

The dimensionless coefficients of power and thrust have been evaluated by dividing thrust and power by $\frac{1}{2}\rho\pi R_\infty^2 V^2$ and $\frac{1}{2}\rho\pi R_\infty^2 V^3$, respectively, where the radius R_∞ of the streamtube far downstream the rotor plane is chosen as reference. In this way we obtain the performance of the turbine if no wake expansion were present. To include the wake expansion we must divide by the rotor disk area:

$$C_{P_r} = C_P \left(\frac{R_\infty}{R}\right)^2 , \tag{62}$$

$$C_{T_r} = C_T \left(\frac{R_\infty}{R}\right)^2 . \tag{63}$$

The additional equations that are needed to evaluate R_∞/R are derived by calculating thrust and torque by means of the blade element momentum theory. Let us consider the velocity triangle in the rotor plane. The magnitude of the relative velocity is the following:

$$u = V\frac{\left(1 - \bar{a_0}\cos^2\vartheta_p\right)^2}{\sin\vartheta_p} , \tag{64}$$

where $\bar{a}_0$ and ϑ_p are the dimensionless wake advance velocity and the flow angle in the rotor plane, respectively. Now, with the Kutta–Joukowski theorem we can write

$$T = \rho B\omega \int_0^R \frac{1 - \bar{a_0}\cos^2\vartheta_p}{1 - \bar{a_0}}\Gamma(r)\,dr , \tag{65}$$

whose dimensionless form is

$$\left(\frac{R_\infty}{R}\right)^2 C_T = \frac{4\bar{w}\,(1 - \bar{w})}{1 - \bar{a_0}} \int_0^1 \left(1 - \bar{a_0}\cos^2\vartheta_p\right) K(x)x dx . \tag{66}$$

Similarly the power coefficient is

$$\left(\frac{R_\infty}{R}\right)^2 C_P = 4\bar{w}\,(1 - \bar{w}) \int_0^1 \left(1 - \bar{a_0}\cos^2\vartheta_p\right) K(x)x dx . \tag{67}$$

By dividing member by member these equations we find the relationship between $\bar{a}_0$ and $\bar{w}$:

$$\bar{a}_0 = \frac{\frac{1}{2}\bar{w} - \frac{\varepsilon}{k}\bar{w}^2}{1 - \bar{w}\left(\frac{1}{2} + \frac{\varepsilon}{k}\right)} . \tag{68}$$

By substituting Eq. (60) into Eq. (66) we can finally write the wake expansion:

$$\left(\frac{R_\infty}{R}\right)^2 = \frac{(1-\bar{w})\,(1-\bar{a}_0 S)}{(1-\bar{a}_0)\left[1-\bar{w}\left(\frac{1}{2}+\frac{\varepsilon}{k}\right)\right]}\ , \tag{69}$$

in which:

$$S = \frac{2}{k}\int_0^1 Kx\cos^2\vartheta_p\,dx = \frac{2}{k}\int_0^1 K\frac{x^3}{x^2+\left(\frac{1-\bar{a}_0}{1-\bar{w}}\right)^2\left(\frac{R_\infty}{R}\right)^2\frac{1}{\lambda_T^2}}\,dx\ . \tag{70}$$

To fully define the problem of the wind turbine blade design, we need to maximize the power coefficient with respect to $\bar{w}$:

$$\frac{d}{d\bar{w}}\left(C_P\,(\bar{w})\left(\frac{R_\infty}{R}\,(\bar{w})\right)^2\right) = 0\ , \tag{71}$$

in which it is necessary to show the explicit dependence of the wake expansion ratio from $\bar{w}$: this requires an iteration procedure, since the equations that relate R_∞/R to S are not explicit. The numerical solution was done as follows: since S and $\bar{w}$ are comprised between 0 and 1, the vectors $\bar{S}_j$ and $\bar{w}_k$ are defined to discretize these quantities; then the function $\bar{a}_0(\bar{w}_k)$ is derived in a discrete form by employing Eq. (68). Then the discrete function $S(\bar{w}_k)$ is derived: for each element of the S vector the following error is evaluated:

$$\Delta = S_j - \sum_{i=1}^{N_v}\Gamma_i\frac{x_i^3\,\Delta x}{x_i^2+\dfrac{\left(1-\bar{w}_k\frac{\varepsilon}{k}\right)\left(1-\bar{a}_{0k}S_j\right)}{\left[1-\bar{w}_k\left(\frac{1}{2}+\frac{\varepsilon}{k}\right)\right]\lambda_T^2}}\ . \tag{72}$$

The value of the function $S(\bar{w}_k)$ is given by the value of the vector S_j which minimizes the difference in Eq. (72). At this point the wake expansion $(\frac{R_\infty}{R}(\bar{w}_k))^2$ can be derived and, thus, the power coefficient is known with respect to $\bar{w}_k$. Finally the maximum value of C_{P_T} can be found.

Once we know the value of $\bar{w}$ which maximizes the power coefficient, the rotor tip speed ratio can be evaluated as well:

$$\lambda = (1-\bar{w})\lambda_T\ . \tag{73}$$

In the present paper it is observed how this model provides unreasonable results for high values of λ_T. In fact when $\lambda_T \to \infty$ the helix angle approaches $\pi/2$, so the wake is similar to a solenoid; thus the induced velocity is purely axial and so $v_z \equiv w$, therefore the mass coefficient and the kinetic energy coefficient become

$$k \to 1\ , \qquad \varepsilon \to 1\ , \qquad S \to 1\ , \tag{74}$$

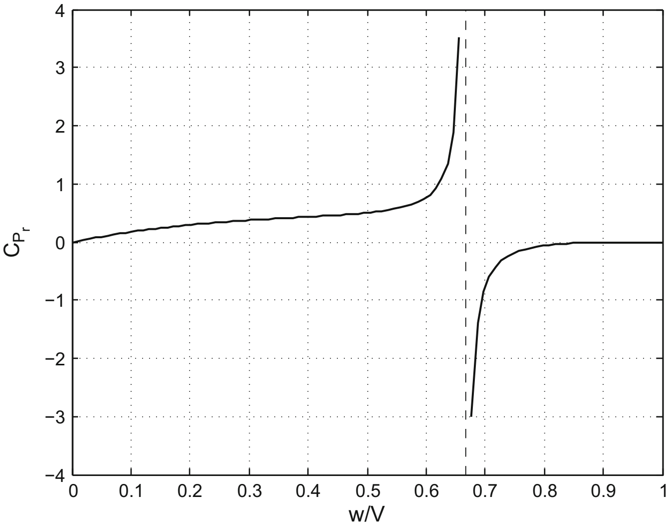


Fig. 37 Divergence of the total power coefficient when $\bar{w}$ approaches $2/3$

thus the power coefficient becomes

$$C_{P_r} = 2\bar{w}\frac{(1-\bar{w})^3}{1-\frac{3}{2}\bar{w}} \ . \tag{75}$$

Clearly this function does not have any maximum but it diverges when $\bar{w}$ approaches $2/3$, as we can see in Fig. 37. A likely explanation of this inconsistency provided in this paper is the following: in the momentum balance we have a term which is related to the wake overpressure with respect to the ambient static pressure $(p - p_0)$. This additional force is not present in the general momentum theory, so we expect a different result between the two models in the case of $\lambda_T \rightarrow \infty$. Hence the C_P does not approach the Betz limit, as shown in Fig. 38. This wake overpressure increases the power extracted by the wind, and together with the fact that the wake cross section increases when λ_T increases, the extracted power tends to diverge. The error due to the presence of this overpressure was underlined by Schouten [15, 16] for the propellers case. However, in this case the wake contracts and so this term does not produce significant errors, while in the case of a wind turbine the equations of Theodorsen tends to be singular. We can state that the fixed pitch rigidly moving helicoidal model is fairly reasonable for the blade loading calculation, but for the performance calculations it needs some corrections. For example, the Okulov and Sørensen model neglects the wake expansion and thus no additional expression is required to evaluate the power coefficient. In fact they derive the performance of the rotor by applying the blade element momentum formulas directly on the rotor plane, without the integral balances that include the

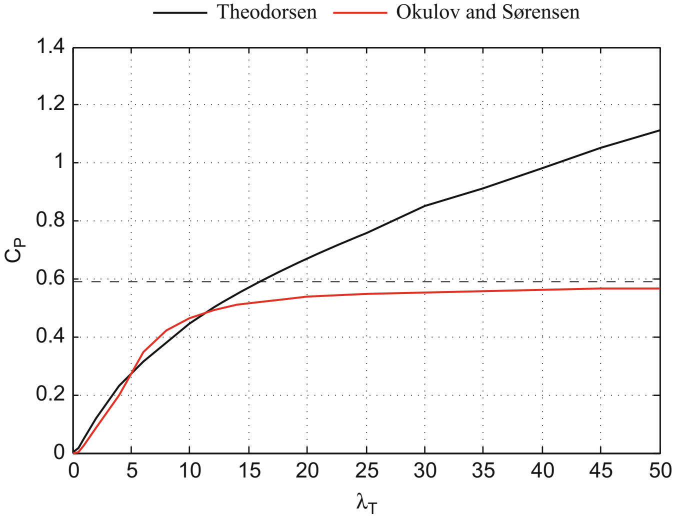


Fig. 38 $C_P - \lambda$ curves: the Theodorsen theory does not approach the Betz limit, unlike the Okulov and Sørensen does

singular term of the overpressure. To relate the rotor induced velocities with the correspondent quantities in the Trefftz plane they simply show how these velocities are the half of the velocities far downstream. The result is that this model is fully consistent with the general momentum theory when the limit case of $\lambda_T \rightarrow \infty$ is analysed.

References

1. Chattot, J.J.: Optimization of wind turbines using helicoidal vortex model. J. Sol. Energy Eng. (2003). doi:10.1115/1.1621675
2. Chattot, J.J.: Effects of blade tip modifications on wind turbine performance using vortex model. Comput. Fluids (2008). doi:10.1016/j.compfluid.2008.01.022
3. Chattot, J.J.: Wind turbine aerodynamics: analysis and design. Int. J. Aerodyn. **1**(3–4) (2011)
4. Cipolla, V., Giuffrida, S.: Utilizzo di codici a pannelli nel progetto preliminare di velivoli PrandtlPlane ultraleggeri; applicazione a nuove configurazioni. Master's Thesis, University of Pisa (2006)
5. Demasi, L., Dipace, A., Monegato, G., Cavallaro, R.: An invariant formulation for the minimum induced drag conditions of non-planar wing systems. National Harbor, Maryland (2014)
6. Foster, S.P., Ribner, H.S.: Ideal efficiency of propellers based on Theodorsen theory. Institute for Aerospace Studies, University of Toronto (1990)
7. Frediani, A., Montanari, G.: Best wing system: an exact solution of the Prandtl's problem. Springer Optimization and Its Applications, vol. 33, pp. 183–211. Springer, New York (2009)
8. Froude, R.E.: On the part played in propulsion by difference in pressure. Transaction of the Institute of Naval Architects, pp. 390–423 (1889)

9. Gaunaa, M., Johansen, J.: Determination of maximum aerodynamic efficiency of wind turbine rotors with winglets. J. Phys. Conf. Ser. (2007). doi:10.1088/1742-6596/75/1/012006
10. Glauert, H.: Airplane propellers. In: Durand, W.F. (ed.) Aerodynamic Theory, vol. 4, pp. 169–360. Springer, Berlin (1935)
11. Goldstein, S.: On the vortex theory of screw propellers. Proc. R. Soc. Lond. **123**, 440–465 (1929)
12. Kroo, I.: Design and analysis of optimally-loaded lifting systems. University of Stanford (1984). http://aero.stanford.edu/Reports/MultOp/multop.html
13. Okulov, V.L.: On the stability of multiple helical vortices. Institute of Thermophysics, Novosibirsk (2004)
14. Rankine, W.J.M.: On the mechanical principles of the action of propellers. Transaction of the Institute of Naval Architects, pp. 13–39 (1865)
15. Schouten, G.: Static pressure in the slipstream of a propeller. J Aircr. **19**(3), 251–253 (1982)
16. Schouten, G.: Theodorsen's ideal propeller performance with ambient pressure in the slipstream. J. Aircr. **30**(3), 417–419 (1993)
17. Shenkar, R.: Design and optimization of planar and non-planar wind turbine blades using vortex theory. Master's Thesis, Technical University of Denmark (2010)
18. Sørensen, J.N.: General Momentum Theory for Horizontal Axis Wind Turbines. Springer, Cham (2015)
19. Sørensen, J.N., Okulov, V.L.: An ideal wind turbine with a finite number of blades. Dokl. Phys. **53**(6), 337–342 (2008)
20. Sørensen, J.N., Okulov, V.L.: Refined Betz limit for rotors with a finite number of blades. Wind Energy **11**(4), 415–426 (2008)
21. Theodore, T.: The Theory of Propellers. Mc-Graw Hill, New York (1948)
22. Yang, K.Y.-L.: Helicopter rotor lift distributions for minimum induced power loss. Master's Thesis, Institute of Technology, Massachusetts, Chap. 3, pp. 42–52 (1993)

A New Paradigm for the Optimum Design of Variable Angle Tow Laminates

Marco Montemurro and Anita Catapano

Abstract In this work the authors propose a new paradigm for the optimum design of variable angle tow (VAT) composites. They propose a generalisation of a multi-scale two-level (MS2L) optimisation strategy already employed to solve optimisation problems of anisotropic structures characterised by a constant stiffness distribution. In the framework of the MS2L methodology, the design problem is split into two sub-problems. At the first step of the strategy the goal is to determine the optimum distribution of the laminate stiffness properties over the structure, while the second step aims at retrieving the optimum fibres-path in each layer meeting all the requirements provided by the problem at hand. The MS2L strategy relies on: (a) the polar formalism for describing the behaviour of the VAT laminate, (b) the isogeometric surfaces for describing the spatial variation of the stiffness properties and (c) a hybrid optimisation tool (genetic- and gradient-based algorithms) to perform the solution search. The effectiveness of the MS2L strategy is proven through a numerical example on the maximisation of the first buckling factor of a VAT plate subject to both mechanical and manufacturability constraints.

1 Introduction

Anisotropic materials, such as fibre-reinforced composite materials, are extensively used in many industrial fields, thanks to their mechanical performances: high stiffness-to-weight and strength-to-weight ratios that lead to a substantial weight saving when compared to metallic alloys. In addition, the recent development of new manufacturing techniques of composite structures, e.g. automated fibre-placement (AFP) machines, allows for going beyond the classical design rules, thus leading the designer to find innovative and more efficient solutions than the classical straight fibres configurations. The use of the AFP technology brought to

M. Montemurro (✉)
Arts et Métiers ParisTech, I2M CNRS UMR 5295, F-33400 Talence, France
e-mail: marco.montemurro@ensam.eu; marco.montemurro@u-bordeaux1.fr

A. Catapano
Bordeaux INP, Université de Bordeaux, I2M CNRS UMR 5295, F-33400 Talence, France
e-mail: anita.catapano@bordeaux-inp.fr

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_14

the emergence of a new class of composite materials: the variable angle tow (VAT) composites, [10, 12]. A modern AFP machine allows the fibre (i.e. the tow) to be placed along a curvilinear path within the constitutive lamina thus implying a point-wise variation of the material properties (stiffness, strength, etc.). Of course, this technology enables the designer to take advantage of the directional properties of composites in the most effective way. The interest of using variable stiffness (VS) laminates is considerably increased during the last years: in the meantime some works on the a posteriori characterisation of the elastic response of such materials have gained a lot of attention from the scientific community of composites materials. For example, [7] deals with the problem of predicting the impact and compression after impact behaviour of VAT laminates while [32] analyses the pre-buckling and buckling mechanisms in VAT laminated plates through a proper evaluation of the non-uniform stress variation within the structure due to the variable stiffness distribution. Although the utilisation of VAT laminates considerably increases the complexity of the design process (mainly due to the large number of design variables involved within the problem), on the other hand, it leads the designer to conceive non-conventional solutions characterised by either a considerable weight saving or enhanced mechanical properties when compared to classical solutions [25–28]. One of the first works that tried to explore the advantages that can be achieved in terms of mechanical performances (stiffness, buckling behaviour, etc.) by using a VS plate in which each ply is characterised by a curvilinear path of the tow (i.e. a VAT configuration) instead of the conventional straight-line fibre format is presented in [12]. The authors make use of a sensitivity analysis and a gradient-based search technique to determine the optimal fibre orientation in a given number of regions of the plate. This work proved that a considerable increment of the buckling load of the structure can be obtained when employing a VAT solution for the layered plate.

The complexity of the design process of a VAT laminated structure is mainly due to two intrinsic properties of VAT composites, i.e. the heterogeneity and the anisotropy that intervene at different scales of the problem and that vary point-wise over the structure. Moreover, a further difficulty is due to the fact that the problem of (optimally) designing a VAT laminate is intrinsically a multi-scale design problem. Indeed, in order to formulate the problem of designing a VAT composite in the most general way, the designer should take into account, within the same design process, the full set of design variables (geometrical and material) governing the behaviour of the structure at each characteristic scale (micro-meso-macro). Up to now no general rules and methods exist for the optimum design of VAT laminates. Only few works on this topic can be found in literature, and all of them always make use of some simplifying hypotheses and rules to get a solution. An exhaustive review focusing on constant and variable stiffness design of composite laminates is presented in [8, 9]. In [1] the first natural frequency of VS composite panels is maximised by considering, on the one hand, the lamination parameters and the classical laminate theory (CLT) for the description of the local stiffness properties of the structure and, on the other hand, a generalised reciprocal approximation algorithm for the resolution of the optimisation problem. This approach is limited to the determination of the stiffness properties of an equivalent homogeneous plate, since the lay-up design phase is not at all considered. In [29] the least-weight design problem of VAT laminates submitted to constraints including the strength and the radius of curvature is considered. The design variables are the layers thickness and fibres angles which are represented by bi-cubic Bezier surfaces and cubic Bezier curves, respectively. A sequential quadratic programming method is used to solve the optimisation problem. A two-level strategy was employed in [34] to design a VAT laminated plate: this work represents the first attempt of applying a multi-scale numerical strategy which aims at determining, at the first level, the optimum local (i.e. point-wise) distribution of the stiffness properties of the structure (in terms of the lamination parameters of the laminate), while at the second level the optimum path (in each constitutive layer) matching locally the lamination parameters resulting from the first step. However, the major drawback of this work actually was in the determination of the curvilinear fibres-path of each layer: the resulting path was discontinuous because the authors had not foreseen a numerical strategy able to simultaneously meet, on the one hand, the continuity of the fibres-path (between adjacent elements) and, on the other hand, the optimum distribution of lamination parameters provided by the first step of the procedure. A further work on the same topic can be found in [35] where the problem of designing variable stiffness composite panels for maximum buckling load is addressed by making use of the generalised reciprocal approximation approach introduced by Abdalla [1]. In [35] the two-level approach was abandoned and the authors stated the problem by directly considering the fibres path in each ply as design variables. However, as in [1], this approach always leads to a discontinuous fibres path and, unlike the strategy proposed in [1], it leads also to the emergence of a new issue: the resulting optimisation problem was highly non-convex since it was formulated directly in the space of the layer orientations (which vary locally over the plate). Accordingly, in [35] the authors conclude that such an issue can be potentially remedied by formulating in a proper way the design problem of VAT laminates in the framework of the two-level strategy and by trying to overcome the issue of the continuity of the fibres-path directly in the first level of the strategy where the design variables are the laminate mechanical properties (e.g. the lamination parameters as in theoretical framework of [34, 35]).

Another issue often addressed by researches on VAT laminates concerns the tow placement technology which could introduce several differences (i.e. imperfections) between the numerical model of the VAT composite and the real structure tailored with the AFP process, if the design methodology does not take into account the manufacturability requirements. To this purpose in [3] an issue linked to the AFP technology is addressed: the overlap of tow-placed courses that increases the ply thickness (the build-up phenomenon) thus affecting the structural response and the surface quality of the laminate. The work of Blom et al. [3] presents a method for designing composite plies with varying fibre angles. The fibre angle distribution per ply is given while, using a streamline analogy, the optimal distributions of fibre courses is determined for minimising the maximum ply thickness or maximising the surface smoothness. An improved research on this topic has been developed in [30] where an algorithm is presented to optimise the fibres path in order to ensure manufacturability. A further work focusing on the development and/or improvement of manufacturing techniques for tailoring VAT laminates in order to minimise the imperfections induced by the fabrication process is presented in [13]. The continuous tow shearing (CTS) technique, utilising the ability to shear dry tows, is proposed as an alternative technique to the well-known AFP process. Later, the work presented in [13] has been improved through the introduction of a computer-aided modelling tool [14] which can create accurate finite element models reflecting the fibre trajectories and thickness variations of VAT composites manufactured using the CTS technique.

As a summary of this non-exhaustive review on VAT composites it can be stated that the main limitations and drawbacks characterising the vast majority of the studies on these materials are:

- a discontinuous distribution of the mechanical parameters (e.g. lamination parameters) describing the elastic response of the laminate over the structure;
- a discontinuous distribution of the local fibres orientation angle within each ply;
- the use of linear/quadratic functions for representing the fibre path (which significantly shrinks the design domain);
- the lack of a proper and efficient multi-scale approach for dealing with the (optimal) design problem of VAT laminates;
- the absence of practical rules for taking into account the manufacturability/ technological constraints since the early stages of the design process;
- the applications which are limited only to ‘academic’ cases and not extended to real-world engineering problems.

To overcome the previous restrictions the present work focuses mainly on the generalisation and extension of the multi-scale bi-level (MS2L) procedure for the optimum design of composite structures (initially introduced in [18, 19]) to the case of VAT composites. The idea of a bi-level (or multi-level) procedure for designing composite structures is not entirely new and has already been used in the past [11]. Up to now this strategy has been employed only by few authors for the optimisation of composite structures but in each study the link between the levels of the procedure and the scales of the problem was never rigorously stated.

The authors and their co-workers already made use of the MS2L procedure for the design and optimisation of several classes of hybrid anisotropic structures in the past [4–6, 16–19, 24]. The MS2L design strategy employed in the previous works is a very general methodology for designing composites structures: it is characterised, on the one hand, by the refusal of the simplifying hypotheses and classical rules usually employed in the framework of the design process of laminates and, on the other hand, by a proper and complete mathematical formalisation of the optimum design problem at each characteristic scale (micro-meso-macro). The MS2L strategy relies on the use of the polar formalism (initially introduced by Verchery [39], and later extended to the case of higher-order theories [21–23]) for the description of the anisotropic behaviour of the composite. The real advantage in using the Verchery’s polar method within the design process of composite structures is in the fact that the elastic response of the structure at the macro-scale is described in terms of tensor invariants, the so-called *polar parameters* having a precise physical meaning (which is linked to the elastic symmetries of the material) [37]. On the other hand, the MS2L strategy relies on the use of a particular genetic algorithm (GA) able to deal with a special class of huge-size optimisation problems (from hundreds to thousands of design variables) defined over a domain of variable dimension, i.e. optimisation problems involving a 'variable number' of design variables [16].

As far as concerns the problem of designing VAT composites, the aim of this paper is twofold. On the one hand, a new paradigm for designing VAT laminates is introduced, while, on the other hand, the MS2L optimisation strategy has been generalised in order to deal with the design problem of VAT composites. Several modifications have been introduced in the theoretical and numerical framework of the MS2L design procedure at both first and second levels. At the first level (laminate macroscopic scale) of the procedure, where the VAT laminate is modelled as an equivalent homogeneous anisotropic plate whose mechanical behaviour is described in terms of polar parameters (which vary locally over the structure) the major modifications are: (1) the utilisation of higher-order theories (first-order shear deformation theory (FSDT) framework [21, 22]) for taking into account the influence of the transverse shear stiffness on the overall mechanical response of VAT composites; (2) the utilisation of B-spline surfaces for obtaining a continuous point-wise variation of the laminate polar parameters. Regarding the second-level problem (laminate mesoscopic scale, i.e. the ply level) the main modifications are: (1) the utilisation of B-spline surfaces for obtaining a continuous point-wise variation of the fibre orientation angle within each ply; (2) a proper mathematical formalisation of the manufacturability constraints linked to the AFP process in the framework of the B-spline representation. All of these modifications imply several advantages for the resolution of the related optimisation problems (both at first and second level of the strategy) that will be detailed in Sects. 3 and 4.

The paper is organised as follows: the design problem and the MS2L strategy are discussed in Sect. 2. The mathematical formulation of the first-level problem is detailed in Sect. 3, while the mathematical statement of the second-level problem (the lay-up design) is presented in Sect. 4. A concise description of the finite element (FE) model of the VAT layered plate is given in Sect. 5, while the numerical results of the optimisation procedure are shown in Sect. 6. Finally, Sect. 7 ends the paper with some concluding remarks.

2 A New Design Paradigm for VAT Laminates

2.1 Description of the Problem

The optimisation strategy presented in this study is applied to a VAT laminated plate composed of a fixed number of plies, hence the total thickness of the plate is fixed a priori. The fibre tow is made of carbon-epoxy pre-preg strips whose elastic properties are listed in Table 1.

Concerning the mechanical behaviour of the VAT plate, further details have to be added in order to clearly define the theoretical framework of this work:

Table 1 Material properties of the carbon-epoxy pre-preg strip, see [21, 22]

Technical constants		Polar parameters of [Q] [a]		Polar parameters of $\left[\hat{Q}\right]$ [b]	
E_1 [MPa]	161000.0	T_0 [MPa]	23793.3868	T [MPa]	5095.4545
E_2 [MPa]	9000.0	T_1 [MPa]	21917.8249	R [MPa]	1004.5454
G_{12} [MPa]	6100.0	R_0 [MPa]	17693.3868	Φ [deg]	90.0
ν_{12}	0.26	R_1 [MPa]	19072.0711		
ν_{23}	0.10	Φ_0 [deg]	0.0		
		Φ_1 [deg]	0.0		
Density and thickness					
ρ [Kg/mm^3]	1.58×10^{-6}				
h_{ply} [mm]	0.125				

[a] In-plane reduced stiffness matrix of the pre-preg strip
[b] Out-of-plane shear stiffness matrix of the pre-preg strip

- the geometry of the laminated structure and the applied boundary conditions (BCs) are known and fixed;
- the VAT plate is composed of identical plies (i.e. same material and thickness);
- the material behaviour is linear elastic;
- the VAT plate is quasi-homogeneous and fully orthotropic [4, 5, 24] point-wise, i.e. these properties apply locally in each point of the structure;
- at the macro-scale (i.e. the scale of the structure) the elastic response of the VAT plate is described in the theoretical framework of the FSDT and the stiffness matrices of the plate (whose components vary point-wise over the structure) are expressed in terms of the laminate polar parameters [21, 22] which constitute also the design variables of the VAT plate at the macroscopic scale.

As far as concerns the mesoscopic scale of the VAT laminate (i.e. that of the constitutive ply) no simplifying hypotheses are made on the rest of the design parameters of the laminated plate, i.e. the design variables of the stack, namely the layer position and orientation angle (which varies point-wise for each layer). Only avoiding the utilisation of a priori assumptions that extremely shrink the solution space (e.g. the utilisation of symmetric balanced stacks to attain membrane/bending uncoupling and membrane orthotropy, respectively) one can hope to obtain the true global optimum for a given problem: this is a key-point in the proposed approach.

2.2 *Description of the Multi-Scale Two-Level Optimisation Strategy*

The main goal of the design strategy is the maximisation of the first buckling load of a VAT plate subject to

- feasibility constraints on the material parameters (i.e. the laminate polar parameters) governing the behaviour of the structure at the macroscopic scale;
- manufacturability constraints on the local radius of the tow (i.e. the local steering) due to the considered AFP technology.

The optimisation procedure is articulated into the following two distinct (but linked) optimisation problems:

1. **First-level problem**. The aim of this phase is the determination of the optimum distribution of the material properties of the VAT structure in order to minimise the considered objective function and to meet, simultaneously, the full set of optimisation constraints provided by the problem at hand. At this level the VAT plate is modelled as an equivalent homogeneous anisotropic continuum whose behaviour at the macro-scale is described in terms of laminate polar parameters, in the theoretical background of the FSDT [21, 22], which vary point-wise over the structure. Indeed the distributions of the laminate polar parameters over the laminated plate constitute the design variables of the first-level problem.
2. **Second-level problem**. The purpose of this design phase is the determination of the optimum lay-up of the laminate composing the structure (the laminate meso-scale) meeting the optimum combination of the polar parameters provided by the first level of the strategy. At this stage, the design variables are the layer orientation angles which vary point-wise in each ply (namely the fibres path) and, if needed, at this stage the designer can add some additional requirements, e.g. constraints on the elastic behaviour of the laminate, manufacturability constraints, strength and damage criteria, etc.

To the best of the authors knowledge only few research activities have been carried out on the application of the bi-level optimisation procedure to the design problem of VAT laminates [6, 29, 36]. Although these works focus only on ‘academic’ cases and benchmarks, they prove that, for a given geometry of the considered structure, the utilisation of a VAT solution allows for obtaining superior mechanical characteristics when compared to a classical multilayer solution composed of unidirectional laminae. This result is due to the elastic behaviour of VAT laminates which fit point-wise the equivalent material properties to the stress and strain fields engendered within the structure. Despite some relevant advances illustrated in [6], the bi-level approach presented in that work for dealing with the problem of the optimum design of VAT composites suffer of the following drawbacks:

- the optimum solution resulting from the first step of the procedure often consists in a discontinuous distribution of the laminate polar parameters which results in a discontinuous fibres-path (for each constitutive layer) for the second-level problem;
- the lack of practical rules and of a very general mathematical formulation for determining a proper fibres-path;
- the manufacturability constraints linked to the AFP process are not taken into account within the design process (i.e. within the problem formulation in the context of the bi-level optimisation procedure).

Accordingly, the optimum solutions illustrated in [6] cannot be manufactured. In order to overcome the difficulties listed above, some major modifications have been introduced within the mathematical formulation of the design/optimisation problem of VAT composites (for each level of the MS2L strategy), especially for taking into account within the design process the manufacturability constraints related to the AFP process. These modifications are detailed for each level of the numerical optimisation strategy in Sects. 3 and 4, respectively.

3 Mathematical Formulation of the First-Level Problem

In order to apply the MS2L numerical optimisation strategy presented in [5, 24] to the case of VAT composites some major modifications have been introduced. Regarding the first-level problem these modifications focus on:

- the utilisation of higher-order theories (in this case the FSDT framework) for taking into account the influence of the transverse shear stiffness on the overall mechanical response of the VAT laminate;
- the utilisation of B-spline surfaces for expressing the variation of the laminate polar parameters over the structure.

The first point represents a very important step forward in the MS2L strategy when applied to every kind of composite structure (classical or VAT) as it allows to properly design thin as well as moderately thick plates.

The second modification leads to important consequences, too. Such consequences constitute just as many advantages for the resolution of the related optimisation problem. Firstly, the utilisation of iso-geometric surfaces leads to a considerable reduction in the number of material design variables (at the macroscale), i.e. the polar parameters defined in each point of the *control net* of the B-spline surface. Secondly, thanks to the *strong convex-hull property* of the B-spline blending functions the optimisation constraints of the problem, related to the specifications of the considered application, can be imposed only on the control points of the net: if they are satisfied on such points they are automatically met over the whole domain.

As previously stated the goal of the first level of the strategy is the maximisation of the buckling load of the VAT laminate by simultaneously satisfying the feasibility constraints on the distribution of the laminate polar parameters over the plate. All of these aspects are detailed in the following subsection.

3.1 Mechanical Design Variables

In the framework of the FSDT theory [33] the constitutive law of the laminated plate (expressed within the global frame of the laminate $R = \{0; x, y, z\}$) can be stated as:

$$\begin{Bmatrix} \{\mathrm{N}\} \\ \{\mathrm{M}\} \end{Bmatrix} = \begin{bmatrix} [\mathrm{A}] & [\mathrm{B}] \\ [\mathrm{B}] & [\mathrm{D}] \end{bmatrix} \begin{Bmatrix} \{\varepsilon_0\} \\ \{\chi_0\} \end{Bmatrix} , \tag{1}$$

$$\{\mathrm{F}\} = [\mathrm{H}]\,\{\gamma_0\} , \tag{2}$$

where [A], [B] and [D] are the membrane, membrane/bending coupling and bending stiffness matrices of the laminate, while [H] is the out-of-plane shear stiffness matrix. {N}, {M} and {F} are the vectors of membrane forces, bending moments and shear forces per unit length, respectively, whilst $\{\varepsilon_0\}$, $\{\chi_0\}$ and $\{\gamma_0\}$ are the vectors of in-plane strains, curvatures and out-of-plane shear strains of the laminate middle plane, respectively, [33].

In order to analyse the elastic response of the multilayer plate the best practice consists in introducing the laminate homogenised stiffness matrices defined as:

$$\begin{aligned} &[\mathrm{A}^*] = \frac{1}{h}[\mathrm{A}] && , \\ &[\mathrm{B}^*] = \frac{2}{h^2}[\mathrm{B}] && , \\ &[\mathrm{D}^*] = \frac{12}{h^3}[\mathrm{D}] && , \\ &[\mathrm{H}^*] = \begin{cases} \dfrac{1}{h}[\mathrm{H}] \quad (\text{basic}) & , \\ \dfrac{12}{5h}[\mathrm{H}] \ (\text{modified}) & . \end{cases} \end{aligned} \tag{3}$$

where h is the total thickness of the laminated plate.

In the framework of the polar formalism it is possible to express the Cartesian components of these matrices in terms of their elastic invariants. To the best of the authors' knowledge, in [21, 22] an invariant representation of the laminate stiffness matrices in the framework of the FSDT has been given for the by using the polar formalism [39] that gives a representation of any planar elasticity-like tensor by means of a complete set of independent invariants, i.e. the *polar parameters*. It can be proven that, also in the FSDT theoretical framework, in the case of a fully orthotropic, quasi-homogeneous laminate the overall number of independent mechanical design variables describing its mechanical response reduces to only three [21, 22]: the anisotropic polar parameters $R_{0K}^{A^*}$ and $R_1^{A^*}$ and the polar angle $\Phi_1^{A^*}$ (this last representing the orientation of the main orthotropy axis) of the homogenised membrane stiffness matrix [A*]. In fact, once the material of the constitutive ply is fixed, the number of polar parameters to be designed remains unchanged when passing from the theoretical framework of the CLT to that of the FSDT; this result is quite surprising and represents a further advantage coming from the utilisation of the polar method. For more details on the polar formalism and its application in the context of the FSDT the reader is addressed to [21, 22, 37].

For a VAT composite the three independent polar parameters (which completely describe the mechanical behaviour of the VAT laminate at the macroscopic scale) must vary point-wise over the structure. As stated beforehand, such a variation is expressed by means of B-spline surfaces. In particular, in the mathematical framework of the B-spline surfaces the variation of the laminate polar parameters can be expressed as:

$$\begin{aligned}
R_{0K}^{A^*}(\xi,\gamma) &= \sum_{i=0}^{n_p}\sum_{j=0}^{m_p} N_{i,p}(\xi)\, N_{j,q}(\gamma)\, R_{0K}^{A^*(i,j)},\\
R_{1}^{A^*}(\xi,\gamma) &= \sum_{i=0}^{n_p}\sum_{j=0}^{m_p} N_{i,p}(\xi)\, N_{j,q}(\gamma)\, R_{1}^{A^*(i,j)},\\
\Phi_{1}^{A^*}(\xi,\gamma) &= \sum_{i=0}^{n_p}\sum_{j=0}^{m_p} N_{i,p}(\xi)\, N_{j,q}(\gamma)\, \Phi_{1}^{A^*(i,j)}.
\end{aligned} \tag{4}$$

Equation (4) fully describes a B-spline surface (in the space of the laminate polar parameters) of degrees p and q along the parametric coordinates ξ and γ, respectively, as depicted in Fig. 1.

The dimensionless coordinates ξ and γ can be arbitrarily defined: a natural choice consists in linking them with the Cartesian coordinates of the laminated plate,

$$\xi = \frac{x}{a}\,, \gamma = \frac{y}{b}\,, \tag{5}$$

where a and b are the lengths of the plate edges along x and y axes, respectively. In Eq. (4) $\{R_{0K}^{A^*(i,j)}, R_{1}^{A^*(i,j)}, \Phi_{1}^{A^*(i,j)}\}$ $(i = 0, \cdots, n_p, j = 0, \cdots, m_p)$ are the values of the laminate polar parameters at the generic control point (the set of $(n_p + 1) \times (m_p + 1)$ control points forms the so-called *control network*), while $N_{i,p}(\xi)$ and $N_{j,q}(\gamma)$ are

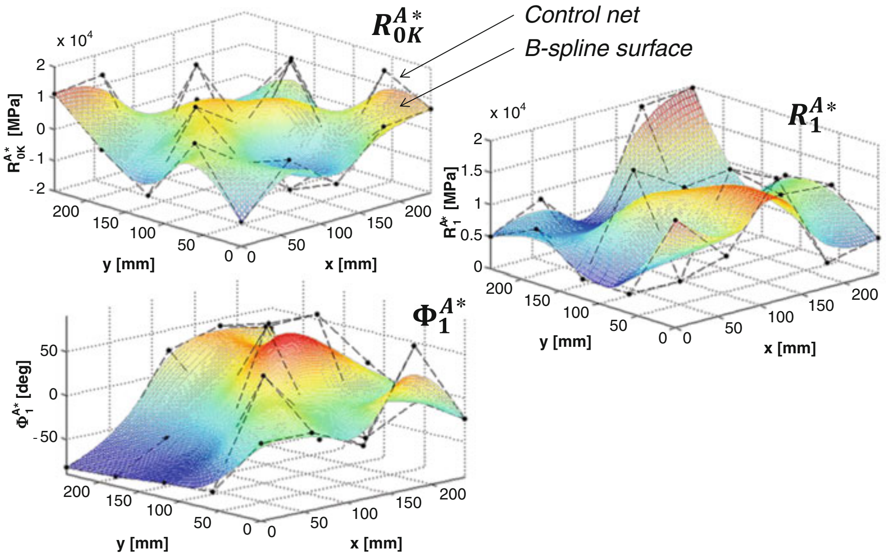


Fig. 1 Example of B-spline surfaces in the space of the laminate polar parameters

the *pth*-degree and *qth*-degree B-spline basis functions (along ξ and γ directions, respectively) defined on the non-periodic, non-uniform knot-vectors:

$$\begin{aligned}\boldsymbol{\Xi} &= \left\{\underbrace{0,\cdots,0}_{p+1}, \Xi_{p+1},\cdots,\Xi_{r-p-1},\underbrace{1,\cdots,1}_{p+1}\right\},\\ \boldsymbol{\Gamma} &= \left\{\underbrace{0,\cdots,0}_{q+1}, \Gamma_{q+1},\cdots,\Gamma_{s-q-1},\underbrace{1,\cdots,1}_{q+1}\right\}.\end{aligned} \tag{6}$$

It is noteworthy that the dimensions of the knot-vectors $\boldsymbol{\Xi}$ and $\boldsymbol{\Gamma}$ are $r+1$ and $s+1$, respectively, with:

$$\begin{aligned} r &= n_p + p + 1\ ,\\ s &= m_p + q + 1\ . \end{aligned} \tag{7}$$

For a deeper insight in the matter the reader is addressed to [31].

As previously stated, the use of iso-geometric surfaces for describing the variation of the mechanical design variables over the structure implies that the three independent polar parameters $\Phi_1^{A^*}$, $R_{0K}^{A^*}$ and $R_1^{A^*}$ have no discontinuity over the plate. Moreover, thanks to the B-spline representation the mechanical design variables (i.e. the laminate polar parameters) must be determined solely on each point of the *control net*, implying in this way a significant reduction in the number of design variables involved within the first-level problem.

Therefore, the optimisation variables of the problem can be grouped into the following vector:

$$\mathbf{x} = \left\{\Phi_1^{A^*(0,0)},\cdots,\Phi_1^{A^*(n_p,m_p)}, R_{0K}^{A^*(0,0)},\cdots,R_{0K}^{A^*(n_p,m_p)}, R_1^{A^*(0,0)},\cdots,R_1^{A^*(n_p,m_p)}\right\}\ . \tag{8}$$

The total number of design variables is hence equal to $3 \times (n_p + 1) \times (m_p + 1)$.

In addition, in the formulation of the optimisation problem for the first level of the strategy, the geometric and feasibility constraints on the polar parameters (which arise from the combination of the layer orientations and positions within the stack) must also be considered. These constraints ensure that the optimum values of the polar parameters resulting from the first step correspond to a feasible laminate that will be designed during the second step of the optimisation strategy, see [38]. Since the laminate is quasi-homogeneous, such constraints can be written only for matrix [A*] as follows:

$$\begin{cases} -R_0 \leq R_{0K}^{A^*} \leq R_0 \ , \\ 0 \leq R_1^{A^*} \leq R_1 \ , \\ 2\left(\dfrac{R_1^{A^*}}{R_1}\right)^2 - 1 - \dfrac{R_{0K}^{A^*}}{R_0} \leq 0 \ . \end{cases} \tag{9}$$

As explained beforehand, thanks to the *strong convex-hull property* these constraints have to be checked only on the points of the control network. If they are met on these points they will be satisfied over the whole domain of the B-spline surface. This aspect represents a further advantage when using the B-spline representation for the mechanical design variables. Moreover, first and second constraints of Eq. (9) can be taken into account as admissible intervals for the relevant optimisation variables, i.e., on $R_{0K}^{A^*\,(i,j)}$ and $R_1^{A^*\,(i,j)}$. Hence, the resulting feasibility constraint on the laminate polar parameters of the generic control point is

$$g_{ij}(\mathbf{x}) = 2\left(\frac{R_1^{A^*\,(i,j)}}{R_1}\right)^2 - 1 - \frac{R_{0K}^{A^*\,(i,j)}}{R_0} \leq 0 \ . \tag{10}$$

with $i = 0, \cdots, n_p$ and $j = 0, \cdots, m_p$. The total number of feasibility constraints to be imposed is thus equal to $(n_p + 1) \times (m_p + 1)$.

For a wide discussion upon the laminate feasibility and geometrical bounds as well as on the importance of the quasi-homogeneity assumption the reader is addressed to [38].

3.2 Mathematical Statement of the Problem

The first-level problem focuses on the definition of the optimal distribution of the laminate polar parameters. In this background, the solution of the structural optimisation problem is searched for an orthotropic quasi-homogeneous (locally, i.e., point-wise) plate subject to given BCs.

Therefore the optimisation problem can be formulated as follows:

$$\begin{aligned} &\min_{\mathbf{x}} \ -\lambda\,(\mathbf{x}) \\ &\text{subject to :} \\ &g_{ij}(\mathbf{x}) \leq 0 \ , (i = 0, \cdots, n_p, \ \ j = 0, \cdots, m_p) \end{aligned} \tag{11}$$

where λ is the first buckling factor of the laminated structure.

3.3 Numerical Strategy

Problem (11) is a non-linear, non-convex problem in terms of the mechanical design variables. Its non-linearity and non-convexity is due to the nature of the objective function, the first buckling factor, that is a non-convex function in terms of the orthotropy orientation. In addition, the complexity of such a problem is also due to the feasibility constraints imposed on the polar parameters of the plate, see Eq. (10). We recall that the overall number of design variables and optimisation constraints for problem (11) is $3 \times (n_p + 1) \times (m_p + 1)$ and $(n_p + 1) \times (m_p + 1)$, respectively.

For the resolution of problem (11) a hybrid optimisation tool, composed of the GA BIANCA [16] interfaced with the MATLAB *fmincon* algorithm [15], coupled with an FE model of the plate (used for numerical calculation of the first buckling load) has been developed, see Fig. 2.

The GA BIANCA was already successfully applied to solve different kinds of real-world engineering problems [5, 24]. As shown in Fig. 2, the optimisation procedure for the first-level problem is split into two phases. During the first phase the GA BIANCA is interfaced with the FE model of the VAT plate: for each individual at each generation, a FE-based buckling analysis is carried out for the evaluation of the first buckling load of the structure. The FE model makes use of the mechanical design variables, given by BIANCA and elaborated by an ANSYS

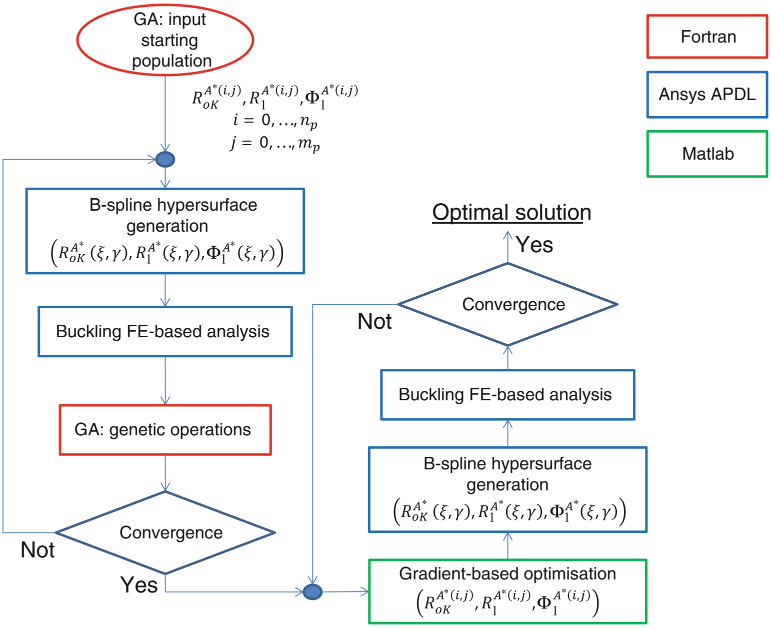

GA: input starting population
$R_{oK}^{A^*(i,j)}, R_1^{A^*(i,j)}, \Phi_1^{A^*(i,j)}$
$i = 0, \ldots, n_p$
$j = 0, \ldots, m_p$
B-spline hypersurface generation
$\left(R_{oK}^{A^*}(\xi,\gamma), R_1^{A^*}(\xi,\gamma), \Phi_1^{A^*}(\xi,\gamma)\right)$
Buckling FE-based analysis
GA: genetic operations
Convergence
Not
Yes
Fortran
Ansys APDL
Matlab
Optimal solution
Yes
Convergence
Not
Buckling FE-based analysis
B-spline hypersurface generation
$\left(R_{oK}^{A^*}(\xi,\gamma), R_1^{A^*}(\xi,\gamma), \Phi_1^{A^*}(\xi,\gamma)\right)$
Gradient-based optimisation
$\left(R_{oK}^{A^*(i,j)}, R_1^{A^*(i,j)}, \Phi_1^{A^*(i,j)}\right)$

Fig. 2 Logical flow of the numerical procedure employed for the solution search of the first-level problem

parametric design language (APDL) macro which generates the B-spline surface representing the distribution of the polar parameters over the VAT plate, in order to calculate the first buckling load of the structure. At the end of the FE analysis, the GA elaborates the results provided by the FE model (in terms of objective and constraint functions) in order to execute the genetic operations. These operations are repeated until the GA meets the user-defined convergence criterion. The generic individual of the GA represents a potential solution for the problem at hand. The genotype of the individual for problem (11) is characterised by $(n_p + 1) \times (m_p + 1)$ chromosomes composed of 3 genes, each one coding a component of the vector of the design variables. Due to the strong non-convex nature of problem (11), the aim of the genetic calculation is to provide a potential sub-optimal point in the design space which constitutes the initial guess for the subsequent phase, i.e. the local optimisation, where the *fmincon* gradient-based algorithm is interfaced with the same FE model of the VAT plate.

4 Mathematical Formulation of the Second-Level Problem

The second-level problem concerns the lay-up design of the VAT laminated plate. The goal of this problem is the determination of at least one stacking sequence satisfying the optimum values of the distribution of the polar parameters over the structure resulting from the first level of the strategy and having the elastic symmetries imposed to the laminate within the formulation of the first-level problem, i.e. quasi-homogeneity and orthotropy.

In the case of a VAT solution the fibres orientation angle varies point-wise in every ply composing the laminate. Hence a proper description of the fibres-path is necessary to formulate and solve the second-level problem of the MS2L strategy. To this purpose, the following modifications have been brought to the second step of the MS2L optimisation procedure:

- the point-wise variation of the fibre orientation (in each ply) is described through the use of a B-spline surface;
- the technological constraint on the minimum radius of curvature of the pre-preg strips is taken into account.

These improvements lead to important advantages in solving the related optimisation problem. In fact, the use of B-spline surfaces allows, as in the case of the first-level problem, to reduce the total number of design variables: in this case it is sufficient to calculate the fibre orientation solely at each point of the B-spline control network. In addition, thanks to the use of iso-geometric blending functions the local steering (i.e. the local radius of curvature of the tow) can be determined easily and introduced in the problem formulation as an optimisation constraint. This last aspect is of paramount importance to obtain a proper formulation of the technological constraints regarding the layout of pre-preg strips in each ply which cannot exceed a given curvature.

Concerning the representation of the fibres-path, the relative B-spline surface for each ply is defined as:

$$\delta_k\left(\xi,\gamma\right)=\sum_{i=0}^{n_p}\sum_{j=0}^{m_p}N_{i,p}\left(\xi\right)N_{j,q}\left(\gamma\right)\delta_k{}^{(i,j)} \text{ with } k=1,\cdots,n. \tag{12}$$

In this case $\delta_k{}^{(i,j)}$ is the orientation angle at the generic control point for the k-th layer, i.e. the design variables of the second-level problem of the MS2L strategy whose overall number is equal to $n\times(n_p+1)\times(m_p+1)$.

In the framework of the polar formalism, the problem of the lay-up design of the VAT laminate can be stated in the form of a constrained minimisation problem:

$$\begin{cases} \min\limits_{\delta_k{}^{(i,j)}} I\left(\delta_k{}^{(i,j)}\right) & k=1,\cdots,n\ , \\ & i=0,\cdots,n_p\ , \\ & j=0,\cdots,m_p\ , \\ r_{adm}-r_{min}\le 0\ . \end{cases} \tag{13}$$

In Eq. (13) r_{adm} is the minimum admissible radius of curvature of the tow whose value depends upon the AFP process, while r_{min} is the local least radius of curvature among all the plies. r_{min} is defined as:

$$\begin{aligned} &r_{min}=\min_k\left[\min_{(x,y)} r_k(x,y)\right]\ , \\ &r_k(x,y)=(\mathbf{t}_k\cdot\nabla\delta_k)^{-1}\ , \quad && k=1,\cdots,n\ , \\ & && x\in[0,\ a]\ , \\ & && y\in[0,\ b]\ . \end{aligned} \tag{14}$$

In Eq. (14) $\mathbf{t}_k$ is the local tangent vector of the angular field $\delta_k(x,y)$ of the k-th ply, while $\nabla\delta_k$ is the gradient of the fibre path with respect to coordinates (x,y), namely

$$\begin{aligned} \mathbf{t}_k &= \{\cos\delta_k\ ,\sin\delta_k\}\ , \\ \nabla\delta_k &= \left\{\frac{1}{a}\frac{\partial\delta_k}{\partial\xi}\ ,\frac{1}{b}\frac{\partial\delta_k}{\partial\gamma}\right\}\ . \end{aligned} \tag{15}$$

In Eq. (13) $I(\delta_k{}^{(i,j)})$ is the overall objective function which is defined as:

$$I\left(\delta_k{}^{(i,j)}\right)=\sum_{i=1}^{6}f_i\left(\delta_k^{(i,j)}\right)\ . \tag{16}$$

where $f_i(\delta_k^{(i,j)})$ are quadratic functions in the space of polar parameters, each one representing a requirement to be satisfied. For the problem at hand the partial objective functions write:

$$
\begin{aligned}
f_1(\delta_k^{(i,j)}) &= \int_0^1\int_0^1 \left[\frac{\Phi_0^{A^*}(\delta_k(\xi,\gamma)) - \Phi_1^{A^*}(\delta_k(\xi,\gamma))}{\pi/4} - K^{A^*(opt)}(\xi,\gamma)\right]^2 d\xi\, d\gamma\ ,\\
f_2(\delta_k^{(i,j)}) &= \int_0^1\int_0^1 \left[\frac{R_0^{A^*}(\delta_k(\xi,\gamma)) - R_0^{A^*(opt)}(\xi,\gamma)}{R_0}\right]^2 d\xi\, d\gamma\ ,\\
f_3(\delta_k^{(i,j)}) &= \int_0^1\int_0^1 \left[\frac{R_1^{A^*}(\delta_k(\xi,\gamma)) - R_1^{A^*(opt)}(\xi,\gamma)}{R_1}\right]^2 d\xi\, d\gamma\ ,\\
f_4(\delta_k^{(i,j)}) &= \int_0^1\int_0^1 \left[\frac{\Phi_1^{A^*}(\delta_k(\xi,\gamma)) - \Phi_1^{A^*(opt)}(\xi,\gamma)}{\pi/4}\right]^2 d\xi\, d\gamma\ ,\\
f_5(\delta_k^{(i,j)}) &= \int_0^1\int_0^1 \left[\frac{||\,[\mathrm{C}]\,(\delta_k(\xi,\gamma))||}{||\,[\mathrm{Q}]\,||}\right]^2 d\xi\, d\gamma\ ,\\
f_6(\delta_k^{(i,j)}) &= \int_0^1\int_0^1 \left[\frac{||\,[\mathrm{B}^*]\,(\delta_k(\xi,\gamma))||}{||\,[\mathrm{Q}]\,||}\right]^2 d\xi\, d\gamma\ ,
\end{aligned}
\tag{17}
$$

where

$$
K^{A^*(opt)}(\xi,\gamma) = \begin{cases} 1 \text{ if} & R_{0K}^{A^*(opt)}(\xi,\gamma) < 0\ ,\\ 0 \text{ otherwise}\ . \end{cases}
\tag{18}
$$

In Eq. (17) $f_1(\delta_k^{(i,j)})$ represents the elastic requirement on the orthotropy of the laminate having the prescribed shape, $f_2(\delta_k^{(i,j)})$, $f_3(\delta_k^{(i,j)})$ and $f_4(\delta_k^{(i,j)})$ are the requirements related to the prescribed values of the optimal polar parameters resulting from the first-level problem, while $f_5(\delta_k^{(i,j)})$ and $f_6(\delta_k^{(i,j)})$ are linked to the quasi-homogeneity condition. For more details on the meaning of the partial objective functions, on the elastic symmetries of the laminate in the framework of the FSDT and on the symbols appearing in Eq. (17), the reader is addressed to [21, 22].

$I(\delta_k^{(i,j)})$ is a positive semi-definite convex function in the space of laminate polar parameters, since it is defined as a sum of convex functions, see Eqs. (16)–(17). Nevertheless, such a function is highly non-convex in the space of plies orientations because the laminate polar parameters depend upon circular functions of the layers orientation angles, see [21, 22]. Moreover, one of the advantages of such a formulation consists in the fact that the absolute minima of $I(\delta_k^{(i,j)})$ are known a priori since they are the zeroes of this function. For more details about the nature of the second-level problem see [16, 17, 19]. Concerning the numerical strategy for solving problem (13) the GA BIANCA has been employed to find a solution also for the second-level problem. In this case, each individual is composed of n chromosomes (one for each ply), each one characterised by $\left(n_p + 1\right) \times \left(m_p + 1\right)$ genes coding the layer orientation angle for each control point of the chromosome-ply.

5 Finite Element Model of the VAT Laminate

In order to determine the current value of the objective function (the first buckling factor) and that of the optimisation constraints of problem (11) a classical eigenvalue buckling analysis must be achieved for the VAT composite. The need to analyse, within the same calculation, different configurations of the VAT plate requires the creation of an *ad-hoc* input file for the FE model that has to be interfaced with the hybrid (GA + gradient-based algorithms) optimisation tool.

The FE model of the VAT laminated plate (see Fig. 3) employed during the first step of the MS2L strategy is built within the ANSYS environment and is made of SHELL281 elements based on the Reissner–Mindlin kinematic model, having 8 nodes and six degrees of freedom (DOFs) per node. The mesh size is chosen after preliminary mesh sensitivity analyses on the convergence of the value of the first buckling load for a given set of BCs. It was observed that a mesh having 2482 DOFs is sufficient to properly evaluate the first buckling load of the structure.

It is noteworthy that the B-spline mathematical formalism has been implemented by the authors into the ANSYS environment by using the ANSYS APDL [2] for creating a set of appropriate macros that were integrated within the FE model of the VAT plate. At this stage, the plate is modelled as an equivalent homogeneous anisotropic plate whose stiffness matrices ([A*], [B*], [D*] and [H*]) vary point-wise, i.e. for each element discretising the real structure. In particular, in order to properly define, for every element of the VAT plate, the correct value of its stiffness properties the following strategy has been employed:

1. for a given set of the laminate polar parameters defined in each control point (the design variables passed from the optimisation tool to the FE model of the VAT plate, see Fig. 3), build the corresponding B-spline surfaces;
2. discretise the plate into N_e elements;

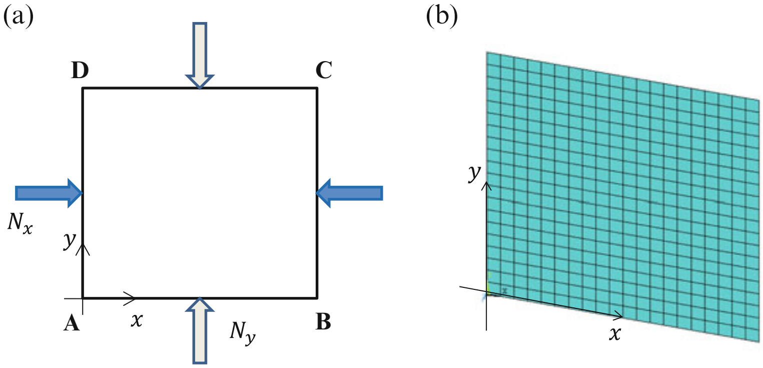


Fig. 3 Geometry of the VAT plate and applied BCs (**a**) and FE model of the structure (**b**)

Table 2 BCs of the FE model of the VAT laminated plate

Sides	BCs
AB, CD	$U_x = 0$
	$U_z = 0$
BC, DA	$U_y = 0$
	$U_z = 0$

3. fix the element index i: for the i-th element retrieve the Cartesian coordinates of its centroid, i.e. (x_e^i, y_e^i) and calculate the corresponding dimensionless coordinates (ξ_e^i, γ_e^i) according to Eq. (5);
4. calculate the laminate polar parameters (and hence the Cartesian components of the stiffness matrices of the laminate) for (ξ_e^i, γ_e^i) and assign the material properties to the element i;
5. repeat points 3 and 4 for each element of the plate.

Finally, the linear buckling analysis is performed using the BCs depicted in Fig. 3 and listed in Table 2.

6 Numerical Example

In this section a meaningful numerical example is considered in order to prove the effectiveness of the MS2L strategy for the optimum design of VAT laminates. As depicted in Fig. 3, a bi-axial compressive load per unit length is applied on the plate edges with a ratio $\frac{N_y}{N_x} = 0.5$. The plate has a square geometry with side length $a = b = 254$ mm and is made of $n = 24$ plies whose material properties are those listed in Table 1. Concerning the first-level problem, the parameters defining the B-spline surfaces which describe the polar parameters distribution over the VAT plate are set as: $n_p = m_p = 4$ (hence five control points along each direction), $p = q = 2$ (degrees of the blending functions along each direction). Moreover, each B-spline is defined over the following uniform knot-vectors:

$$\begin{aligned} \boldsymbol{\Xi} &= \left\{0, 0, 0, \tfrac{1}{3}, \tfrac{2}{3}, 1, 1, 1\right\}, \\ \boldsymbol{\Gamma} &= \left\{0, 0, 0, \tfrac{1}{3}, \tfrac{2}{3}, 1, 1, 1\right\}. \end{aligned} \tag{19}$$

Accordingly, for the first-level problem the overall number of design variables and optimisation constraints is 75 and 25, respectively. The mechanical design variables together with their nature and bounds for the first-level problem are listed in Table 3.

Concerning the second-level problem, the parameters defining the B-spline surface which describes the point-wise variation of the fibre orientation angle (for each ply) are the same as those employed during the first step of the strategy. This means that the overall number of design variables for the second-level problem is significant and equal to 600 (i.e. 25 orientation angles defined in each control point

Table 3 Design space of the first-level problem

Design variable	Type	Lower bound	Upper bound
$R_{0K}^{A*(i,j)}$ [MPa]	Continuous	−17693.3868	17693.3868
$R_1^{A*(i,j)}$ [MPa]	Continuous	0.0	19072.0711
$\Phi_1^{A*(i,j)}$ [deg]	Continuous	−90.0	90.0

Table 4 Genetic parameters of the GA BIANCA for both first- and second-level problems

Genetic parameters	1st level problem	2nd level problem
No. of populations	1	1
No. of individuals	500	2000
No. of generations	200	1000
Crossover probability	0.85	0.85
Mutation probability	0.002	0.0005
Selection operator	Roulette-wheel	Roulette-wheel
Elitism operator	Active	Active

per layer), while there is only one optimisation constraint, see Eq. (13). In addition, the reference value for the minimum admissible radius of curvature of the tow, i.e. r_{adm} is set equal to 30 mm.

It must be highlighted the fact that $\delta_k^{(i,j)}$ are continuous variables in the range $[-90°, \ 90°]$.

Regarding the setting of the genetic parameters for the GA BIANCA utilised for both first- and second-level problems they are listed in Table 4. Moreover, concerning the constraint-handling technique for both levels of the strategy the automatic dynamic Penalisation (ADP) method has been employed, see [20]. For more details on the numerical techniques developed within the new version of BIANCA and the meaning of the values of the different parameters tuning the GA the reader is addressed to [16, 19].

As far as concerns the *fmincon* optimisation tool employed for the local solution search at the end of the first step, the numerical algorithm chosen to carry out the calculations is the *active-set* method with non-linear constraints. For more details on the gradient-based approaches implemented into MATLAB, the reader is addressed to [15].

Before starting the multi-scale optimisation process a reference structure must be defined in order to establish a reference value for the first buckling factor of the plate. The reference structure is still a square plate of side $a = b = 254$ mm composed of 24 unidirectional fibre-reinforced laminae whose material properties are those listed in Table 1. The stacking sequence of the reference solution is $[0/-45/0/45/90/45/0/-45/90/45/90/-45]_s$. The choice of the reference solution has been oriented towards a symmetric quasi-isotropic stack, of common use in real-world engineering applications, which constitutes a ‘good’ compromise between weight and stiffness requirements (in terms of buckling load): such a configuration is characterised by a buckling factor $\lambda_{ref} = 81.525$ when $N_x = 1$ N/mm and $N_y = 0.5$ N/mm.

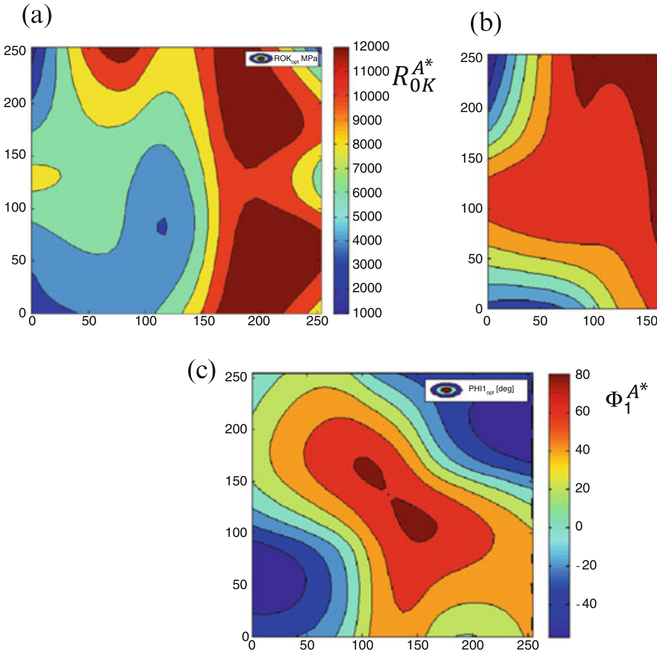


Fig. 4 Optimal distribution of the polar parameters ${R_{0K}}^{A^*}$ (**a**), ${R_1}^{A^*}$ (**b**) and ${\Phi_1}^{A^*}$ (**c**) over the VAT plate resulting from the first-level optimisation problem

Table 5 Optimum value of ${R_{0K}}^{A^*}$ [MPa] for each control point of the B-spline surface

n_p \ m_p	0	1	2	3	4
0	3591.3747	3668.8094	9592.3869	4497.2954	975.9237
1	4004.9094	8579.8092	7560.9021	9286.1572	17011.3438
2	6579.9750	805.5577	3036.0234	5777.9895	6901.0129
3	16467.7981	16448.7345	15040.6238	16596.7375	16926.6619
4	8529.0554	15762.0163	4406.7537	14321.3969	2829.6760

Concerning the first-level problem, the optimum distribution of the laminate polar parameters over the VAT plate is illustrated in Fig. 4, while the optimum value of the mechanical design variables for each control point are listed in Tables 5, 6, 7. On the other hand, concerning the solution of the second-level problem, an illustration of the optimum fibres-path for the firsts four layers (for sake of synthesis) is depicted in Fig. 5. It is noteworthy that the optimal solution found at the end of the MS2L design procedure is characterised by a buckling factor of 173.94 which is about 114 % higher than the reference counterpart and, in the meantime, satisfies the technological constraint on the minimum (local) radius of curvature of the tow imposed by the AFP process.

Table 6 Optimum value of ${R_1}^{A^*}$ [MPa] for each control point of the B-spline surface

m_p n_p	0	1	2	3	4
0	5205.1699	12784.5159	16065.5107	5068.0708	1932.6002
1	3447.8699	16298.3907	16011.8612	16593.1744	17332.2704
2	13511.7905	13789.0773	14452.1311	15045.4781	15809.2434
3	14966.4897	18733.6564	18342.5388	18722.0860	18735.1483
4	10265.5066	17842.4603	14882.8151	16217.0759	10244.8019

Table 7 Optimum value of ${\Phi_1}^{A^*}$ [deg] for each control point of the B-spline surface

m_p n_p	0	1	2	3	4
0	−20.9816	−84.5402	11.1669	38.4351	−0.9851
1	−30.7624	−80.1076	56.6921	86.1616	53.2841
2	79.6062	82.6653	89.9998	88.5717	−26.4919
3	−3.9816	60.5155	89.3887	−88.5577	−28.6826
4	45.3655	57.9974	30.2596	−84.8941	−26.7837

From a careful analysis of the optimum configuration of the VAT laminated plate provided by the MS2L procedure, it is possible to deduce the following facts.

- The polar parameters distribution resulting from the first step of the strategy is totally asymmetric. Symmetric solutions are, of course, possible: it is sufficient to impose the symmetry condition directly on the values of the laminate polar parameters at the points of the control network of the B-spline surfaces. However, in order to state and solve the optimisation problem in the most general case, in this study we prefer of not imposing such a condition.
- When looking at the optimum distribution of the laminate polar parameters (Fig. 4), one can notice that the laminate is always characterised by an ordinary orthotropy shape with $K^{A^*} = 0$ because $R_{0K}^{A^*}(\xi, \gamma)$ is strictly positive over the laminated plate.
- Unlike the vast majority of works reported in literature [40], the optimum fibres-path for each ply is very general. In the framework of the proposed approach, the point-wise variation of the fibre orientation angle in every lamina does not follow simple linear or parabolic variations (with respect to laminate global frame) as in [40], rather it is described by a general B-spline surface, see Eq. (12). This fact, together with the very general formulation of problem (13), allows the designer to find (at the cost of a considerable computation effort) an optimum stack meeting all the requirements (i.e. elastic and manufacturability constraints) provided by problem (13), without the need of a further post-processing treatment to simplify the trajectory of the tows in order to comply with the constraints imposed by the AFP process.

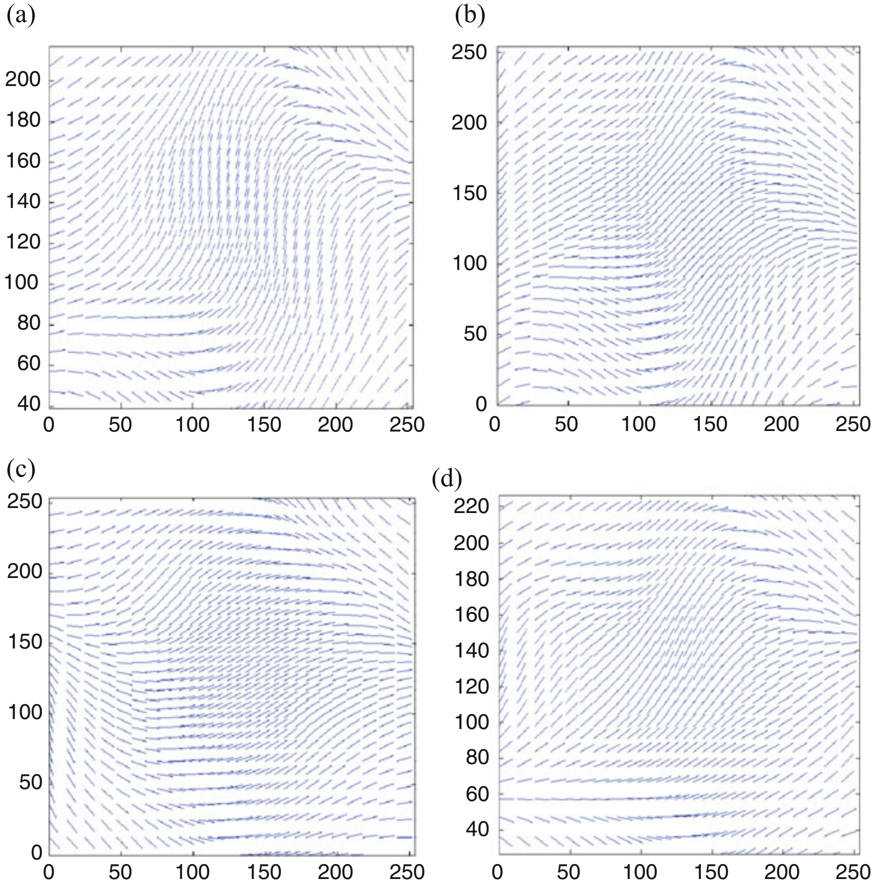


Fig. 5 Optimum fibres-path for the firsts four layers of the VAT plate resulting from the second-level optimisation problem, 1*st* ply (**a**), 2*nd* ply (**b**), 3*rd* ply (**c**) and 4*th* ply (**d**)

- Finally, the optimum fibres-path (for each layer) found at the end of the second step of the MS2L procedure does not need of a further step for the reconstruction of the CAD model because the variation of the fibres-path is described by a B-spline surface which is fully compatible with several standard file formats (IGES, STL and STEP), allows in this way a rapid exchange of information among the CAD tool and the software of the AFP process.

7 Conclusions and Perspectives

In this work a new paradigm for the design and optimisation of VAT composite structures is presented. This paradigm essentially relies on the utilisation of a MS2L optimisation procedure characterised by several features that make it an original, effective and general method for the multi-scale design of complex VAT structures. In the present work this strategy has been employed to deal with the problem of the maximisation of the buckling factor of a VAT plate subject to both mechanical and manufacturability constraints. On the one hand, the design process is not submitted to restrictions: any parameter characterising the VAT composite (at each scale) is an optimisation variable. This allows the designer to look for a *true global minimum*, hard to be obtained otherwise. On the other hand, both the formulation of the design problem and the MS2L optimisation strategy have been generalised and improved in order to be applied to the problem of designing a VS composite.

In the framework of the MS2L design methodology several modifications have been introduced for both first- and second-level problems.

Concerning the first-level problem the main modifications are: (1) the use of higher-order theories (introduced as result of [21–23]) for taking into account the influence of the transverse shear stiffness on the overall mechanical response of VAT composites and (2) the utilisation of B-spline surfaces for describing the distribution of the laminate polar parameters over the structure which allow for a continuous point-wise variation of the laminate stiffness matrices. This last aspect leads to some important advantages for the resolution of the related optimisation problem. Firstly, the utilisation of B-spline surfaces leads to a considerable reduction in the number of design variables (the polar parameters have to be defined solely in each point of the control network of the B-spline surface). Secondly, thanks to the *strong convex-hull* property of the B-spline blending functions the optimisation constraints of the problem (related to the specifications of the considered application) can be imposed only on the control points of the network: if they are satisfied on such points they are automatically met over the whole domain.

For the second-level problem the major modifications are: (1) the utilisation of B-spline basis functions for obtaining a continuous point-wise variation of the fibre orientation angle within each ply; (2) a proper mathematical formalisation of the manufacturability constraints linked to the AFP process in the framework of the B-spline representation. Also in the second step of the procedure, these modifications imply some important consequences. On the one hand, the utilisation of B-spline surfaces leads to an important reduction of the number of design variables (the orientation angles defined in each control point of the layer), while, on the other hand, the B-spline mathematical formalism allows to express in a closed analytical form the manufacturability constraints linked to the AFP process. All of these modifications allow to go beyond the main restrictions characterising the design activities and research studies on VAT composites that one can find in literature.

Finally, the improved version of the MS2L strategy has been tested through a meaningful numerical example which proved its effectiveness. The optimisation tool allows to find an optimum VAT laminate characterised by a significant increment of the first buckling factor (about the 114 %) when compared to a reference classical solution (composed of unidirectional plies).

Concerning the perspectives of this work, there are still some theoretical, numerical and technical aspects and features that need to be deeply investigated and developed in order to make the proposed approach a very general and comprehensive strategy able to provide solutions that are both efficient (true optimal configurations) and manufacturable. Of course, this action passes through a real understanding of the potential and the technological restrictions linked to the AFP process. Currently, only the technological constraint on the tow steering has been integrated in the MS2L strategy. A step forward can be realised by properly formalising and including into the design problem other kinds of manufacturability constraints: tow gap and overlapping, the variation of the fibre volume fraction due to imperfections, etc. Moreover, in the framework of the MS2L optimisation procedure proposed in this work the manufacturability constraint linked to the minimum admissible radius of curvature of the tow has been integrated only within the second step of the design procedure. Actually, when using such an approach, there is no warranty that the optimisation algorithm could find an optimum fibres-path able to meet, on the one hand, the optimum distribution of the laminate polar parameters resulting from the first step of the strategy and, on the other hand, the manufacturability constraint related to the tow steering condition. To overcome such an issue, the formulation of the first-level problem should be modified accordingly, i.e. by integrating the manufacturability constraints since the first stage of the MS2L strategy. In addition, from a numerical point of view, the designer could be interested in optimising also the number of design variables (i.e. the number of the parameters tuning the shape of the B-spline surfaces) involved into both levels of the MS2L procedure: this point can be easily taken into account by exploiting the original features of the GA BIANCA. Finally, further modifications may also be considered in the formulation of the design problem depending on the nature of the considered application, e.g. by including constraints on inter- and intra-laminar damage, variability effects linked to the fabrication process, costs, etc.

Research is ongoing on all of the previous aspects.

References

1. Abdalla, M.M., Setoodeh, S., Gürdal, Z.: Design of variable stiffness composite panels for maximum fundamental frequency using lamination parameters. Compos. Struct. **81**, 283–291 (2007)
2. Ansys: ANSYS Mechanical APDL Basic Analysis Guide. Release 15.0, ANSYS, Inc., Southpointe, 275 Technology Drive, Canonsburg, PA 15317 (2013)
3. Blom, A.W., Abdalla, M.M., Gürdal, Z.: Optimization of course locations in fiber-placed panels for general fiber angle distributions. Compos. Sci. Tec. **70**, 564–570 (2010)

4. Catapano, A., Montemurro, M.: A multi-scale approach for the optimum design of sandwich plates with honeycomb core. Part I: homogenisation of core properties. Compos. Struct. **118**, 664–676 (2014)
5. Catapano, A., Montemurro, M.: A multi-scale approach for the optimum design of sandwich plates with honeycomb core. Part II: the optimisation strategy. Compos. Struct. **118**, 677–690 (2014)
6. Catapano, A., Desmorat, B., Vannucci, A.: Stiffness and strength optimization of the anisotropy distribution for laminated structures. J. Optim. Theory App. **167**(1), 118–146 (2015)
7. Dang, T.D., Hallet, S.R.: A numerical study on impact and compression after impact behaviour of variable angle tow laminates. Compos. Struct. **96**, 194–206 (2013)
8. Ghiasi, H., Pasini, D., Lessard, L.: Optimum stacking sequence design of composite materials. Part I: Constant stiffness design. Compos. Struct. **90**, 1–11 (2009)
9. Ghiasi, H., Fayazbakhsh, K., Pasini, D., Lessard, L.: Optimum stacking sequence design of composite materials. Part II: Variable stiffness design. Compos. Struct. **93**, 1–13 (2010)
10. Gürdal, Z., Tatting, B.F., Wu, K.C.: Variable stiffness panels: Effects of stiffness variation on the in-plane and buckling responses. Compos. Part A-Appl. S. **39**(9), 11–22 (2008)
11. Hammer, V.B., Bendsoe, M.P., Lipton, R., Pedersen, P.: Parametrization in laminate design for optimal compliance. Int. J. Solids Struct. **34**(4), 415–434 (1997)
12. Hyer, M.W., Lee, H.H.: The use of curvilinear fiber format to improve buckling resistance of composite plates with central circular holes. Compos. Struct. **18**, 239–261 (1991)
13. Kim, B.C., Weaver, P.M., Potter, K.: Manufacturing characteristics of the continuous tow shearing method for manufacturing of variable angle tow composites. Compos. Part A-Appl. S. **61**, 141–151 (2014)
14. Kim, B.C., Weaver, P.M., Potter, K.: Computer aided modelling of variable angle tow composites manufactured by continuous tow shearing. Compos. Struct. **129**, 256–267 (2015)
15. Mathworks: Optimization Toolbox (2016). http://it.mathworks.com/help/optim/index.html
16. Montemurro, M.: Optimal design of advanced engineering modular systems through a new genetic approach. PhD thesis, UPMC, Paris VI, France (2012). http://tel.archives-ouvertes.fr/tel-00955533
17. Montemurro, M., Vincenti, A., Vannucci, P.: Design of elastic properties of laminates with minimum number of plies. Mech. Compos. Mater. **48**, 369–390 (2012)
18. Montemurro, M., Vincenti, A., Vannucci, P.: A two-level procedure for the global optimum design of composite modular structures - application to the design of an aircraft wing. Part 1: theoretical formulation. J. Optim. Theory App. **155**(1), 1–23 (2012)
19. Montemurro, M., Vincenti, A., Vannucci, P.: A two-level procedure for the global optimum design of composite modular structures - application to the design of an aircraft wing. Part 2: numerical aspects and examples. J. Optim. Theory App. **155**(1), 24–53 (2012)
20. Montemurro, M., Vincenti, A., Vannucci, P.: The automatic dynamic penalisation method (ADP) for handling constraints with genetic algorithms. Comput. Method Appl. M **256**, 70–87 (2013)
21. Montemurro, M.: An extension of the polar method to the first-order shear deformation theory of laminates. Compos. Struct. **127**, 328–339 (2015)
22. Montemurro, M.: Corrigendum to “an extension of the polar method to the first-order shear deformation theory of laminates” [Compos Struct 127 (2015) 328–339]. Compos. Struct. **131**, 1143–1144 (2015)
23. Montemurro, M.: The polar analysis of the third-order shear deformation theory of laminates. Compos. Struct. **131**, 775–789 (2015)
24. Montemurro, M., Catapano, A., Doroszewski, D.: A multi-scale approach for the simultaneous shape and material optimisation of sandwich panels with cellular core. Compos. Part B-Eng. **91**, 458–472 (2016)
25. Nagendra, S., Kodiyalam, S., Davis, J., Parthasarathy, V.: Optimization of tow fiber paths for composite design. In: Proceedings of the AIAA/ASME/ASCE/AHS/ASC 36th Structures, Structural Dynamics and Materials Conference, vol. AIAA 95-1275, New Orleans, LA (1995)

26. Nagendra, S., Jestin, D., Gürdal, Z., Haftka, R.T., Watson, L.T.: Improved genetic algorithm for the design of stiffened composite panels. Comput. Struct. **58**(3), 543–555 (1996)
27. Nik, M.A., Fayazbakhsh, K., Pasini, D., Lessard, L.: Surrogate-based multi-objective optimization of a composite laminate with curvilinear fibers. Compos. Struct. **94**, 2306–2313 (2012)
28. Nik, M.A., Fayazbakhsh, K., Pasini, D., Lessard, L.: Optimization of variable stiffness composites with embedded defects induced by automated fiber placement. Compos. Struct. **107**, 160–166 (2014)
29. Parnas, L., Oral, S., Ceyhan, U.: Optimum design of composite structures with curved fiber courses. Compos. Sci. Technol. **63**, 1071–1082 (2003)
30. Peeters, D.M.J., Hesse, S., Abdalla, M.M.: Stacking sequence optimisation of variable stiffness laminates with manufacturing constraints. Compos. Struct. **125**, 596–604 (2015)
31. Piegl, L., Tiller, W.: The NURBS Book. Springer, New York (1997)
32. Raju, G., Wu, Z., Kim, B.C., Weaver, P.M.: Pre-buckling and buckling analysis of variable angle tow plates with general boundary conditions. Compos. Struct. **94**, 2961–2970 (2012)
33. Reddy, J.N.: Mechanics of Composite Laminated Plates and Shells: Theory and Analysis. CRC Press, Boca Raton (2003)
34. Setoodeh, S., Abdalla, M.M., Gürdal, Z.: Design of variable stiffness laminates using lamination parameters. Compos. Part B-Eng. **37**, 301–309 (2006)
35. Setoodeh, S., Abdalla, M.M., IJsselmuiden, S.T., Gürdal, Z.: Design of variable stiffness composite panels for maximum buckling load. Compos. Struct. **87**, 109–117 (2006)
36. Van Campen, J.M.J.F., Kassapoglou, C., Gürdal, Z.: Generating realistic laminate fiber angle distributions for optimal variable stiffness laminates. Compos. Part B-Eng. **43**(2), 354–360 (2011)
37. Vannucci, P.: Plane anisotropy by the polar method. Meccanica **40**, 437–454 (2005)
38. Vannucci, P.: A note on the elastic and geometric bounds for composite laminates. J. Elasticity **112**, 199–215 (2013)
39. Verchery, G.: Les invariants des tenseurs d'ordre 4 du type de l'élasticité. In: Proc. of Colloque Euromech 115, VIllard-de-Lans, France (1979)
40. Wu, Z., Raju, G., Weaver, P.M.: Framework for the buckling optimisation of variable-angle-tow composite plates. AIAA J **53**(12), 3788–3804 (2015)

Numerical Study of a Monolithic Fluid–Structure Formulation

Olivier Pironneau

Abstract The conservation laws of continuum mechanic are naturally written in an Eulerian frame where the difference between a fluid and a solid is only in the expression of the stress tensors, usually with Newton's hypothesis for the fluids and Helmholtz potentials of energy for hyperelastic solids. There are currently two favored approaches to Fluid Structured Interactions (FSI) both working with the equations for the solid in the initial domain; one uses an ALE formulation for the fluid and the other matches the fluid–structure interfaces using Lagrange multipliers and the immersed boundary method. By contrast the proposed formulation works in the frame of physically deformed solids and proposes a discretization where the structures have large displacements computed in the deformed domain together with the fluid in the same; in such a monolithic formulation velocities of solids and fluids are computed all at once in a single variational formulation by a semi-implicit in time and the finite element method. Besides the simplicity of the formulation the advantage is a single algorithm for a variety of problems including multi-fluids, free boundaries, and FSI. The idea is not new but the progress of mesh generators renders this approach feasible and even reasonably robust. In this article the method and its discretization are presented, stability is discussed showing in a loose fashion were are the difficulties and why one is able to show convergence of monolithic algorithms on fixed domains for fluids in compliant shell vessels restricted to small displacements. A numerical section discusses implementation issues and presents a few simple tests.

AMS Classification 65M60 (74F10 74S30 76D05 76M25).

O. Pironneau (✉)
Sorbonne Universités, UPMC (Paris VI), Laboratoire Jacques-Louis Lions, Paris, France
e-mail: olivier.pironneau@upmc.fr

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_15

1 Introduction

In an earlier paper [5] the author and his coauthors proposed a method to compute a fluid in a vessel modeled as a shell with normal displacements as in Nobile and Vergara [23]. It was argued that since the model is valid for small displacements only, one may as well use a transpiration approximation for the fluid and do the full computation in a fixed domain. As we were able to prove convergence, an interesting question arose: what is so special about the model that one could prove existence and convergence of the numerical scheme?

This paper answers partially the question: what makes FSI really hard is the moving domain. The same is true of free boundary problems for the Navier-Stokes equations. So it was the transpiration approximation for the moving part which made the analysis possible in [5].

Turning to ALE to work on a fixed domain both for the fluid and the solid is a popular solution [11], but the difficulty is transferred to the mesh [19] and the matching conditions at the fluid–solid interface [17]. Even more so with immersed boundary methods (IBM) [9, 24], although the convergence analysis is more advanced [2].

Furthermore, iterative solvers for FSI which rely on alternative solutions of the fluid and the structure parts are subject to the added mass effect and require special solvers [4, 10].

Every so often it is not a bad idea, I guess, to rethink fundamentals and check that what is taken for granted in numerical analysis is still true in the face of hardware and software progress.

So, is there an alternative to ALE and IBM? One old method [1] has resurfaced recently, the so-called *actualized Lagrangian methods* for computing structures [16, 20] (see also [8] although different from the present study because it deals mostly with membranes).

Continuum mechanics doesn't distinguish between solids and fluids till it comes to the constitutive equations. This has been exploited in many studies but most often in the context of ALE [14, 18].

In the present study we investigate what Stephan Turek [14] calls a monolithic formulation but here in an Eulerian framework, following the displaced geometry of the fluid and the solid.

To the specialist it may appear to be a back to square one idea and it is true: there is nothing new here from the modeling view-point; everyone knows that it can be done. What is new is that the mesh generators are now robust and agile at following complex motions of objects, making feasible an Eulerian numerical method.

The first difficulty with Eulerian methods comes from the hyperbolic character of the equations for the displacement of solids while those for the fluid are parabolic in time for the velocity. So let us begin by showing that a wave equation for a displacement can be reformulated as a seemingly parabolic equation for its velocity.

1.1 Preliminaries on the Wave Equation

At the core of the numerical scheme proposed here for large displacements is the following rewriting of the first and second order finite difference in time schemes for the wave equation:

$$\begin{aligned}&\partial_{tt}d - \Delta d = f,\ \forall x \in \Omega,\quad \forall t \in (0,T);\ \ d(x,t) = 0\ \forall x \in \partial\Omega;\ \forall t \in (0,T)\\ &d_{|t=0} = \partial_t d_{|t=0} = 0,\ \forall x \in \Omega.\end{aligned} \tag{1}$$

For $\alpha = 0, \theta = 1$ or $\alpha = 1, \theta \in [\frac{1}{4}, \frac{1}{2}]$ the following scheme is unconditionally stable on a uniform finite difference grid in 1D (see, for instance, [21]):

$$\frac{d_j^{n+1} - 2d_j^n + d_j^{n-1}}{\delta t^2} - \Delta_j[\theta d^{n+1} + \alpha((1-2\theta)d^n + \theta d^{n-1})] = f^n,\ \ d^0 = d^1 = 0,$$

where $\Delta_j d = (d_{j+1} - 2d_j + d_{j-1})/\delta x^2$. With $\alpha = 0$ it is first order in time; it is second order when $\alpha = 1$.

By introducing $u_j^{n+1} = \frac{1}{\delta t}(d_j^{n+1} - d_j^n)$, these schemes can be rewritten as

$$\frac{u_j^{n+1} - u_j^n}{\delta t} - \Delta_j d^n - \theta\delta t \Delta_j(u^{n+1} - \alpha u^n) = f^n,\ \ d_j^{n+1} = d_j^n + \delta t u_j^{n+1},$$

and initialized by $u^0 = d^0 = 0$. Evidently this is also unconditionally stable and first order when $\alpha = 0,\ \theta = 1$ and second order when $\alpha = 1,\ \theta \in [\frac{1}{4}, \frac{1}{2}]$.

2 General Laws of Continuum Mechanics

Consider a time dependent computational domain Ω_t made of a fluid region Ω_t^f and a solid region Ω_t^s: $\overline{\Omega}_t = \overline{\Omega}_t^f \cup \overline{\Omega}_t^s$, $\Omega_t^f \cap \Omega_t^s = \emptyset$ at all times. The fluid–structure interface is denoted $\Sigma_t = \overline{\Omega}_t^f \cap \overline{\Omega}_t^s$ and the boundary of Ω_t is $\partial\Omega_t$.

At initial time Ω_0^f and Ω_0^s are prescribed. The following notations are standard [1, 6, 14, 18, 22]:

- $\mathbf{X} : \Omega_0 \times (0,T) \mapsto \Omega_t$: $\mathbf{X}(x^0, t)$, the Lagrangian position at t of x^0.
- $\mathbf{u} = \partial_t \mathbf{X}$, the velocity of the deformation,
- $\mathbf{F}_{ji} = \partial_{x_i^0}\mathbf{X}_j$, the transposed gradient of the deformation,
- $J = \det_{\mathbf{F}}$, the Jacobian of the deformation.

Remark 1. We use a notation for the gradient which is the transposed of the one found in engineering Anglo-Saxon books (see [6] for instance). Here the gradient of a scalar being a column vector the gradient of a vector is the row of vectors of the derivative of its components.

We denote by tr_A and $\det_A$ the trace and determinant of A. As usual the following quantities are introduced:

- the density $\rho(x,t) = \mathbf{1}_{\Omega_t^f}\rho^f(x,t) + \mathbf{1}_{\Omega_t^s}\rho^s(x,t)$, at $x \in \Omega_t,\ t \in (0,T)$,
- the stress tensor $\sigma(x,t) = \mathbf{1}_{\Omega_t^f}\sigma^f(x,t) + \mathbf{1}_{\Omega_t^s}\sigma^s(x,t)$,
- $\mathbf{f}(x,t)$,$\mathbf{B}(x,t)$ the density of volumic and surfacic forces at x,t.
- $\mathbf{d} = \mathbf{X}(x^0,t) - x^0$, the displacement.

Finally and unless specified all spatial derivatives are with respect to $x \in \Omega_t$ and not with respect to $x^0 \in \Omega_0$. If ϕ is a function of $x = \mathbf{X}(x^0,t),\ x^0 \in \Omega_0$,

$$\nabla_{x^0}\phi = [\partial_{x_i^0}\phi] = [\partial_{x_i^0}\mathbf{X}_j\partial_{x_j}\phi] = \mathbf{F}^T\nabla\phi.$$

When $\mathbf{X}$ is one-to-one and invertible, $\mathbf{d}$ and $\mathbf{F}$ can be seen as functions of (x,t) instead of (x^0,t). They are related by

$$\mathbf{F}^T = \nabla_{x^0}\mathbf{X} = \nabla_{x^0}(\mathbf{d}+x^0) = \nabla_{x^0}\mathbf{d} + \mathbf{I} = \mathbf{F}^T\nabla\mathbf{d} + \mathbf{I}, \quad \Rightarrow \quad \mathbf{F} = (\mathbf{I} - \nabla\mathbf{d})^{-T}$$

Time derivatives are related by

$$D_t\phi := \frac{d}{dt}\phi(\mathbf{X}(x_0,t),t) = \partial_t\phi(x,t) + \mathbf{u}\cdot\nabla\phi(x,t).$$

It is convenient to introduce the notation

$$\mathrm{D}\mathbf{u} = \nabla\mathbf{u} + \nabla\mathbf{u}^T.$$

Conservation of momentum and conservation of mass take the same form for the fluid and the solid:

$$\rho D_t\mathbf{u} = \mathbf{f} + \nabla\cdot\sigma, \quad \frac{d}{dt}(J\rho) = 0,$$

So $J\rho = \rho_0$ at all times and

$$J^{-1}\rho_0 D_t\mathbf{u} = f + \nabla\cdot\sigma \mathrm{in}\Omega_t,\ \forall t \in (0,T), \tag{2}$$

with continuity of $\mathbf{u}$ and of $\sigma\cdot\mathbf{n}$ at the fluid–structure interface Σ when $\mathbf{B} = 0$. There are also unwritten constraints pertaining to the realizability of the map $\mathbf{X}$ (see [6, 22]). Finally incompressibility implies $J = 1$ and so $\rho = \rho_0$ constant.

2.1 Constitutive Equations

- For a Newtonian incompressible fluid : $\sigma = -p^f\mathbf{I} + \mu^f\mathrm{D}\mathbf{u}$
- For a hyperelastic incompressible material : $\sigma = -p^s\mathbf{I} + \rho^s\partial_{\mathbf{F}}\Psi\mathbf{F}^T$

where Ψ is the Helmholtz potential which, in the case of a Mooney–Rivlin two dimensional material, is [6]

$$\Psi(\mathbf{F}) = c_1 \text{tr}_{\mathbf{F}^T\mathbf{F}} + c_2(\text{tr}_{(\mathbf{F}^T\mathbf{F})^2} - \text{tr}^2_{\mathbf{F}^T\mathbf{F}}).$$

For a compressible Mooney–Rivlin material the same holds but without p^s. Here we will only consider incompressible material, but everything said below can be adapted easily.

2.2 Computation of the Mooney–Rivlin 2D Stress Tensor

It is readily seen that

$$\partial_{\mathbf{F}} \text{tr}_{\mathbf{F}^T\mathbf{F}} = ((\partial_{\mathbf{F}_{ij}} \sum_{m,n} F^2_{m,n})) = 2\mathbf{F}$$

Similarly

$$\partial_{\mathbf{F}} \text{tr}_{(\mathbf{F}^T\mathbf{F})^2} = ((\partial_{\mathbf{F}_{ij}} \sum_{n,m,p,k} F_{n,k}F_{n,m}F_{p,m}F_{p,k})) = 4\mathbf{F}\mathbf{F}^T\mathbf{F}$$

Hence

$$\Psi(\mathbf{F}) = c_1 \text{tr}_{\mathbf{F}^T\mathbf{F}} + c_2(\text{tr}_{(\mathbf{F}^T\mathbf{F})^2} - \text{tr}^2_{\mathbf{F}^T\mathbf{F}}) \Rightarrow \partial_{\mathbf{F}}\Psi = 2c_1\mathbf{F} + c_2(4\mathbf{F}\mathbf{F}^T\mathbf{F} - 4\text{tr}_{\mathbf{F}^T\mathbf{F}}\mathbf{F})$$

Let $\mathbf{B} := \mathbf{F}\mathbf{F}^T = \left((\mathbf{I} - \nabla\mathbf{d})(\mathbf{I} - \nabla\mathbf{d})^T\right)^{-1}$, $b :=\text{det}_{\mathbf{B}}$, $c := \text{tr}_{\mathbf{B}} = \text{tr}_{\mathbf{F}^T\mathbf{F}}$. Then

$$\partial_{\mathbf{F}}\Psi\mathbf{F}^T = (2c_1 - 4c_2c)\mathbf{B} + 4c_2\mathbf{B}^2.$$

Now by the Cayley-Hamilton theorem $\mathbf{B}^2 = c\mathbf{B} - b\mathbf{I}$ so

$$\partial_{\mathbf{F}}\Psi\mathbf{F}^T = 2c_1\mathbf{B} - 4c_2b\mathbf{I} = 2c_1\mathbf{F}\mathbf{F}^T - 4c_2\text{det}_{\mathbf{F}\mathbf{F}^T}\mathbf{I}$$

By the Cayley-Hamilton theorem again

$$\mathbf{B} = c\mathbf{I} - b\mathbf{B}^{-1} = c\mathbf{I} - b(\mathbf{I} - \nabla\mathbf{d} - \nabla\mathbf{d}^T + \nabla\mathbf{d}\nabla\mathbf{d}^T).$$

So

$$\begin{aligned}\partial_{\mathbf{F}}\Psi\mathbf{F}^T &= (2c_1(c-b) - 4c_2b)\mathbf{I} + 2c_1b(\text{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T)\\ &= 2c_1\text{det}_{\mathbf{F}\mathbf{F}^T}(\text{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) + (2c_1\text{tr}_{\mathbf{F}\mathbf{F}^T} - (2c_1 + 4c_2)\text{det}_{\mathbf{F}\mathbf{F}^T})\mathbf{I}\end{aligned}$$

Hence an incompressible two dimensional Mooney–Rivlin material will have, for some α,

$$\partial_{\mathbf{F}}\Psi\mathbf{F}^T = 2c_1(\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) + \alpha\mathbf{I}.$$

3 Monolithic Variational Formulation in 2D

Let $\Gamma \subset \partial\Omega$ the part of the boundary on which the solid is clamped or the fluid has a no-slip condition. For incompressible material the final fluid–structure formulation in two dimensions is

Find $(\mathbf{u}, p)$ with $\mathbf{u}_{|\Gamma} = 0$, $\mathbf{d}$ and $\Omega_t^r,\ r = s,f$ solutions of

$$\begin{aligned}\int_{\Omega_t}\Big[\rho D_t\mathbf{u}\cdot\hat{\mathbf{u}} - p\nabla\cdot\hat{\mathbf{u}} - \hat{p}\nabla\cdot\mathbf{u} + \mathbf{1}_{\Omega_t^f}\frac{\nu}{2}\mathrm{D}\mathbf{u} : \mathrm{D}\hat{\mathbf{u}}\\ + \mathbf{1}_{\Omega_t^s}\tilde{c}_1(\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) : \mathrm{D}\hat{\mathbf{u}}\Big] = \int_{\Omega_t} f\cdot\hat{\mathbf{u}}\\ \mathrm{D}_t\mathbf{d} = \mathbf{u},\end{aligned} \tag{3}$$

for all $(\hat{\mathbf{u}}, \hat{p})$ with $\hat{\mathbf{u}}_{|\Gamma} = 0$, where Ω_t^s and Ω_t^f are defined incrementally by

$$\frac{dX}{\mathrm{d}\tau} = \mathbf{u}(X(\tau), \tau),\ \ X(t) \in \Omega_t^r \ \Rightarrow\ X(\tau) \in \Omega_\tau^r\ \ \forall\tau \in (0, T),\ r = s,f$$

Initial conditions are: $\mathbf{u}$ given, $\mathbf{d} = 0$, Ω_0^r given, $r = s,f$.

We have used the notation $\mathbf{B} : \mathbf{C} = \mathrm{tr}_{\mathbf{B}^T\mathbf{C}}$ and $\tilde{c}_1 := \rho^s c_1$.

3.1 Conservation of Energy

Proposition 1.

$$\frac{d}{dt}\int_{\Omega_t}\frac{\rho}{2}|\mathbf{u}|^2 + \frac{\nu}{2}\int_{\Omega_t^f}|\mathrm{D}\mathbf{u}|^2 + \frac{d}{dt}\int_{\Omega_0^s}\Psi(\mathbf{I} + \nabla_{x^0}\mathbf{d}^T) = \int_{\Omega_t} f\cdot\mathbf{u}$$

When Ψ is convex, an existence of solution result can be gained from this equality (see [12] for example).

Proof. Choosing $\hat{\mathbf{u}} = \mathbf{u}$, $\hat{p} = -p$ will give the proposition provided

$$\int_{\Omega_t^s}(\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) : \mathrm{D}\partial_t\mathbf{d} = \frac{d}{dt}\int_{\Omega_0^s}\Psi(\nabla_{x^0}\mathbf{X}).$$

By construction

$$\int_{\Omega_t^s} c_0(\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) : \mathrm{D}\hat{\mathbf{u}} = \int_{\Omega_t^s} (\partial_{\mathbf{F}}\Psi(\mathbf{F})\mathbf{F}^T - \alpha\mathbf{I}) : \mathrm{D}\hat{\mathbf{u}} = \int_{\Omega_0^s} \partial_{\mathbf{F}}\Psi(\mathbf{F}) : \mathrm{D}_{x^0}\hat{\mathbf{u}}$$

Now as $\frac{d}{dt}\Psi(\mathbf{F}) = \partial_{\mathbf{F}}\Psi(\mathbf{F}) : \partial_t\mathbf{F}$ and $\nabla_{x^0}\mathbf{u} = \partial_t\nabla_{x^0}\mathbf{d} = \partial_t\mathbf{F}^T$,

$$\int_{\Omega_t^s} c_0(\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) : \mathrm{D}\mathbf{u} = \int_{\Omega_0^s} \frac{d}{dt}\Psi(\mathbf{F}) = \frac{d}{dt}\int_{\Omega_0^s} \Psi(\mathbf{I} + \nabla_{x^0}\mathbf{d}^T).$$

3.2 Discretization in Time

It is natural to use the following discretization

$$\int_{\Omega_t} (\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) : \mathrm{D}\hat{\mathbf{u}} \approx \int_{\Omega_n} (\mathrm{D}\mathbf{d}^{n+1} - \nabla\mathbf{d}^{n+1}\nabla\mathbf{d}^{n+1^T}) : \mathrm{D}\hat{\mathbf{u}}$$

with $\mathbf{d}^{n+1} = \mathbf{d}^n + \delta t\mathbf{u}^{n+1}$. Hence

$$\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T \approx \mathrm{D}\mathbf{d}^n - \nabla\mathbf{d}^n\nabla\mathbf{d}^{n^T} + \delta t(\mathrm{D}\mathbf{u}^{n+1} - \nabla\mathbf{u}^{n+1}\nabla\mathbf{d}^{n^T} - \nabla\mathbf{d}^n\nabla\mathbf{u}^{n+1^T}) + o(\delta t).$$

So if X^n is a first order approximation of $X(t^{n+1} - \delta t)$ defined by

$$\dot{X} = \mathbf{u}(X(\tau), \tau), \; X(t^{n+1}) = x$$

such as $X^n(x) = x - \delta t\mathbf{u}^n(x)$, a consistent first order scheme is to find $\mathbf{u}^{n+1}, p^{n+1}$ such that $\mathbf{u}^{n+1} = 0$ on Γ and for all $\hat{\mathbf{u}}, \hat{p}$ with $\hat{\mathbf{u}}_{|\Gamma} = 0$,

$$\begin{aligned}
&\int_{\Omega_n} \Big[\rho^n \frac{\mathbf{u}^{n+1} - \mathbf{u}^n \circ X^n}{\delta t} \cdot \hat{\mathbf{u}} - p^{n+1}\nabla \cdot \hat{\mathbf{u}} - \hat{p}\nabla \cdot \mathbf{u}^{n+1} + \mathbf{1}_{\Omega_n^f}\frac{\nu}{2}\mathrm{D}\mathbf{u}^{n+1} : \mathrm{D}\hat{\mathbf{u}} \\
&+\tilde{c}_1\mathbf{1}_{\Omega_n^s}[\mathrm{D}\mathbf{d}^n - \nabla\mathbf{d}^{n^T}\nabla\mathbf{d}^n + \delta t(\mathrm{D}\mathbf{u}^{n+1} - \nabla\mathbf{u}^{n+1}\nabla\mathbf{d}^{n^T} - \nabla\mathbf{d}^n\nabla\mathbf{u}^{n+1^T})] : \mathrm{D}\hat{\mathbf{u}}\Big] \\
&= \int_{\Omega_n} f\hat{\mathbf{u}}, \\
&\mathbf{d}^{n+1} = \mathbf{d}^n \circ X^n + \delta t\mathbf{u}^{n+1}, \; \Omega_{n+1}^r = \{x + \delta t\mathbf{u}^{n+1}(x) : x \in \Omega_n^r\}, \; r = s,f \qquad (4)
\end{aligned}$$

3.3 Spatial Discretization with Finite Elements

Let $\mathscr{T}_h^0$ be a triangulation of the initial domain. Spatial discretization can be done with Lagrangian triangular elements of degree 2 for the space V_h of velocities and displacements and Lagrangian triangular elements of degree 1 for the pressure space Q_h. A small penalization with parameter ϵ must be added to impose uniqueness of the pressure when a direct linear solver is used.

At each time step one must find $\mathbf{u}_h^{n+1}, p_h^{n+1} : \forall \hat{\mathbf{u}}_h \in V_{0h}, \forall \hat{p}_h \in Q_h$

$$\int_{\Omega_n} \Big[\rho^n \frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^n \circ X^n}{\delta t} \cdot \hat{\mathbf{u}}_h - p_h^{n+1} \nabla \cdot \hat{\mathbf{u}}_h - \hat{p}_h \nabla \cdot \mathbf{u}_h^{n+1} + \mathbf{1}_{\Omega_n^f} \frac{\nu}{2} \mathrm{D}\mathbf{u}_h^{n+1} : \mathrm{D}\hat{\mathbf{u}}_h$$
$$+ \tilde{c}_1 \mathbf{1}_{\Omega_n^s} [\mathrm{D}\mathbf{d}_h^n - \nabla \mathbf{d}^{n^T} \nabla \mathbf{d}_h^n + \delta t (\mathrm{D}\mathbf{u}_h^{n+1} - \nabla \mathbf{u}_h^{n+1} \nabla \mathbf{d}_h^{nT} - \nabla \mathbf{d}_h^n \nabla \mathbf{u}_h^{n+1^T})] : \mathrm{D}\hat{\mathbf{u}}_h$$
$$+ \epsilon p_h \hat{p}_h \Big] = \int_{\Omega_n} f \hat{\mathbf{u}}_h, \tag{5}$$

Then the triangulation must be updated by $\mathbf{u}_h^{n+1}$ by moving each vertex from q_i^n to $q_i^{n+1} := q_i^n + \delta t \mathbf{u}_h^{n+1}$.

Let $\mathbf{d}_i^n := \mathbf{d}^n(q_i)$ then

$$\mathbf{d}^n \circ X^n(q_i^{n+1}) = \mathbf{d}^n(q_i^n + \delta t \mathbf{u}_h^{n+1} - \delta t \mathbf{u}_h^{n+1}) = \mathbf{d}^n(q_i^n)$$

so to implement $\mathbf{d}_h^{n+1} = \mathbf{d}_h^n \circ X^n + \delta t \mathbf{u}_h^{n+1}$ it suffices to copy the array of values of $\mathbf{d}_h^n$ in the array of values of $\mathbf{d}_h^{n+1}$ and add $\delta t \mathbf{u}_h^{n+1}(q_i^n)$ to the array.

The vertices in the fluid are moved by $\tilde{\mathbf{u}}$ solution of $-\Delta \tilde{\mathbf{u}} = 0$ in the fluid and $\tilde{\mathbf{u}} = \mathbf{u}$ on Σ and zero on other boundaries. Moving the vertices of $\mathscr{T}_h^n$ gives a new triangulation $\mathscr{T}_h^{n+1}$.

3.4 Map Preserving Scheme

It is important to understand on which triangulation each variable is defined. One possibility is to assume that Ω_{n+1} is constructed by successive iterations such that $\Omega_n = \{x - \delta t \mathbf{u}_h^{n+1}(x),\ x \in \Omega_{n+1}\}$. Then $\mathbf{X}^{n+1}(x_0)$ is such that

$$\mathbf{X}^{n+1}(x_0) = \mathbf{X}^n(x_0) + \delta t \mathbf{u}_h^{n+1}(\mathbf{X}^{n+1}(x_0))$$

So $\mathbf{u}_h^{n+1}$ and $\mathbf{d}_h^{n+1}$ live on $\mathscr{T}_h^{n+1}$ and $\mathbf{u}_h^n$ and $\mathbf{d}_h^n$ live on $\mathscr{T}_h^n$.

Recall the notation $X^{n+1}(x) = x - \mathbf{u}^{n+1}(x) + o(\delta t)$, not to be confused with $\mathbf{X}^{n+1}$. Let $x = \mathbf{X}^{n+1}(x_0) \in \Omega_{n+1}$, then $X^{n+1}(x) \in \Omega_n$; let $\mathbf{F}^n(x) = (\nabla_{x_0} \mathbf{X}^n(x_0))^T$. Then

$$\nabla_{x_0} \mathbf{X}^{n+1}(x_0) = \nabla_{x_0} \mathbf{X}^n(x_0) + \delta t \nabla_{x_0} \mathbf{X}^{n+1}(x_0) \nabla \mathbf{u}_h^{n+1}(x), \tag{6}$$

Hence $\mathbf{F}^{n+1} = [\mathbf{I} - \delta t \nabla \mathbf{u}_h^{n+1}]^{-T} \mathbf{F}^n \circ X^{n+1}$ and

$$\mathbf{F}^{n+1} \mathbf{F}^{n+1^T} = [\mathbf{I} - \delta t \nabla \mathbf{u}_h^{n+1}]^{-T} (\mathbf{F}^n \mathbf{F}^{n^T}) \circ X^{n+1} [\mathbf{I} - \delta t \nabla \mathbf{u}_h^{n+1}(x)]^{-1} \tag{7}$$

As above, the Cayley-Hamilton theorem and incompressibility imply that

$$[\mathbf{I} - \delta t \nabla \mathbf{u}_h^{n+1}]^{-1} = \mathbf{I} + \delta t \nabla \mathbf{u}_h^{n+1} + o(\delta t)$$

Hence $\mathbf{B}^n(x) = \mathbf{F}^n(x)\mathbf{F}^{nT}(x)$ satisfies

$$\mathbf{B}^{n+1} = [\mathbf{I} + \delta t \nabla \mathbf{u}_h^{n+1^T}]\mathbf{B}^n \circ X^{n+1}[\mathbf{I} + \delta t \nabla \mathbf{u}_h^{n+1}] + o(\delta t) \tag{8}$$

This formula can be used in the variational formulation instead of $\nabla \mathbf{d}$. Its numerical performance is similar to the one that uses $\mathbf{d}$ but it preserves energy… at the cost of an iterative adjustment of Ω_{n+1}.

Remark 2. When Ω^s_{n+1} is defined by: $\Omega^s_{n+1} = \{x + \delta t \mathbf{u}^{n+1}(x) \ : \ x \in \Omega^s_n\}$ and all integrals are computed on Ω_n, we can also compute a map preserving $\mathbf{F}$ because

$$\mathbf{X}^{n+1}(x) = \mathbf{X}^n(x) + \delta t \mathbf{u}_h^{n+1}(x) \ \Rightarrow \ \nabla_{x^0}\mathbf{X}^{n+1} = \nabla_{x^0}\mathbf{X}^n + \delta t \nabla_{x^0}\mathbf{X}^n \nabla \mathbf{u}_h^{n+1}$$

which implies that $\mathbf{F}^{n+1} = (\mathbf{I} + \delta t \nabla \mathbf{u}_h^{n+1^T})\mathbf{F}^n$. However this defines $\mathbf{X}^{n+1}$ on Ω^s_n and something must be added to the algorithm to project $\mathbf{X}^{n+1}$ on Ω^s_{n+1}.

3.5 Remark on the Construction of Second Order Accurate Schemes

To build a second order scheme for (3) one would have to

- use a second order characteristic method [3, 25] for $D_t\mathbf{u}$,
- approximate $\mathbf{u}$ by $(\mathbf{u}^{n+1} + \mathbf{u}^n)/2$ in the fluid viscous term
- and $\mathbf{d}$ by $(\mathbf{d}^{n+1} + \mathbf{d}^{n-1})/2 = \mathbf{d}^n + \delta t(\mathbf{u}^{n+1} - \mathbf{u}^n)/2$.
- But one must also approximate Ω_t by $\Omega_{n+\frac{1}{2}}$.

The last item is difficult; the construction of second order schemes is discussed in particular in the thesis of Hauret [12] with the Newmark mid-point scheme [15]. It seems rather difficult to prove the same here, so we postpone it to a future study.

3.6 Perturbation About an Equilibrium

An equilibrium is reached when $p, \mathbf{d}$ and $\Omega^s = \{x^0 + \mathbf{d}(x^0) \ : \ x^0 \in \Omega^s_0\}$ are such that

$$\int_\Omega \Big[-p\nabla\cdot\hat{\mathbf{u}} + \tilde{c}_1 \mathbf{1}_{\Omega^s}[\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T] : \mathrm{D}\hat{\mathbf{u}} \Big] = \int_\Omega f\hat{\mathbf{u}}, \ \forall\hat{\mathbf{u}}, \tag{9}$$

Assuming for clarity that $\rho^s = \rho^f$ and Ω is independent of t, if we prime all variations about that state, they much be such that $\partial_t\mathbf{d}' = \mathbf{u}'$ and

$$\begin{aligned}&\int_\Omega \Big[\rho D_t\mathbf{u}'\cdot\hat{\mathbf{u}} - p'\nabla\cdot\hat{\mathbf{v}} + \hat{p}\nabla\cdot\mathbf{u}'\\ &+\mathbf{1}_{\Omega^f}\frac{\nu}{2}\mathrm{D}\mathbf{u}' : \mathrm{D}\hat{\mathbf{u}} + \tilde{c}_1\mathbf{1}_{\Omega^s}(\mathrm{D}\mathbf{d}' - \nabla\mathbf{d}'\nabla\mathbf{d}^T - \nabla\mathbf{d}\nabla\mathbf{d}'^T) : \mathrm{D}\hat{\mathbf{u}}\Big]\\ &+\tilde{c}_1\int_\Sigma [\mathbf{u}'\cdot\mathbf{n}(\mathrm{D}\mathbf{d} - \nabla\mathbf{d}\nabla\mathbf{d}^T) : \mathrm{D}\hat{\mathbf{u}}] = \int_\Omega f'\hat{\mathbf{u}},\end{aligned} \tag{10}$$

because, provided Σ is smooth, $\left(\int_{\Omega_t^s}\phi\right)' = \int_{\Omega^s}\phi' + \int_\Sigma \phi\mathbf{u}'\cdot\mathbf{n}$. On this formulation we believe that it is not difficult to prove existence, uniqueness, and convergence of the time scheme as in [5].

4 On the Stability of the Scheme

For clarity we drop the subscript h.

Consider the scheme based on (8) and iterated so as to replace Ω_n by Ω_{n+1} Then

$$\begin{aligned}&\int_\Omega f\hat{\mathbf{u}} = \int_\Omega \Big[[\rho^n\frac{\mathbf{u}^{n+1} - \mathbf{u}^n\circ X^n}{\delta t} - p^{n+1}\nabla\cdot\hat{\mathbf{u}} - \hat{p}\nabla\cdot\mathbf{u}^{n+1}] : \hat{\mathbf{u}}\\ &+\mathbf{1}_{\Omega^f_{n+1}}\frac{\nu}{2}\mathrm{D}\mathbf{u}^{n+1} : \mathrm{D}\hat{\mathbf{u}} + \tilde{c}_1\mathbf{1}_{\Omega^s_{n+1}}\mathbf{F}^{n+1}\mathbf{F}^{n+1^T} : \mathrm{D}\hat{\mathbf{u}}\Big]\end{aligned} \tag{11}$$

Now suppose for clarity that $\rho^f = \rho^s$ and $\mathbf{f} = 0$; the general case will be analyzed later. Choosing $\hat{p} = -p^{n+1}$, $\hat{\mathbf{u}} = \mathbf{u}^{n+1}$ and assuming Ω independent of t,

$$\begin{aligned}0 = \int_\Omega &\Big[\rho\frac{1}{\delta t}(|\mathbf{u}^{n+1}|^2 - \mathbf{u}^n\circ X^n\cdot\mathbf{u}^{n+1}) + \mathbf{1}_{\Omega^f_{n+1}}\frac{\nu}{2}|\mathrm{D}\mathbf{u}^{n+1}|^2\\ &+\frac{\tilde{c}_1}{\delta t}\mathbf{1}_{\Omega^s_{n+1}}\mathbf{F}^{n+1}\mathbf{F}^{n+1^T} : \mathrm{D}(\mathbf{d}^{n+1} - \mathbf{d}^n\circ X^n)\Big]\end{aligned} \tag{12}$$

Lemma 1.

$$2(\mathbf{u}^{n+1} - \mathbf{u}^n\circ X^n)\cdot\mathbf{u}^{n+1} = |\mathbf{u}^{n+1}|^2 - |\mathbf{u}^n\circ X^n|^2 + |\mathbf{u}^{n+1} - \mathbf{u}^n\circ X^n|^2 \geq |\mathbf{u}^{n+1}|^2 - |\mathbf{u}^n|^2$$

Now let us analyze the last term in (12). As $\nabla\cdot\mathbf{d} = 0$,

$$\tilde{c}_1\int_{\Omega^s}\mathbf{F}^{n+1}\mathbf{F}^{n+1^T} : \mathrm{D}(\mathbf{d}^{n+1} - \mathbf{d}^n) = \int_{\Omega_0^s}\partial_{\mathbf{F}}\Psi(\mathbf{F}^{n+1}) : (\mathbf{F}^{n+1} - \mathbf{F}^n) \tag{13}$$

By the convexity of Ψ

$$\int_{\Omega_0^s}\Psi(\mathbf{F}^{n+1}) - \int_{\Omega_0^s}\Psi(\mathbf{F}^n) \leq \int_{\Omega_0^s}\partial_{\mathbf{F}}\Psi(\mathbf{F}^{n+1}) : (\mathbf{F}^{n+1} - \mathbf{F}^n)$$

Finally

$$\frac{1}{2}\|\mathbf{u}^{n+1}\|^2_{\Omega} + \frac{\nu\delta t}{2}\|\mathrm{D}\mathbf{u}^{n+1}\|^2_{\Omega^f} + \int_{\Omega^s_0}\Psi(\mathbf{F}^{n+1}) \le \frac{1}{2}\|\mathbf{u}^n\|^2_{\Omega} + \int_{\Omega^s_0}\Psi(\mathbf{F}^n) \tag{14}$$

4.1 Case : $\rho^s \neq \rho^f$

Consider Ω^r_t, $r = s$ or f and

$$\mathcal{N}(u,\hat{u}) := \int_{\Omega^r_t}\rho^r[\partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u}]\cdot\hat{\mathbf{u}} \approx \rho^r\int_{\Omega^r_n}[\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\delta t} + \mathbf{u}^{n+1}\cdot\nabla\mathbf{u}^{n+1}]\cdot\hat{\mathbf{u}} \tag{15}$$

Let $\hat{\mathbf{u}} = \mathbf{u}^{n+1}$, recall that $\int_{\Omega^r_n}(|\mathbf{u}^{n+1}|^2 - \mathbf{u}^n\cdot\mathbf{u}^{n+1}) \ge \int_{\Omega^r_n}\frac{1}{2}(|\mathbf{u}^{n+1}|^2 - |\mathbf{u}^n|^2)$. Now when $\nabla\cdot\mathbf{u}^{n+1} = 0$,

$$2\int_{\Omega^r_n}\mathbf{u}^{n+1}\cdot\nabla\mathbf{u}^{n+1}\cdot\mathbf{u}^{n+1} = \int_{\Omega^r_n}\mathbf{u}^{n+1}\cdot\nabla|\mathbf{u}^{n+1}|^2 = -\int_{\Omega^r_n}\nabla\cdot(\mathbf{u}^{n+1})|\mathbf{u}^{n+1}|^2$$
$$+\int_{\partial\Omega^r_n}|\mathbf{u}^{n+1}|^2\mathbf{u}^{n+1}\cdot\mathbf{n} = \frac{1}{\delta t}\Big[\int_{\Omega^r_{n+1}}|\mathbf{u}^{n+1}|^2 - \int_{\Omega^r_n}|\mathbf{u}^{n+1}|^2\Big] + o(\delta t)$$

Hence $\mathcal{N}(\mathbf{u}^{n+1},\mathbf{u}^{n+1}) \ge \frac{1}{2\delta t}\Big[\int_{\Omega^r_{n+1}}\rho^r\mathbf{u}^{n+1}|^2 - \int_{\Omega^r_n}\rho^r|\mathbf{u}^n|^2\Big]$. Consequently

$$\int_{\Omega_n}\rho[\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\delta t} + \mathbf{u}^{n+1}\cdot\nabla\mathbf{u}^{n+1}]\cdot\mathbf{u}^{n+1} \ge \frac{1}{2\delta t}\Big[\int_{\Omega_{n+1}}\rho^{n+1}\mathbf{u}^{n+1}|^2 - \int_{\Omega_n}\rho^n|\mathbf{u}^n|^2\Big]$$

5 Formulation in the Initial Domain for the Solid

We wish to compare the formulation in the moving domain with the formulation in the fixed domain for a single hyperelastic incompressible structure. Recall that the deformation map satisfies

$$\partial_{tt}\mathbf{X} - \nabla_{x^0}\cdot((-p^s\mathbf{I} + \partial_{\mathbf{F}}\Psi\mathbf{F}^T)J\mathbf{F}^{-T}) = \tilde{\mathbf{f}} := \frac{1}{\rho^s}\mathbf{f},\quad J := \det\nabla_{x^0}\mathbf{X} = 1,$$
$$\partial_{\mathbf{F}}\Psi\mathbf{F}^T = 2\tilde{c}_1\mathbf{F}\mathbf{F}^T + \alpha\mathbf{I} = 2\tilde{c}_1(\mathrm{D}_{x^0}\mathbf{d} + \nabla_{x^0}\mathbf{d}^T\nabla_{x^0}\mathbf{d}) + \tilde{\alpha}\mathbf{I},\quad \partial_{\mathbf{F}}\Psi = 2\tilde{c}_1\nabla\mathbf{d}^T + \tilde{\alpha}\mathbf{F}^{-T}$$

Integrated on the initial domain and assuming no additional surface constraint on $\partial\Omega_0$, the variational formulation is

$$\int_{\Omega_0^s}\Big[(\partial_{tt}\mathbf{d})\cdot\hat{\mathbf{d}} + (-pJ\mathbf{I}\mathbf{F}^{-T} + 2\tilde{c}_1\nabla_{x^0}\mathbf{d} : \nabla_{x^0}\hat{\mathbf{d}} + (J-1)\hat{p}\Big] = \int_{\Omega_0^s}\tilde{\mathbf{f}}\cdot\hat{\mathbf{d}}.$$

for all $\hat{\mathbf{d}}$ zero on Γ. A fully implicit numerical scheme is [12, 14]:

$$\begin{aligned}\int_{\Omega_0}\Big[&\frac{\mathbf{d}^{n+1}-2\mathbf{d}^n+\mathbf{d}^{n-1}}{\delta t^2}\cdot\hat{\mathbf{d}} + 2c_1\nabla_{x^0}\tilde{\mathbf{d}} : \nabla_{x^0}\hat{\mathbf{d}}\\ &-p^{n+1}J^{n+1}\mathbf{F}^{n+1^{-T}} : \nabla_{x^0}\hat{\mathbf{d}} + (J^{n+1}-1)\hat{p}\Big] = \int_{\Omega_0}\tilde{\mathbf{f}}\cdot\hat{\mathbf{d}}\end{aligned}$$

with $\tilde{\mathbf{d}} = \frac{1}{2}(\theta\mathbf{d}^{n+1} + (2-\theta)\mathbf{d}^{n-1})$ and $\theta = 1$ or 2 and with $\mathbf{d}^{n+1} = \mathbf{d}^n + \delta t\mathbf{u}^{n+1}$. Neglecting the terms in δt^2 leads to

$$\begin{aligned}&\int_{\Omega_0}\Big[\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\delta t}\cdot\hat{\mathbf{u}} + \delta tc_1\nabla_{x^0}(\theta\mathbf{u}^{n+1}-(2-\theta)\mathbf{u}^n) : \nabla_{x^0}\hat{\mathbf{u}} + 2c_1\nabla_{x^0}\mathbf{d}^n : \nabla_{x^0}\hat{\mathbf{u}}\\ &-p^{n+1}[\mathscr{F}(1,\mathbf{d}^n) + \delta t\mathscr{F}(0,\mathbf{u}^{n+1})] : \nabla_{x^0}\hat{\mathbf{u}} + (\det_{\mathscr{F}(1,\mathbf{d}^n)}-1)\hat{p}\\ &+\delta t(\partial_{x^0}d_1\partial_{y^0}u_2 - \partial_{x^0}d_2\partial_{y^0}u_1 + \partial_{x^0}d_1\partial_{y^0}u_2 - \partial_{x^0}u_2\partial_{y^0}d_1 + \partial_{x^0}u_1 + \partial_{y^0}u_2)^{n+1}\hat{p}\Big]\\ &= \int_{\Omega_0}\tilde{\mathbf{f}}^n\cdot\hat{\mathbf{u}} \qquad \text{where}\mathscr{F}(\alpha,\mathbf{d}) = \begin{pmatrix}\alpha+\partial_{x^0}d_1, -\partial_{x^0}d_2\\ -\partial_{y^0}d_1, \alpha+\partial_{y^0}d_2\end{pmatrix}\end{aligned}$$

Remark 3. To be consistent with the formulation in the moving domain one ought to have written $\nabla\cdot u = 0$, instead of $J = 1$, i.e., $(F^{-T}\nabla_{x^0})\cdot\mathbf{u} = 0$ or equivalently $J(F^{-T}\nabla_{x^0})\cdot\mathbf{u} = 0$, i.e.,

$$\int_{\Omega_0}\Big[\hat{p}\mathscr{F}(1,\mathbf{d}^n) : \nabla_{x^0}\mathbf{u}^{n+1}\Big] = 0, \quad \forall\hat{p},$$

leading to the formulation

$$\begin{aligned}&\int_{\Omega_0}\Big[\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\delta t}\cdot\hat{\mathbf{u}} + \delta tc_1\nabla_{x^0}(\theta\mathbf{u}^{n+1}-(2-\theta)\mathbf{u}^n) : \nabla_{x^0}\hat{\mathbf{u}} + 2c_1\nabla_{x^0}\mathbf{d}^n : \nabla_{x^0}\hat{\mathbf{u}}\\ &-p^{n+1}\mathscr{F}(1,\mathbf{d}^n) : \nabla_{x^0}\mathbf{u}^{n+1} - \hat{p}\mathscr{F}(1,\mathbf{d}^n) : \nabla_{x^0}\mathbf{u}^{n+1}\Big] = \int_{\Omega_0}\tilde{\mathbf{f}}^n\cdot\hat{\mathbf{u}}\end{aligned} \tag{16}$$

Remark 4. The following identity is true for all vector valued functions in $H_0^1(\Omega)^2$ (see Costabel et al. [7]):

$$\int_\Omega\nabla u : \nabla v = \int_\Omega[\frac{1}{2}(\nabla u + \nabla u^T) : (\nabla v + \nabla v^T) - \nabla\cdot u\nabla\cdot v] \tag{17}$$

It changes the boundary condition on $\partial\Omega\backslash\Gamma$ to use it in (16), i.e.,

$$2c_1\nabla_{x^0}\mathbf{d}^n : \nabla_{x^0}\hat{\mathbf{u}} \text{ changed to } \mathrm{D}_{x^0}\mathbf{d}^n : \mathrm{D}_{x^0}\hat{\mathbf{u}} - 2c_1\nabla_{x^0}\cdot\mathbf{d}^n\nabla_{x^0}\cdot\hat{\mathbf{u}}$$

leads to a new boundary condition. The Mooney–Rivlin model may be degenerate in 2D for incompressible material or the boundary condition implied by this formulation may be more appropriate. This point needs to be studied further.

6 Numerical Tests

6.1 Comparisons for an Incompressible Beam

The monolithic method set in the moving domain is compared with the more classical method (16) set in the initial domain in the case of a single rectangular elastic beam of length to width ratio equal to 10 and bent by its own weight from rest and clamped either on both side or on one side.

The spatial discretization is performed by using Lagrangian Triangular Finite Elements with polynomials of degree 2 for the velocities and displacements and degree 1 for the pressures. `FreeFem++` [13] has been used to implement the algorithms. All linear systems are solved by direct solvers of the library `MUMPS`.

The method in the fixed initial domain did not seem to work with (16) but it did with (17); we have used the second order approximation, $\theta = 1$.

The penalization parameter is $\epsilon = 0.01$ for the formulation in the initial domain and 10^{-6} for the one in the moving domain (Fig. 1).

The other parameters are

$$E = 2.15,\ \sigma = 0.29,\ \mu = \frac{E}{2(1+\sigma)},\ \rho^s = 1,\ c_1 = \frac{\mu}{2} = 0.417,\ T = 45,\ \delta t = 1.$$

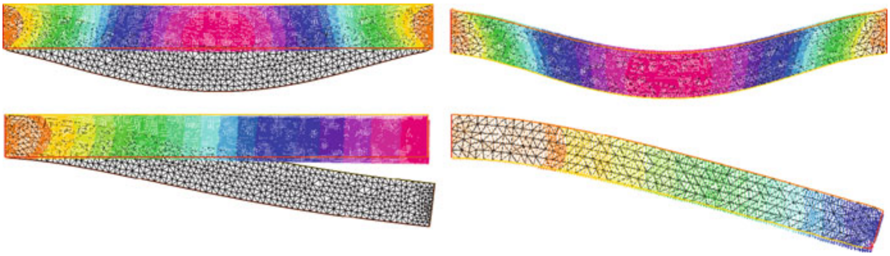


Fig. 1 *Top*: Maximum bent under its own weight for the formulation in the initial domain (*left*) and for the formulation in the moving domain (*right*). *Bottom*: Free fall under its own weight of the same solid clamped on the left only; position at time 50 computed in the initial domain (*left*) and the moving domain (*right*). For the full swing see Fig. 3

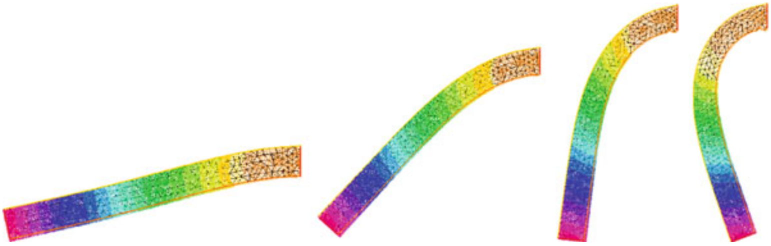


Fig. 2 Free fall under its own weight of the same hyperelastic solid clamped on the left only; position at iteration 50, 100, 150, 200

The gravity differs :

1. when the beam is clamped at both sides, it is set to $f = -0.02$,
2. and when the beam is clamped on the left only $f = -0.002$.

The beam clamped at both side went through one and a half oscillation cycle during the 45 time steps while the one clamped on the left only went through a half cycle. With the method using the initial domain there is a slight change of surface area, going from 10 initially to 10.57 after 45 time iterations while the change is less than 1 % for the other one.

6.2 Fall of a Hyperelastic Beam Clamped on One Side

The same beam clamped on one side only is allowed to fall under its own weight for 200 iterations. The results obtained by the method set in the moving domain are shown in Fig. 2. All parameters have the same value except the time step which is 0.5.

7 Monolithic Fluid–Structure Interaction

Now the fall of a similar hyperelastic incompressible beam is studied in a rectangular box filled with a fluid. As the beam is clamped on the right and free to fall under its own weight in so doing it compels the fluid to move. The results are shown in Fig. 3. The beam is a rectangle 8 × 1 initially. Its Mooney–Rivlin coefficient is $c_1 = 0.2$. Its density is 1 and the gravity force is -0.2. The fluid has $\rho^f = 0.5$ and $\nu = 0.1$. The computation stops at$T = 30$ after 300 time steps. In the fluid part an auxiliary Laplace equation is solved to move the inner vertices at each time step:

$$-\Delta \mathbf{v} = 0, \text{in} \Omega_t^f \quad \mathbf{v}_{|\Sigma} = \mathbf{u}, \quad \mathbf{v}_{|\partial\Omega\setminus\Sigma} = 0.$$

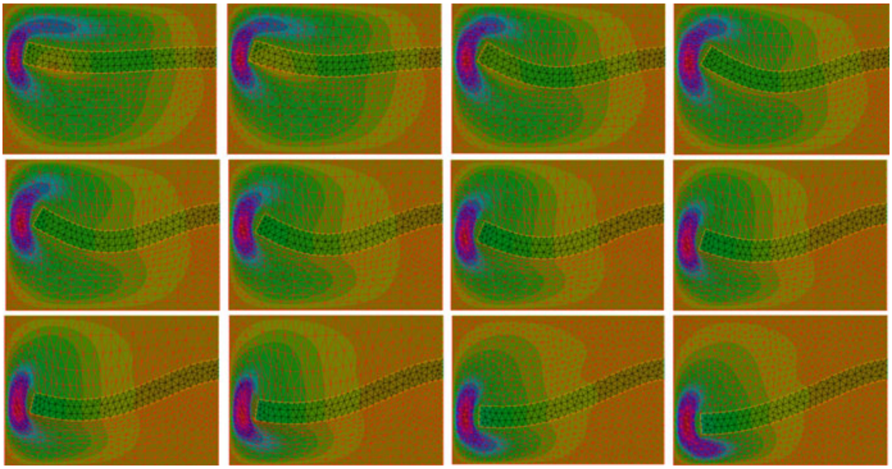


Fig. 3 Free fall of the same beam, clamped on the right, in a fluid initially at rest. Every 10th step the geometry and the norm of velocities are shown. At the 110th time step the mesh is unusable, so `adaptmesh()` is called

The vertices of the mesh in the solid part are moved by $\mathbf{u}^{n+1}\delta t$ by calling `movemesh()`, the mesh moving function of `FreeFem++`.

The mesh is moved by calling the `FreeFem++` function `movemesh()`. If it flips over a triangle, the function `adaptmesh()` is called. The second example is the free motion of a hyperelastic incompressible solid submitted to its own weight and to the force due to the rotation of the fluid induced by the sliding of the lower horizontal boundary at unit speed. Initially the solid is a disk. The fluid domain is a rectangle of size 10 × 7 and $\nu = 0.1$, $\rho^f = 0.5$ while $\rho^s = 1$. The gravity is -0.2 (Fig. 4).

8 Free Surface Flow with the Same Code

A rotating fluid driven by the lower boundary has a free surface on top. It is subject to its own weight and it is allowed to slip on the two vertical boundaries. Results are shown on the left in Fig. 5. The following data have been used:

$$\rho^1 = 1, f = -0.1\rho,\ \delta t = 1., \text{Bottom velocity } 1,\ \nu = 0.1.$$

With the same data two layers of fluids with different densities, the top one having $\rho^2 = 2$, rotate under the action of the sliding lower boundary. No-slip conditions is applied to the vertical walls. Here when `movemesh()` flip over a triangle, it is detected with `checkmovemesh()` and then a call to `adaptmesh(th,h,IsMetric=1)` rebuild the mesh `th` with mesh size average h.

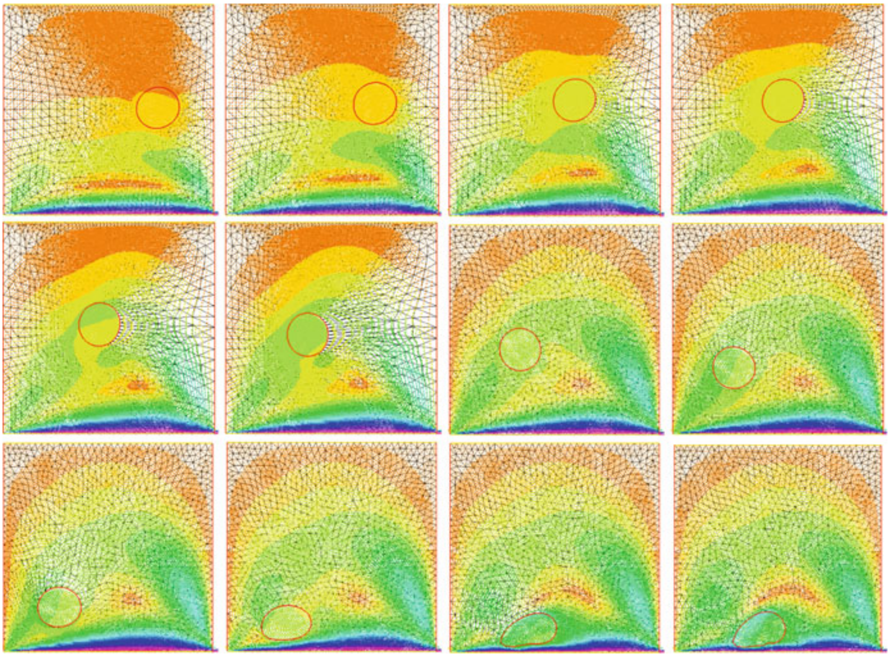


Fig. 4 Free motion of a hyperelastic incompressible solid submitted to its own weight and to the force due to the rotation of the fluid induced by the sliding lower horizontal boundary

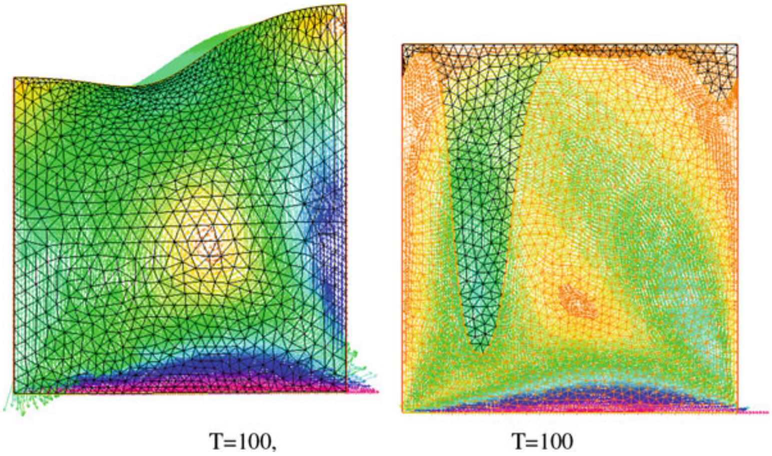


Fig. 5 A free boundary problem (*left*) and a two fluid problem where the fluid above is twice heavier than the one below. Both problems are solved by the same Eulerian algorithm

Once more we stress the fact that all cases have been computed with the same computer program—given in appendix for the case of Fig. 4—without modification of the core algorithm.

9 Conclusion

A fully Eulerian fluid–structure formulation has been presented and an attempt at deriving an implicit unconditionally stable monolithic finite element discretization has been proposed and studied. The method has been implemented with FreeFem++. It is reasonably robust but needs to be made unconditionally stable by implicit iterations on the moving domain, a path currently under investigation.

Appendix: The FreeFem++ Script

```
// FSI with same variable for fluid and structure
border a(t=10,3)  { x=0; y=t;};  //  left
border b(t=0,10) { x=t; y=3 ;};  //  bottom
border c(t=3,7)  { x=10; y=t ;}; //  right low
border d(t=10,1) { x=t; y=7; };  //  low beam
border e(t=7,8) { x=1; y=t;  };  // left beam
border f(t=1,10) { x=t; y=8; };  // top beam
border g(t=8,10) { x=10; y=t;};  // right up
border hh(t=10,0) { x=t; y=10 ;}; //  top
border ee(t=7,8) { x=10; y=t;  };  // left beam
int m=1;
mesh th = buildmesh( a(m*30)+b(m*20)+c(m*16)+d(m*30)+e(m*5)
              +f(m*30)
+g(m*5)+hh(m*20)+ee(m*5));
real h=0.3;
int fluid=th(1,4).region, beam=th(9,7.5).region;

fespace V2h(th,P2);
fespace Vh(th,P1);
fespace Wh(th,P1);
Vh p,ph;
V2h u=0,v=0,uh,vh,d1=0,d2=0, uold=0, vold=0,
               uu,vv, uuold=0,vvold=0;
real nu=0.1;
real E = 2.15;
real sigma = 0.29;
real mu = E/(2*(1+sigma));
real c1=2*mu/2;
real penal=1e-6;
//real lambda = E*sigma/((1+sigma)*(1-2*sigma));
real gravity = -0.2;
real rhof=0.5, rhos=1.;
```

```
macro div(u,v) ( dx(u)+dy(v) ) // EOM
macro DD(u,v)  [[2*dx(u),div(v,u)],[div(v,u),2*dy(v)]] // EOM
macro Grad(u,v)[[dx(u),dy(u)],[dx(v),dy(v)]] // EOM
Vh g11=1,g12=0,g21=0,g22=1, g11aux,g22aux,g12aux,g21aux,
    f11,f12,f21,f22;
macro G[[g11,g12],[g12,g22]]//EOM

int NN=100;
real T=300, dt=T/NN;

problem aa([u,v,p],[uh,vh,ph]) =
int2d(th,beam)( rhos*[u,v]'*[uh,vh]/dt - div(uh,vh)*p
- div(u,v)*ph+ penal*p*ph
+dt*c1*trace(DD(uh,vh)*(DD(u,v)-Grad(u,v)*Grad(d1,d2)'
- Grad(d1,d2)*Grad(u,v)')))
   + int2d(th,beam) ( -rhos*gravity*vh +c1*trace(DD(uh,vh)
    *(DD(d1,d2)
- Grad(d1,d2)*Grad(d1,d2)'))
-  rhos*[uold,vold]'*[uh,vh]/dt)
  +  int2d(th,fluid)(rhof*[u,v]'*[uh,vh]/dt- div(uh,vh)*p
     -div(u,v)*ph
+ penal*p*ph + nu/2*trace(DD(uh,vh)'*DD(u,v)))
   -  int2d(th,fluid)(rhof*gravity*vh
+rhof*[convect([uuold,vvold],-dt,uuold),
convect([uuold,vvold],-dt,vvold)]'*[uh,vh]/dt)
 + on(1,3,7,8,9,u=0,v=0) + on(2,u=0,v=0) ;

// Computation time loop
for(int n=0;n<NN;n++){
aa;
solve bb([uu,vv],[uh,vh]) = int2d(th,fluid)(
trace(Grad(uu,vv)*Grad(uh,vh)') )
+ int2d(th,beam)(10000*[uu,vv]'*[uh,vh])
- int2d(th,beam)(10000*[u,v]'*[uh,vh])
+ on(1,2,3,7,8,uu=0,vv=0) + on(4,5,6,9,uu=u,vv=v);
  real mintcc = checkmovemesh(th,[x,y])/5.;
  real mint = checkmovemesh(th,[x+uu*dt,y+vv*dt]);
  uh=d1;
  vh=d2;
  if (mint<mintcc) {
    th=adaptmesh(th,h,IsMetric=1) ;// plot(th);
    }
  else {
th = movemesh(th,[x+uu*dt,y+vv*dt]);
  d1=0; d1[]=uh[]+dt*u[];
  d2=0; d2[]=vh[]+dt*v[];
  uold=0;  uold[]=u[];
  vold=0;  vold[]=v[];
uuold=u;
vvold=v;
  f11=1+dt*dx(uold); f12= dt*dx(vold); f21=dt*dy(uold);
  f22=1+dt*dy(vold);
     g11aux=g11*f11+g12*f21;
 g22aux=g12*f12+g22*f22;
 g12aux=g11*f12+g12*f22 ;
 g21aux=g12*f11+g12*f21 ;
     g11=f11*g11aux+f21*g21aux;
 g22=f12*g12aux+f22*g22aux;
 g12=f11*g12aux+f21*g22aux ;
  }
if((n/10)*10==n) plot(th,[uold,vold]);
vh=det(Grad(d1,d2));
cout<<n*dt<<" <- time, det d -> " << vh[].max<<
" pmax= "<<ph[].max<<" area= "<<int2d(th,beam)(1.)<<endl;
}
}
```

Acknowledgements The author thanks Frédéric Hecht and Patrice Hauret for very valuable discussions and comments.

References

1. Bathe, K.J.: Finite Element Procedures. Prentice-Hall, Englewood Cliffs, New-Jersey (1996)
2. Boffi, D., Cavallini, N., Gastaldi, L.: The finite element immersed boundary method with distributed lagrange multiplier. SIAM J. Numer. Anal. **53**(6), 2584–2604 (2015)
3. Boukir, K., Maday, Y., Metivet, B.: A high order characteristics method for the incompressible Navier-Stokes equations. Comp. Methods Appl. Math. Eng. **116**, 211–218 (1994)
4. Bukaca, M., Canic, S., Glowinski, R., Tambacac, J., Quainia, A.: Fluid-structure interaction in blood flow capturing non-zero longitudinal structure displacement. J. Comput. Phys. **235**, 515–541 (2013)
5. Chacón-Rebollo, T., Girault, V., Murat, F., Pironneau, O.: Analysis of a coupled fluid-structure model with applications to hemodynamics. SIAM J. Numer. Anal. **54**(2), 994–1019 (2016)
6. Ciarlet, P.G.: Mathematical Elasticity. North Holland, Amsterdam (1988)
7. Costabel, M., Dauge, M.: Singularities of electromagnetic fields in polyhedral domains. Preprint 97–19, Université de Rennes 1 (1997). http://www.maths.univ-rennes1.fr/~dauge/
8. Cottet, G.H., Maitre, E., Milcent, T.: Eulerian formulation and level set models for incompressible fluid-structure interaction. M2AN Math. Model. Numer. Anal. **42**(3), 471–492 (2008)
9. Coupez, Th., Silva, L., Hachem, E.: Implicit boundary and adaptive anisotropic meshes. In: Peretto, S., Formaggia, L. (eds.) New Challenges in Grid Generation and Adaptivity for Scientific Computing. SEMA-SIMAI Springer Series, vol. 5. Springer, Cham (2015)
10. Fernandez, M.A., Mullaert, J., Vidrascu, M.: Explicit Robin-Neumann schemes for the coupling of incompressible fluids with thin-walled structures. Comp. Methods Appl. Mech. Eng. **267**, 566–593 (2013)
11. Formaggia, L., Quarteroni, A., Veneziani, A.: Cardiovasuclar Mathematics. Springer MS&A Series. Springer, Berlin (2009)
12. Hauret, P.: Méthodes numériques pour la dynamique des structures non-linéaires incompressibles à deux échelles. Doctoral thesis, Ecole Polytechnique (2004)
13. Hecht, F.: New development in FreeFem++. J. Numer. Math. **20**, 251–265 (2012). www.FreeFem.org
14. Hron, J., Turek, S.: A monolithic fem solver for an ALE formulation of fluid'structure interaction with configuration for numerical benchmarking. In: Wesseling, P., Onate, E., Periaux, J. (eds.) European Conference on Computational Fluid Dynamics ECCOMAS CFD 2006. TU Delft, The Netherlands (2006)

15. Kane, C., Marsden, J.E., Ortiz, M., West, M.: Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. Int. J. Numer. Methods Eng. **49**, 1295–1325 (2000)
16. Léger, S.: Méthode lagrangienne actualisée pour des problèmes hyperélastiques en très grandes déformations. Thèse de doctorat, Université Laval (2014)
17. Le Tallec, P., Hauret, P.: Energy conservation in fluid-structure interactions. In: Neittanmaki, P., Kuznetsov, Y., Pironneau, O. (eds.) Numerical Methods for Scientific Computing, Variational Problems and Applications. CIMNE, Barcelona, (2003)
18. Le Tallec, P., Mouro, J.: Fluid structure interaction with large structural displacements. Comput. Methods Appl. Mech. Eng. **190**(24–25), 3039–3068 (2001)
19. Liu, J.: A second-order changing-connectivity ALE scheme and its application to FSI with large convection of fluids and near contact of structures. J. Comput. Phys. **304** 380–423 (2016)
20. Liu, I.-S., Cipolatti, R., Rincon, M.A.: Incremental linear approximation for finite elasticity. In: Proceedings of the ICNAAM 2006. Wiley, Weinheim (2006)
21. Lucquin, B., Pironneau, O.: Introduction to Scientific Computing. Wiley, New York (1996)
22. Marsden, J., Hughes, T.J.R.: Mathematical Foundations of Elasticity. Dover Publications, New York (1993)
23. Nobile, F., Vergara, C.: An effective fluid-structure interaction formulation for vascular dynamics by generalized robin conditions. SIAM J. Sci. Comput. **30**(2), 731–763 (2008)
24. Peskin, C.S.: The immersed boundary method. Acta Numer. **11**, 479–517 (2002)
25. Pironneau, O.: Finite Element Methods for Fluids. Wiley, New York (1989)

A New, General Neighboring Optimal Guidance for Aerospace Vehicles

Mauro Pontani

Abstract This work describes and applies the recently introduced, general-purpose perturbative guidance termed variable-time-domain neighboring optimal guidance, which is capable of driving an aerospace vehicle along a specified nominal, optimal path. This goal is achieved by minimizing the second differential of the objective function (related to the flight time) along the perturbed trajectory. This minimization principle leads to deriving all the corrective maneuvers, in the context of an iterative closed-loop guidance scheme. Original analytical developments, based on optimal control theory and adoption of a variable time domain, constitute the theoretical foundation for several original features. The real-time feedback guidance at hand is exempt from the main disadvantages of similar algorithms proposed in the past, such as the occurrence of singularities for the gain matrices. The variable-time-domain neighboring optimal guidance algorithm is applied to two typical aerospace maneuvers: (1) minimum-time climbing path of a Boeing 727 aircraft and (2) interception of fixed and moving targets. Perturbations arising from nonnominal propulsive thrust or atmospheric density and from errors in the initial conditions are included in the dynamical simulations. Extensive Monte Carlo tests are performed, and unequivocally prove the effectiveness and accuracy of the variable-time-domain neighboring optimal guidance algorithm.

1 Introduction

The problem of driving an aerospace vehicle along a specified path leading to fulfilling the boundary conditions associated with the mission specifications requires defining the corrective actions aimed at compensating nonnominal flight conditions. This means that a feedback control law, or, equivalently, a closed-loop guidance algorithm, is to be defined, on the basis of the current state of the vehicle, evaluated at prescribed sampling times.

M. Pontani (✉)
Sapienza University of Rome, Rome, Italy
e-mail: mauro.pontani@uniroma1.it

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_16

Traditionally, two different approaches to guidance exist. Adaptive algorithms compute the flight trajectory at the beginning of each guidance interval, on the basis of feasibility or optimality criteria [3, 19]. Perturbative algorithms assume a specified nominal trajectory, and define the feedback control corrections aimed at maintaining the vehicle in the proximity of the nominal path [7, 9].

Neighboring Optimal Guidance (NOG) is a perturbative guidance concept that relies on the analytical second order optimality conditions, in order to find the corrective control actions in the neighborhood of the reference trajectory. This is an optimal trajectory that satisfies the first and second-order optimality conditions. In general, the neighboring optimal path originates from a perturbed state and is associated with the minimization of the second differential of the objective function. Several time-varying gain matrices, referring to the nominal trajectory, are defined, computed offline, and stored in the onboard computer. Only a limited number of works have been devoted to studying neighboring optimal guidance [1, 4–6, 18, 21]. In particular, a thorough treatment of NOG is due to Chuang [5], who proposed a simple formula for updating the time of flight, and used a very basic strategy to evaluate the gain matrices when the time of flight exceeds its nominal value. Hull [6, 7] supplied further relevant contributions to the topic, using a vector that contains the unknown parameters to optimize and proposing an analytical formulation for the update of the time of flight, albeit only at the initial time. A common difficulty encountered in implementing the NOG consists in the fact that the gain matrices become singular while approaching the final time. As a result, the real-time correction of the time of flight can lead to numerical difficulties so relevant to cause the failure of the guidance algorithm.

This work describes and applies the recently introduced [15, 16], general-purpose variable-time-domain neighboring optimal guidance algorithm (VTD-NOG), on the basis of the general theory of NOG described in [7]. Some fundamental, original features of VTD-NOG are aimed at overcoming the main difficulties related to the use of former NOG schemes, in particular the occurrence of singularities and the lack of an efficient law for the iterative real-time update of the time of flight. This is achieved by adopting a normalized time domain, which leads to defining a novel updating law for the time of flight, a new termination criterion, and a new analytical formulation for the sweep method. Two applications are considered, for the purpose of illustrating the new guidance algorithm: (1) minimum-time-to-climb path of a Boeing 727 aircraft and (2) interception of fixed and moving targets. Specifically, perturbations arising from the imperfect knowledge of the propulsive thrust and from errors in the initial conditions are included in the dynamical modeling. In addition, atmospheric density fluctuations are modeled for application (1). Extensive Monte Carlo (MC) tests are performed, with the intent of demonstrating the effectiveness and accuracy of the variable-time-domain neighboring optimal guidance algorithm.

2 Nominal Trajectory

The nominal trajectory of aerospace vehicles is computed in the absence of any perturbation. For the purpose of applying a neighboring optimal guidance, the nominal path is required to be an optimal trajectory that minimizes a specified objective function.

In general, the spacecraft trajectory is described through the time-varying, n-dimensional state vector $\mathbf{x}(t)$ and controlled through the time-varying, m-dimensional control vector $\mathbf{u}(t)$; the dynamical evolution over the time interval $[t_0, t_f]$ (with t_0 set to 0 and t_f unspecified) depends also on the time-independent, $\tilde{p}$-dimensional parameter vector $\tilde{\mathbf{a}}$. The governing state equations have the general form

$$\dot{\mathbf{x}} = \tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u}, \tilde{\mathbf{a}}, t) \tag{1}$$

and are subject to q boundary conditions

$$\boldsymbol{\psi}(\mathbf{x}_0, \mathbf{x}_f, \tilde{\mathbf{a}}, t_f) = \mathbf{0} \tag{2}$$

where the subscripts "0" and "f" refer to t_0 and t_f . A feasible trajectory is a solution that obeys the state equations (1) and satisfies the boundary conditions (2).

The problem at hand can be reformulated by using the dimensionless normalized time τ defined as

$$\tau := \frac{t}{t_f} \quad \Rightarrow \quad \tau_0 \equiv 0 \leq \tau \leq 1 \equiv \tau_f \tag{3}$$

Let the dot denote the derivative with respect to τ hence forward. If $\mathbf{a} := \begin{bmatrix} \tilde{\mathbf{a}} & t_f \end{bmatrix}^T$ (and $p := \tilde{p} + 1$), the state equations (1) are rewritten as

$$\dot{\mathbf{x}} = t_f \tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u}, \tilde{\mathbf{a}}, t_f \tau) =: \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{a}, \tau) \tag{4}$$

The objective functional to minimize has the following general form:

$$J = \phi(\mathbf{x}_0, \mathbf{x}_f, \mathbf{a}) + \int_0^1 L\left[\mathbf{x}(\tau), \mathbf{u}(\tau), \mathbf{a}, \tau\right] d\tau \tag{5}$$

The spacecraft trajectory optimization problem consists in identifying a feasible solution that minimizes the objective functional J, through selection of the optimal control law $\mathbf{u}^*(t)$ and the optimal parameter vector $\mathbf{a}^*$, i.e.

$$\{\mathbf{u}^*, \mathbf{a}^*\} = \arg \min_{\{\mathbf{u}, \mathbf{a}\}} J \tag{6}$$

2.1 First-Order Necessary Conditions for a Local Extremal

In order to state the necessary conditions for optimality, a Hamiltonian H and a function of the boundary conditions Φ are defined as [7]

$$H(\mathbf{x}, \mathbf{u}, \mathbf{a}, \boldsymbol{\lambda}, \tau) := L + \boldsymbol{\lambda}^T \mathbf{f} \qquad \Phi(\mathbf{x}_0, \mathbf{x}_f, \mathbf{a}, \boldsymbol{\upsilon}) := \phi + \boldsymbol{\upsilon}^T \boldsymbol{\psi} \tag{7}$$

where the time-varying, n-dimensional costate vector $\boldsymbol{\lambda}(\tau)$ and the time-independent, q-dimensional vector $\boldsymbol{\upsilon}$ are the adjoint variables conjugate to the state equations (4) and to the conditions (2), respectively.

In the presence of an optimal (locally minimizing) solution, the following conditions hold:

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} H \tag{8}$$

$$\dot{\boldsymbol{\lambda}} = -\left[\frac{\partial H}{\partial \mathbf{x}}\right]^T \tag{9}$$

$$\boldsymbol{\lambda}_0 = -\left[\frac{\partial \Phi}{\partial \mathbf{x}_0}\right]^T \qquad \boldsymbol{\lambda}_f = \left[\frac{\partial \Phi}{\partial \mathbf{x}_f}\right]^T \tag{10}$$

$$\left[\frac{\partial \Phi}{\partial \mathbf{a}}\right]^T + \int_0^1 \left[\frac{\partial H}{\partial \mathbf{a}}\right]^T d\tau = \mathbf{0} \tag{11}$$

For the very general Hamiltonian (7) the Pontryagin minimum principle (8) yields the control variables as functions of the adjoint variables and the state variables; the relations (9) are the adjoint (or costate) equations, together with the related boundary conditions (10); (11) is equivalent to p algebraic scalar equations. If the control $\mathbf{u}$ is unconstrained, then (8) implies that H is stationary with respect to $\mathbf{u}$ along the optimal path, i.e.

$$H_{\mathbf{u}}^* = \mathbf{0}^T \tag{12}$$

Equations (8) through (11) are well established in optimal control theory (and are proven, for instance, in [7]), and allow translating the optimal control problem into a two-point boundary-value problem. Unknowns are the state $\mathbf{x}$, the parameter vector $\mathbf{a}$, and the adjoint variables $\boldsymbol{\lambda}$ and $\boldsymbol{\upsilon}$ (while the optimal control $\mathbf{u}^*$ is given by (8), as previously remarked). It is straightforward to demonstrate that the condition (11) is equivalent to

$$\boldsymbol{\mu}_f - \left[\frac{\partial \Phi}{\partial \mathbf{a}}\right]^T = \mathbf{0}, \quad \text{with} \quad \dot{\boldsymbol{\mu}} = -\left[\frac{\partial H}{\partial \mathbf{a}}\right]^T \quad \text{and} \quad \boldsymbol{\mu}_0 = \mathbf{0} \tag{13}$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_f$ are, respectively, the initial and final value (at $\tau = 1$) of the time-varying $(p \times 1)$-vector $\boldsymbol{\mu}$.

2.2 Second-Order Sufficient Conditions for a Local Minimum

The derivation of the second-order optimality conditions involves the definition of an admissible comparison path, located in the neighborhood of the (local) nominal, optimal solution, associated with the state $\mathbf{x}^*$, costate $\boldsymbol{\lambda}^*$, and control $\mathbf{u}^*$. By definition, an admissible comparison path is a feasible trajectory that satisfies the equations of motion and the boundary conditions. A neighboring optimal path is an admissible comparison trajectory that satisfies also the optimality conditions. The nonexistence of alternative neighboring optimal paths is to be proven in order to guarantee optimality of the nominal solution [7, 8].

The first second-order condition is the Clebsch-Legendre sufficient condition for a minimum [7, 8], i.e. $H^*_{\mathbf{uu}} > 0$ (positive definiteness of $H^*_{\mathbf{uu}}$). In the necessary (weak) form the Hessian $H^*_{\mathbf{uu}}$ must be positive semidefinite.

In general, a neighboring optimal path located in the proximity of the optimal solution fulfills the feasibility equations (4) and (2) and the optimality conditions (8)–(11) to first order. This means that the state and costate displacements $\{\delta\mathbf{x}, \delta\boldsymbol{\lambda}\}$ (from the optimal solution) satisfy the linear equations deriving from (4) and (9),

$$\delta\dot{\mathbf{x}} = \mathbf{f}_{\mathbf{x}}\delta\mathbf{x} + \mathbf{f}_{\mathbf{u}}\delta\mathbf{u} + \mathbf{f}_{\mathbf{a}}\delta\mathbf{a} \tag{14}$$

$$\delta\dot{\boldsymbol{\lambda}} = -H_{\mathbf{xx}}\delta\mathbf{x} - H_{\mathbf{xu}}\delta\mathbf{u} - H_{\mathbf{x}\boldsymbol{\lambda}}\delta\boldsymbol{\lambda} - H_{\mathbf{xa}}\delta\mathbf{a} \tag{15}$$

in conjunction with the respective linear boundary conditions, derived from (2) and (10),

$$\boldsymbol{\psi}_{\mathbf{x}_f}\delta\mathbf{x}_f + \boldsymbol{\psi}_{\mathbf{x}_0}\delta\mathbf{x}_0 + \boldsymbol{\psi}_{\mathbf{a}}\delta\mathbf{a} = \mathbf{0} \tag{16}$$

$$\delta\boldsymbol{\lambda}_0 = -\Phi_{\mathbf{x}_0\mathbf{x}_0}\delta\mathbf{x}_0 - \Phi_{\mathbf{x}_0\mathbf{a}}\delta\mathbf{a} - \boldsymbol{\psi}^T_{\mathbf{x}_0}d\boldsymbol{\upsilon} \tag{17}$$

$$\delta\boldsymbol{\lambda}_f = \Phi_{\mathbf{x}_f\mathbf{x}_f}\delta\mathbf{x}_f + \Phi_{\mathbf{x}_f\mathbf{a}}\delta\mathbf{a} + \boldsymbol{\psi}^T_{\mathbf{x}_f}d\boldsymbol{\upsilon} \tag{18}$$

The fact that the Hamiltonian is stationary with respect to $\mathbf{u}$, i.e. $H^*_{\mathbf{u}} = \mathbf{0}^T$, yields

$$H_{\mathbf{ux}}\delta\mathbf{x} + H_{\mathbf{uu}}\delta\mathbf{u} + H_{\mathbf{ua}}\delta\mathbf{a} + H_{\mathbf{x}\boldsymbol{\lambda}}\delta\boldsymbol{\lambda} = \mathbf{0} \tag{19}$$

Under the assumption that the Clebsch-Legendre condition is satisfied, (19) is solved for $\delta\mathbf{u}$

$$\delta\mathbf{u} = -H^{-1}_{\mathbf{uu}}\left(H_{\mathbf{ux}}\delta\mathbf{x} + H_{\mathbf{ua}}\delta\mathbf{a} + H_{\mathbf{x}\boldsymbol{\lambda}}\delta\boldsymbol{\lambda}\right) \tag{20}$$

The parameter condition (11) is replaced by (13), leading to the following relations:

$$\delta\dot{\boldsymbol{\mu}} = -H_{\mathbf{ax}}\delta\mathbf{x} - H_{\mathbf{au}}\delta\mathbf{u} - H_{\mathbf{aa}}\delta\mathbf{a} - H_{\mathbf{a}\boldsymbol{\lambda}}\delta\boldsymbol{\lambda}, \quad \text{with} \tag{21}$$

$$\delta\boldsymbol{\mu}_0 = \mathbf{0}, \quad \delta\boldsymbol{\mu}_f = -\Phi_{\mathbf{ax}_f}\delta\mathbf{x}_f - \Phi_{\mathbf{aa}}\delta\mathbf{a} - \boldsymbol{\psi}_{\mathbf{a}}^T d\boldsymbol{\upsilon} \tag{22}$$

where (22) is written under the assumption that $\Phi_{\mathbf{ax}_0} = \mathbf{0}$, condition that is met for the problems at hand. It is relatively straightforward to recognize that solving the equation system (14)–(19) and (21)–(22) is equivalent to solving the *accessory optimization problem* [7, 8], which consists in minimizing the second differential d^2J. The solution process involves the definition of the sweep variables, through the following relations:

$$\delta\boldsymbol{\lambda} = \mathbf{S}\delta\mathbf{x} + \mathbf{R}d\boldsymbol{\upsilon} + \mathbf{m}d\mathbf{a}, \tag{23}$$

$$\mathbf{0} = \mathbf{R}\delta\mathbf{x} + \mathbf{Q}d\boldsymbol{\upsilon} + \mathbf{n}d\mathbf{a}, \tag{24}$$

$$\delta\boldsymbol{\mu} = \mathbf{m}\delta\mathbf{x} + \mathbf{n}^T d\boldsymbol{\upsilon} + \boldsymbol{\alpha}d\mathbf{a} \tag{25}$$

The matrices $\mathbf{S}$, $\mathbf{R}$, $\mathbf{m}$, $\mathbf{Q}$, $\mathbf{n}$, and $\boldsymbol{\lambda}$ must satisfy the sweep equations (not reported for the sake of conciseness), in conjunction with the respective boundary conditions (prescribed at the final time) [7, 8]. The variations $d\boldsymbol{\upsilon}$ and $d\mathbf{a}$ can be solved simultaneously at τ_0 (at which $\delta\boldsymbol{\mu}_0 = \mathbf{0}$, cf. (22)), to yield

$$\begin{bmatrix} d\boldsymbol{\upsilon} \\ d\mathbf{a} \end{bmatrix} = -\mathbf{V}_0^{-1}\mathbf{U}_0^T\delta\mathbf{x}_0 \quad \text{where} \quad \mathbf{U} := [\mathbf{R} \;\; \mathbf{m}] \quad \text{and} \quad \mathbf{V} := \begin{bmatrix} \mathbf{Q} & \mathbf{n} \\ \mathbf{n}^T & \boldsymbol{\alpha} \end{bmatrix} \tag{26}$$

If (26) is used at τ_0, then $\delta\boldsymbol{\lambda}_0 = \left(\mathbf{S}_0 - \mathbf{U}_0\mathbf{V}_0^{-1}\mathbf{U}_0^T\right)\delta\mathbf{x}_0$. Letting $\hat{\mathbf{S}} = \mathbf{S} - \mathbf{U}\mathbf{V}^{-1}\mathbf{U}^T$, the same sweep equation satisfied by $\mathbf{S}$ turns out to hold also for $\hat{\mathbf{S}}$, with boundary condition $\mathbf{S} \to \mathbf{0}$ as $\tau \to \tau_f \, (= 1)$. From the previous relation on $\delta\boldsymbol{\lambda}_0$ and $\delta\mathbf{x}_0$ one can conclude that $\delta\boldsymbol{\lambda}_0 \to \mathbf{0}$ as $\delta\mathbf{x}_0 \to \mathbf{0}$, unless $\hat{\mathbf{S}}$ tends to infinity at an internal time $\bar{\tau}$ $(\tau_0 \leq \bar{\tau} < \tau_f)$, which is referred to as conjugate point. If $\delta\boldsymbol{\lambda}_0 \to \mathbf{0}$ and $\delta\mathbf{x}_0 \to \mathbf{0}$ then also $\delta\mathbf{u} \to \mathbf{0}$. In the end, if $\hat{\mathbf{S}} < \infty$, then no neighboring optimal path exists. This is the *Jacobi condition*. The use of $\mathbf{S}$ is not effective for the purpose of guaranteeing optimality. In fact, cases exist for which $\mathbf{S}$ becomes singular, while $\hat{\mathbf{S}}$ remains finite [7, 8], and this fully justifies the use of $\hat{\mathbf{S}}$.

It is worth remarking that, with the exception of the displacements $\{\delta\mathbf{x}, \delta\mathbf{u}, \delta\mathbf{a}, \;\delta\boldsymbol{\upsilon}, \delta\boldsymbol{\lambda}, \delta\boldsymbol{\mu}, \delta\mathbf{x}_0, \delta\mathbf{x}_f\}$, all the vectors and matrices reported in this section are evaluated along the nominal, optimal trajectory.

3 Variable-Time-Domain Neighboring Optimal Guidance

The iterative Variable-Time-Domain Neighboring Optimal Guidance (VTD-NOG) uses the optimal trajectory as the reference path, with the final intent of determining the control correction at each sampling time $\{t_k\}_{k=0,\ldots,n_S}$. These are the times at which the displacement between the actual trajectory, associated with $\mathbf{x}$, and the nominal trajectory, corresponding to $\mathbf{x}^*$, is evaluated, to yield $d\mathbf{x}_k \equiv \delta\mathbf{x}_k = \mathbf{x}(t_k) - \mathbf{x}_k^*(t_k)$. The total number of sampling times, n_S, is unspecified, whereas the actual time interval between two successive sampling times is given and denoted with Δt_S, $\Delta t_S = t_{k+1} - t_k$. It is apparent that a fundamental ingredient needed to implement VTD-NOG is the formula for determining the overall time of flight $t_f^{(k)}$ at time t_k. This is equivalent to finding the time-to-go $\left(t_f^{(k)} - t_k\right)$ at t_k. The following subsection is focused on this issue.

3.1 *Time-to-Go Updating Law and Termination Criterion*

The fundamental principle that underlies the VTD-NOG scheme consists in finding the control correction $\delta\mathbf{u}(\tau)$ in the generic interval $[\tau_k, \tau_{k+1}]$ such that the second differential of J is minimized,

$$d^2J = \int_{\tau_k}^{1} \begin{bmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \\ d\mathbf{a} \end{bmatrix}^T \begin{bmatrix} H_{\mathbf{xx}} & H_{\mathbf{xu}} & H_{\mathbf{xa}} \\ H_{\mathbf{ux}} & H_{\mathbf{uu}} & H_{\mathbf{ua}} \\ H_{\mathbf{ax}} & H_{\mathbf{au}} & H_{\mathbf{aa}} \end{bmatrix} \begin{bmatrix} \delta\mathbf{x} \\ \delta\mathbf{u} \\ d\mathbf{a} \end{bmatrix} d\tau \\ + \begin{bmatrix} d\mathbf{x}_k \\ d\mathbf{x}_f \\ d\mathbf{a} \end{bmatrix}^T \begin{bmatrix} \Phi_{\mathbf{x}_k\mathbf{x}_k} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times p} \\ \mathbf{0}_{n\times n} & \Phi_{\mathbf{x}_f\mathbf{x}_f} & \Phi_{\mathbf{x}_f\mathbf{a}} \\ \mathbf{0}_{p\times n} & \Phi_{\mathbf{a}\mathbf{x}_f} & \Phi_{\mathbf{aa}} \end{bmatrix} \begin{bmatrix} d\mathbf{x}_k \\ d\mathbf{x}_f \\ d\mathbf{a} \end{bmatrix} \tag{27}$$

while holding the first-order expansions of the state equations, the related final conditions, and the parameter condition (i.e., the second of (22)). In contrast, the first of (22) cannot be used, because in general $\delta\boldsymbol{\mu}_k \neq \mathbf{0}$ at τ_k. Minimizing the objective (27) is equivalent to solving the accessory optimization problem, defined in the interval $[\tau_k, 1]$. This means that the relations reported in Sect. 2.2 need to be extended to the generic interval $[\tau_k, 1]$.

Other than the linear expansion of the state and costate equations, the related boundary conditions, and the second relation of (22), also Eqs. (23)–(25), (19), and (20) remain unchanged. However, now (26) is to be evaluated at τ_k and becomes

$$\begin{bmatrix} d\boldsymbol{\upsilon} \\ d\mathbf{a} \end{bmatrix} = -\mathbf{V}_k^{-1}\mathbf{U}_k^T\delta\mathbf{x}_k - \mathbf{V}_k^{-1}\boldsymbol{\Theta}\delta\boldsymbol{\mu}_k, \quad \text{with} \quad \boldsymbol{\Theta} = \begin{bmatrix} \mathbf{0}_{q\times p} \\ \mathbf{I}_{p\times p} \end{bmatrix} \tag{28}$$

because $\delta\boldsymbol{\mu}_k \neq \mathbf{0}$ (unlike $\delta\boldsymbol{\mu}_0 = \mathbf{0}$). The latter relation supplies the corrections $d\boldsymbol{v}$ and $d\mathbf{a}$ at τ_k as functions of the gain matrices $\mathbf{U}$ and $\mathbf{V}$ (defined in (26)), evaluated at τ_k, and $\delta\boldsymbol{\mu}_k$ (coming from the numerical integration of (21) in the preceding interval $[\tau_{k-1}, \tau_k]$). Equation (28) contains the updating law of the total flight time t_f, which is included as a component of $\mathbf{a}$. Hence, if $dt_f^{(k)}$ denotes the correction on t_f^* evaluated at τ_k, then $t_f^{(k)} = t_f^* + dt_f^{(k)}$. As the sampling interval Δt_S is specified, the general formula for τ_k is

$$\tau_k = \sum_{j=0}^{k-1} \frac{\Delta t_S}{t_f^{(j)}} \tag{29}$$

The overall number of intervals n_S is found at the first occurrence of the following condition, associated with the termination of VTD-NOG:

$$\sum_{j=0}^{n_S-1} \frac{\Delta t_S}{t_f^{(j)}} \geq 1 \qquad \Rightarrow \qquad \tau_{n_S} = 1 \tag{30}$$

It is worth stressing that the updating formula (28) derives directly from the natural extension of the accessory optimization problem to the time interval $[\tau_k, 1]$. In addition, the introduction of the normalized time τ now reveals its great utility. In fact, all the gain matrices are defined in the normalized interval [0,1] and cannot become singular. Moreover, the limiting values $\{\tau_k\}_{k=0,\ldots,n_S-1}$ are dynamically calculated at each sampling time using (29), while the sampling instants in the actual time domain are specified and equally spaced (cf. Fig. 1). Also the termination criterion (30) has a logical, consistent definition, and corresponds to the upper bound of the interval [0,1], to which τ is constrained.

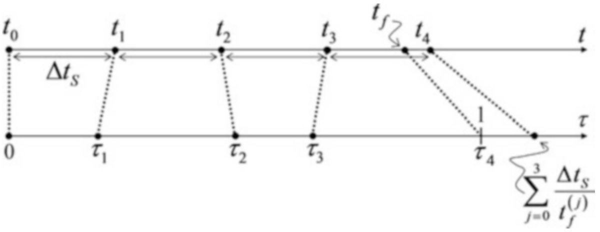


Fig. 1 Illustrative sketch of the relations between the two time domains

3.2 Modified Sweep Method

The definition of a neighboring optimal path requires the numerical backward integration of the sweep equations [7]. However, as previously remarked, the matrix $\hat{\mathbf{S}}$ has practical utility, because $\mathbf{S}$ may become singular while $\hat{\mathbf{S}}$ remains finite in the interval [0,1[. Therefore, a suitable integration technique is based on using the classical sweep equations in the interval $[\tau_{sw}, 1]$ (where τ_{sw} is sufficiently close to $\tau_f = 1$) and then switching to $\hat{\mathbf{S}}$. However, due to (28), new relations are to be derived for $\hat{\mathbf{S}}$ and the related matrices.

With this intent, the first step consists in combining (28) with (23)–(24), and leads to obtaining

$$\delta\boldsymbol{\lambda} = \left(\hat{\mathbf{S}} - \mathbf{W}\mathbf{m}^T\right)\delta\mathbf{x} - \mathbf{W}\mathbf{n}^T d\boldsymbol{\upsilon} - \mathbf{W}\boldsymbol{\alpha} d\mathbf{a}, \quad \text{with} \quad \mathbf{W} := \mathbf{U}\mathbf{V}^{-1}\boldsymbol{\Theta} \tag{31}$$

This relation replaces (23).

Equation (31) is to be employed repeatedly in the derivation of new sweep equations. The related analytical developments are described in full detail in [15], and lead to attaining the following modified sweep equations:

$$\dot{\hat{\mathbf{S}}} = -\hat{\mathbf{S}}\mathbf{A} + \hat{\mathbf{S}}\mathbf{B}\hat{\mathbf{S}} + \left[\hat{\mathbf{S}}\mathbf{D}\boldsymbol{\alpha}^{-1} + \mathbf{W}\mathbf{F}\boldsymbol{\alpha}^{-1} + \mathbf{E}\boldsymbol{\alpha}^{-1}\right]\mathbf{m}^T - \mathbf{W}\mathbf{E}^T - \mathbf{W}\mathbf{D}^T\hat{\mathbf{S}} - \mathbf{C} - \mathbf{A}^T\hat{\mathbf{S}} \tag{32}$$

$$\dot{\mathbf{R}}^T = \mathbf{R}^T\mathbf{B}\hat{\mathbf{S}} - \mathbf{R}^T\mathbf{A} - \mathbf{R}^T\mathbf{B}\mathbf{W}\mathbf{m}^T \tag{33}$$

$$\dot{\mathbf{m}}^T = -\mathbf{m}^T\mathbf{A} + \mathbf{m}^T\mathbf{B}\hat{\mathbf{S}} - \mathbf{m}^T\mathbf{B}\mathbf{W}\mathbf{m}^T - \mathbf{E}^T - \mathbf{D}^T\hat{\mathbf{S}} + \mathbf{D}^T\mathbf{W}\mathbf{m}^T \tag{34}$$

$$\dot{\mathbf{Q}} = -\mathbf{R}^T\mathbf{B}\mathbf{W}\mathbf{n}^T \tag{35}$$

$$\dot{\mathbf{n}} = -\mathbf{R}^T\left(\mathbf{D} + \mathbf{B}\mathbf{W}\boldsymbol{\alpha}\right) \tag{36}$$

$$\dot{\boldsymbol{\alpha}} = \mathbf{D}^T\mathbf{W}\boldsymbol{\alpha} - \mathbf{F} - \mathbf{m}^T\mathbf{B}\mathbf{W}\boldsymbol{\alpha} - \mathbf{m}^T\mathbf{D} \tag{37}$$

The gain matrices involved in the sweep method, i.e. $\mathbf{S}$, $\hat{\mathbf{S}}$, $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{n}$, $\mathbf{m}$, and $\boldsymbol{\alpha}$, can be integrated backward in two steps:

1. in $[\tau_{sw}, 1]$ the equations of the classical sweep method [7, 15], with the respective boundary conditions, are used.
2. in $[0, \tau_{sw}]$ (32) through (37) are used; $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{n}$, $\mathbf{m}$, and $\boldsymbol{\alpha}$ are continuous across the time τ_{sw}, whereas $\hat{\mathbf{S}}$ is given by $\hat{\mathbf{S}} := \mathbf{S} - \mathbf{U}\mathbf{V}^{-1}\mathbf{U}^T$; in this work, τ_{sw} is set to 0.99.

3.3 Preliminary Offline Computations and Algorithm Structure

The implementation of VTD-NOG requires several preliminary computations that can be completed offline and stored in the onboard computer. First of all, the optimal trajectory is to be determined, together with the related state, costate, and control variables, which are assumed as the nominal ones. In the time domain τ these can be either available analytically or represented as sequences of equally spaced values, e.g. $\mathbf{u}_i^* = \mathbf{u}^*(\tau_i)$ ($i = 0, \ldots, n_D$; $\tau_0 = 0$ and $\tau_{n_D} = 1$). However, in the presence of perturbations, VTD-NOG determines the control corrections $\delta\mathbf{u}(\tau)$ in each interval $[\tau_k, \tau_{k+1}]$, where the values $\{\tau_k\}$ never coincide with the equally spaced values $\{\tau_i\}$ used for $\mathbf{u}_i^*$. Hence, regardless of the number of points used to represent the control correction $\delta\mathbf{u}(\tau)$ in $[\tau_k, \tau_{k+1}]$, it is apparent that a suitable interpolation is to be adopted for the control variable $\mathbf{u}^*$ (provided that no analytical expression is available). In this way, the value of $\mathbf{u}^*$ can be evaluated at any arbitrary time in the interval $0 \le \tau \le 1$. For the same reason also the nominal state $\mathbf{x}^*$ and costate $\boldsymbol{\lambda}^*$ need to be interpolated. If a sufficiently large number of points is selected (e.g., $n_D = 1001$), then piecewise linear interpolation is a suitable option. The successive step is the analytical derivation of the matrices

$$\left\{\begin{array}{l}\mathbf{f}_{\mathbf{x}}, \mathbf{f}_{\mathbf{u}}, \mathbf{f}_{\mathbf{a}}, H_{\mathbf{xx}}, H_{\mathbf{xu}}, H_{\mathbf{x}\boldsymbol{\lambda}}, H_{\mathbf{xa}}, H_{\mathbf{ux}}, H_{\mathbf{uu}}, H_{\mathbf{ua}}, H_{\mathbf{u}\boldsymbol{\lambda}}, H_{\mathbf{ax}}, H_{\mathbf{au}}, H_{\mathbf{aa}}, H_{\mathbf{a}\boldsymbol{\lambda}}, \\ \boldsymbol{\psi}_{\mathbf{x}_f}, \boldsymbol{\psi}_{\mathbf{x}_0}, \boldsymbol{\psi}_{\mathbf{a}}, \Phi_{\mathbf{x}_0\mathbf{x}_0}, \Phi_{\mathbf{x}_0\mathbf{a}}, \Phi_{\mathbf{x}_f\mathbf{x}_f}, \Phi_{\mathbf{x}_f\mathbf{a}}, \Phi_{\mathbf{ax}_f}, \Phi_{\mathbf{aa}}\end{array}\right\} \tag{38}$$

Then, they are evaluated along the nominal trajectory and linearly interpolated, as well as **A**, **B**, **C**, **D**, **E**, and **F**, whose expressions are reported in [17]. Subsequently, the two-step backward integration of the sweep equations described in Sect. 3.2 is performed, and yields the gain matrices $\hat{\mathbf{S}}$, **R**, **m**, **Q**, **n**, and $\boldsymbol{\alpha}$, using also the analytic expressions of **W**, **U**, and **V**. The linear interpolation of all the matrices not yet interpolated concludes the preliminary computations.

On the basis of the optimal reference path, at each time τ_k the VTD-NOG algorithm determines the time of flight and the control correction. More specifically, after setting the actual sampling time interval Δt_S, at each τ_k ($k = 0, \ldots, n_S - 1$; $\tau_0 = 0$) the following steps implement the feedback guidance scheme:

1. Evaluate $\delta\mathbf{x}_k$.
2. Assume the value of $\delta\boldsymbol{\mu}$ calculated at the end of the previous interval $[\tau_{k-1}, \tau_k]$ as $\delta\boldsymbol{\mu}_k$ ($\delta\boldsymbol{\mu}_0 = \mathbf{0}$).
3. Calculate the correction $dt_f^{(k)}$ and the updated time of flight $t_f^{(k)}$.
4. Calculate the limiting value τ_{k+1}.
5. Evaluate $\delta\boldsymbol{\lambda}_k$.
6. Integrate numerically the linear differential system composed of (14), (15), and (21).
7. Determine the control correction $\delta\mathbf{u}(\tau)$ in $[\tau_k, \tau_{k+1}]$ through (20).
8. Points 1 through 7 are repeated after increasing k by 1, until (30) is satisfied.

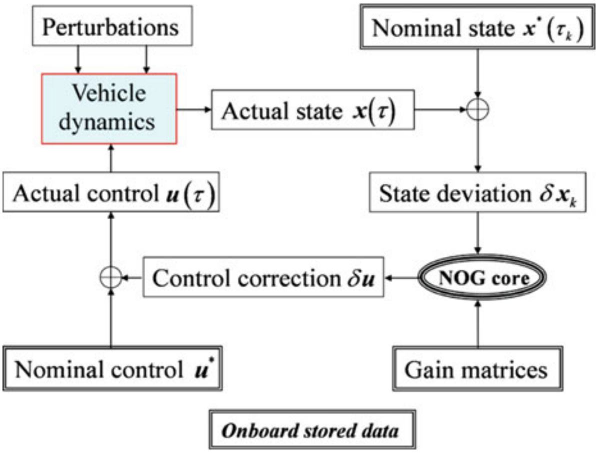


Fig. 2 Block diagram of VTD-NOG

Figure 2 portrays a block diagram that illustrates the feedback structure of VTD-NOG. The control and flight time corrections depend on the state displacement $\delta\mathbf{x}$ (evaluated at specified times) through the time-varying gain matrices, computed offline and stored onboard.

4 Minimum-Time-to-Climb Path of a Boeing 727 Aircraft

As a first example, the variable-time-domain neighboring optimal guidance is applied to the minimum-time ascent path of a Boeing 727 aircraft, which is a mid-size commercial jet aircraft. Its propulsive and aerodynamics characteristics are interpolated on the basis of real data, and come from [2].

4.1 Problem Definition

The aircraft motion is assumed to occur in the vertical plane. In addition, due to the low flight altitude, the flat-Earth approximation is adopted, and the gravitational force is considered constant ($g = 9.80665\,\text{m/s}^2$). In light of these assumptions, the equations of motion are [20]

$$z\prime = v \sin\gamma \tag{39}$$

$$x\prime = v \cos\gamma \tag{40}$$

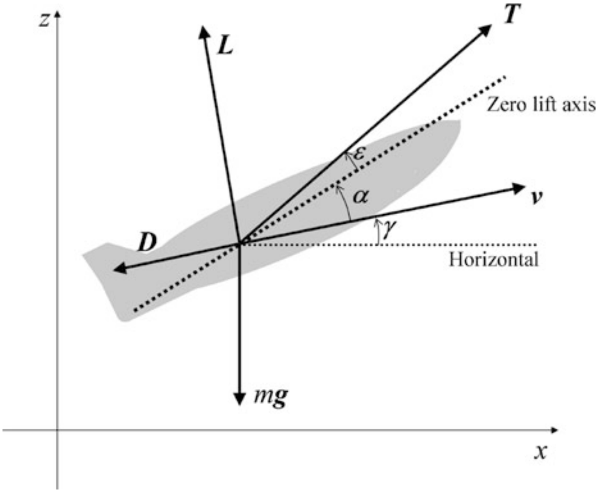


Fig. 3 Thrust, lift, and drag forces, with the related angles

$$\gamma\prime = -\frac{g}{v}\cos\gamma + \frac{L}{mv} + \frac{T}{mv}\sin(\alpha + \epsilon) \tag{41}$$

$$v\prime = -g\sin\gamma - \frac{D}{m} + \frac{T}{m}\cos(\alpha + \epsilon) \tag{42}$$

where z, x, γ, and v denote, respectively, the altitude, range, flight path angle, and velocity of the aircraft at hand, and $\prime$ is the derivative with respect to the actual time t; D and L represent the magnitudes of the aerodynamic drag and lift (whose direction is illustrated in Fig. 3), whereas T denotes the thrust magnitude, and m is the mass, which is assumed constant and equal to 81,647 kg. The angle ϵ is portrayed in Fig. 3 as well, and identifies the thrust direction with respect to the zero lift axis; α is the angle of attack. The two aerodynamic forces are functions of (1) the (dimensionless) lift and drag coefficients c_L and c_D, (2) the atmospheric density ρ, (3) the instantaneous velocity v, and (4) the reference surface area S, equal to 145 m^2 [2], according to the following relations:

$$L = \frac{1}{2}c_L \rho S v^2 \qquad \text{and} \qquad D = \frac{1}{2}c_D \rho S v^2 \tag{43}$$

Due to low altitude, the atmospheric density at sea level is used for the entire time of flight. The two coefficients c_L and c_D depend on the angle of attack and are interpolated in the following fashion [2]:

$$c_L = c_{L0} + c_{L1}\alpha + c_{L2}\left(\alpha - \alpha_1\right)^2 \tag{44}$$

$$c_D = c_{D0} + c_{D1}\alpha + c_{D2}\alpha^2 \tag{45}$$

where the values of the constant quantities $\{c_{L0}, c_{L1}, c_{L2}, \alpha_1, c_{D0}, c_{D1}, c_{D2}\}$ are reported in [2]. Lastly, the thrust magnitude T depends on the instantaneous velocity, and is interpolated as well,

$$T = c_{T0} + c_{T1} v + c_{T2} v^2 \tag{46}$$

where the constant values of $\{c_{T0}, c_{T1}, c_{T2}\}$ are again reported in [2].

The minimum-time-to-climb problem consists in finding the optimal time history of the control angle α that minimizes the time t_f needed to reach a given altitude in horizontal flight, with a prescribed velocity. This means that the objective function is simply

$$J = t_f \tag{47}$$

(the initial time is set to 0). The final conditions are partially specified,

$$z_f = \bar{z}_f \,(= 609.6\,\text{m}), \quad \gamma_f = 0\,\text{deg}, \quad v_f = \bar{v}_f \,(= 128\,\text{m/s}) \tag{48}$$

whereas the initial conditions are completely known,

$$z_0 = 0\,\text{m}, \quad x_0 = 0\,\text{m}, \quad \gamma_0 = 0\,\text{deg}, \quad v_0 = 128\,\text{m/s} \tag{49}$$

As the range x does not appear in the right-hand sides of the equations of motion nor in the final conditions, x is ignorable; as a result, the state vector $\mathbf{x}$ is given by $\mathbf{x} = [z \quad \gamma \quad v]^T$, while the control vector $\mathbf{u}$ includes only α ($\mathbf{u} \equiv \alpha$). The state equations can be rewritten in terms of τ-derivatives,

$$\dot{z} = t_f v \sin\gamma \tag{50}$$

$$\dot{\gamma} = t_f \left[-\frac{g}{v}\cos\gamma + \frac{L}{mv} + \frac{T}{mv}\sin(\alpha + \epsilon) \right] \tag{51}$$

$$\dot{v} = t_f \left[-g\sin\gamma - \frac{D}{m} + \frac{T}{m}\cos(\alpha + \epsilon) \right] \tag{52}$$

The right-hand sides of (50)–(52) form the vector $\mathbf{f}$.

4.2 Optimal Trajectory

The VTD-NOG algorithm requires the preliminary determination of the optimal trajectory, which is assumed as the nominal path, together with the related optimal control, state, and costate vectors (cf. Sect. 2).

For the dynamical system at hand the Hamiltonian H and the function Φ are

$$\begin{aligned} H = \lambda_1 t_f v \sin\gamma + \lambda_2 t_f \left[-\tfrac{g}{v}\cos\gamma + \tfrac{L}{mv} + \tfrac{T}{mv}\sin(\alpha+\epsilon)\right] \\ +\lambda_3 t_f \left[-g\sin\gamma - \tfrac{D}{m} + \tfrac{T}{m}\cos(\alpha+\epsilon)\right] \end{aligned} \tag{53}$$

$$\Phi = \upsilon_1(z_f - \bar{z}_f) + \upsilon_2\gamma_f + \upsilon_3(v_f - \bar{v}_f) \tag{54}$$

The adjoint equations assume the form

$$\dot{\lambda}_1 = 0 \qquad \Rightarrow \qquad \lambda_1 = \lambda_{1,0} \tag{55}$$

$$\dot{\lambda}_2 = t_f\left[-\lambda_1 v\cos\gamma - \frac{\lambda_2 g}{v}\sin\gamma + \lambda_3 g\cos\gamma\right] \tag{56}$$

$$\begin{aligned} \dot{\lambda}_3 = t_f\left\{-\lambda_1\sin\gamma - \tfrac{\lambda_2}{mv}\left[\tfrac{\partial T}{\partial v}\sin(\alpha+\epsilon) + \tfrac{\partial L}{\partial v} - \tfrac{T}{v}\sin(\alpha+\epsilon) - \tfrac{L}{v}\right]\right\} \\ +t_f\left\{-\lambda_2\tfrac{g\cos\gamma}{v^2} - \lambda_3\left[-\tfrac{\partial D}{\partial v} + \tfrac{\partial T}{\partial v}\cos(\alpha+\epsilon)\right]\right\} \end{aligned} \tag{57}$$

The respective boundary conditions (10) do not add any further information (since the state components are completely specified at the final time), and therefore they are not reported. It is worth remarking that the derivatives of L, D, and T with respect to α and v can be easily expressed using (43)–(46) and are continuous. Moreover, the fact that H is stationary with respect to α at the optimal solution yields

$$\lambda_2\frac{T}{mv}\cos(\alpha+\epsilon) + \frac{\lambda_2}{mv}\frac{\partial L}{\partial\alpha} - \frac{\lambda_3}{m}\frac{\partial D}{\partial\alpha} - \lambda_3\frac{T}{m}\sin(\alpha+\epsilon) = 0 \tag{58}$$

Finally, the parameter condition (11) leads to

$$\int_0^1 \boldsymbol{\lambda}^T\frac{\partial\mathbf{f}}{\partial t_f}d\tau + 1 = 0 \tag{59}$$

However, the parameter condition can be proven to be ignorable. As a first step, the components of $\boldsymbol{\lambda}$ are homogeneous in the adjoint equations (55)–(57). This circumstance implies that if an optimization algorithm is capable of finding some initial value of $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}_0 = k_\lambda\boldsymbol{\lambda}_0^*$ ($k_\lambda > 0$) (where * denotes optimality), then the same proportionality holds between $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*$ at any τ. Moreover, the control $\mathbf{u}$ can be found through (58), which yields the same solution if $\boldsymbol{\lambda}$ replaces $\boldsymbol{\lambda}^*$. This circumstance implies that if $\boldsymbol{\lambda}$ is proportional to $\boldsymbol{\lambda}^*$ then the final conditions are fulfilled at the minimum final time t_f^*. In contrast, the parameter condition is violated, because the integral of (59) turns out to be

$$\int_0^1 \boldsymbol{\lambda}^T \frac{\partial \mathbf{f}}{\partial t_f} d\tau = k_\lambda \int_0^1 \boldsymbol{\lambda}^{*T} \frac{\partial \mathbf{f}}{\partial t_f} d\tau = -k_\lambda \neq -1 \tag{60}$$

Therefore, provided that the proportionality condition holds, the optimal control $\mathbf{u}^*$ can be determined without considering the parameter condition, which is ignorable as an equality constraint and can be replaced by the inequality constraint

$$\int_0^1 \boldsymbol{\lambda}^T \frac{\partial \mathbf{f}}{\partial t_f} d\tau < 0 \tag{61}$$

because k_λ is an arbitrary positive constant. Once the (nonoptimal) values of the costate variables (fulfilling the proportionality condition) have been determined, the correct (optimal) values can be recovered after calculating k_λ by means of (60).

In the end, the problem of determining the minimum-time-to-climb path can be reformulated as a two-point boundary-value problem, in which the unknowns are the initial values of three adjoint variables, as well as the time of flight, i.e. $\{\lambda_{1,0}, \lambda_{2,0}, \lambda_{3,0}, t_f\}$. The boundary conditions are represented by (48), accompanied by the inequality constraint (61). Once the optimal parameter set has been determined, the state and costate equations can be integrated, using (58) to express the control angle α as a function of the adjoint variables.

The optimal parameter set is determined by means of a simple implementation of swarming algorithm (PSO). This is a heuristic optimization technique, based on the use of a population of individuals (or particles). Selection of the globally optimal parameters is the result of a number of iterations, in which the individuals share their information. This optimization approach is extremely intuitive and easy-to-implement. Nevertheless, in the scientific literature several papers [10–14] prove that the use of this method is effective for solving trajectory optimization problems. A set of canonical units is employed for the problem at hand: the distance unit (DU) and time unit (TU) are

$$\text{DU} = \frac{2m}{\rho S} \quad \text{and} \quad \text{TU} = \sqrt{\frac{2m}{\rho g S}} \tag{62}$$

The search space is defined by the inequalities $-1 \leq \lambda_{k,0} \leq 1$ and $1\,\text{TU} \leq t_f \leq 6\,\text{TU}$. It is worth remarking that ignorability of the parameter condition allows defining arbitrarily the range in which the initial values of the adjoint variables are sought. Their correct values (fulfilling also the parameter condition (59)) can be recovered a posteriori, as discussed previously. PSO is used with the intent of minimizing the objective

$$\tilde{J} = \sum_{k=1}^{3} |d_k| \tag{63}$$

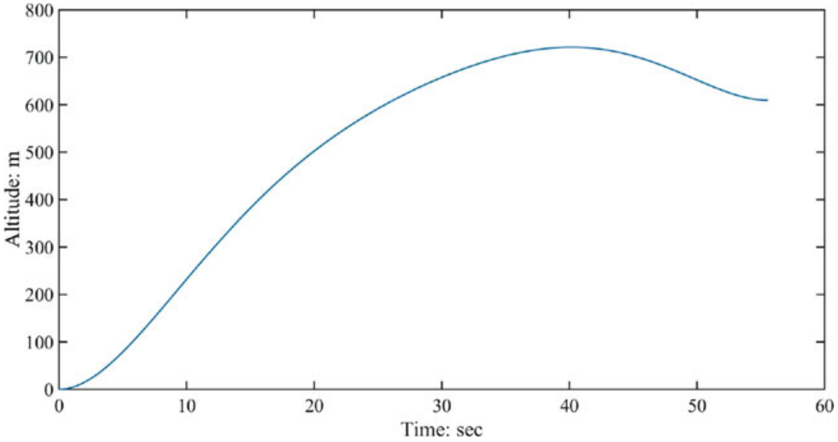


Fig. 4 727 aircraft: nominal altitude

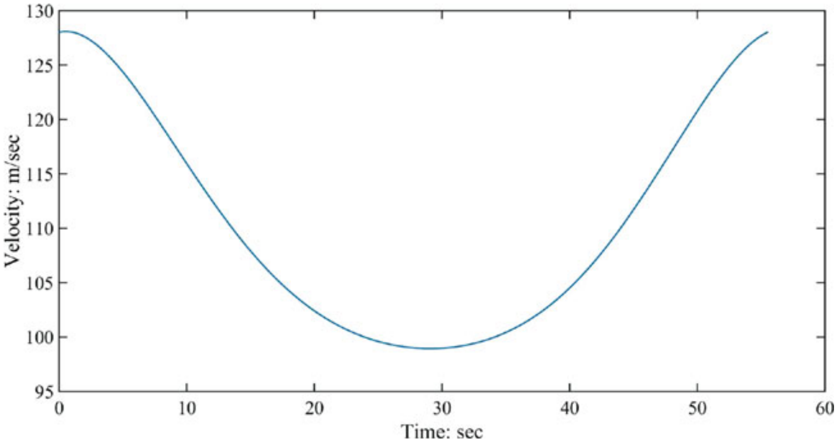


Fig. 5 727 aircraft: nominal velocity

where each term d_k represents a final constraint violation. While minimization of (63) ensures feasibility, enforcement of the necessary conditions for optimality guarantees that the solution found by PSO is (at least locally) optimal.

The PSO algorithm yields a solution associated with $\tilde{J} = 8.458 \times 10^{-7}$. The corresponding optimal time histories of the state and control components are portrayed in Figs. 4, 5, 6, and 7. From their inspection it is apparent that the minimum climbing path is composed of two phases: an initial ascent phase, up to an altitude greater than the final one, followed by a diving phase. The minimum time to climb equals 55.5 s.

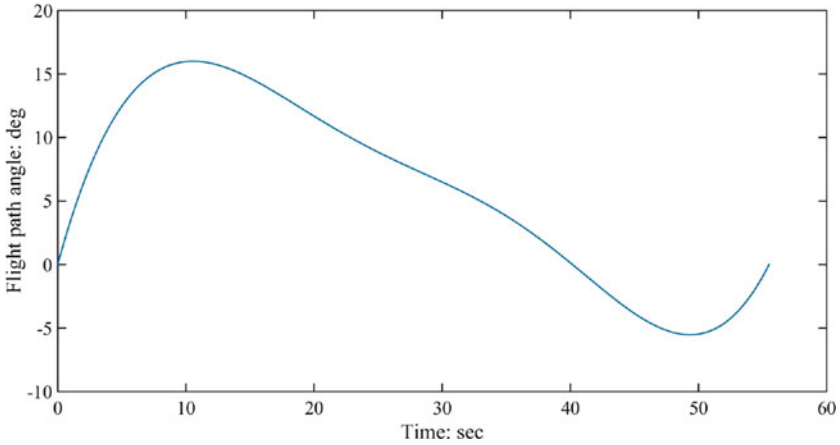


Fig. 6 727 aircraft: nominal flight path angle

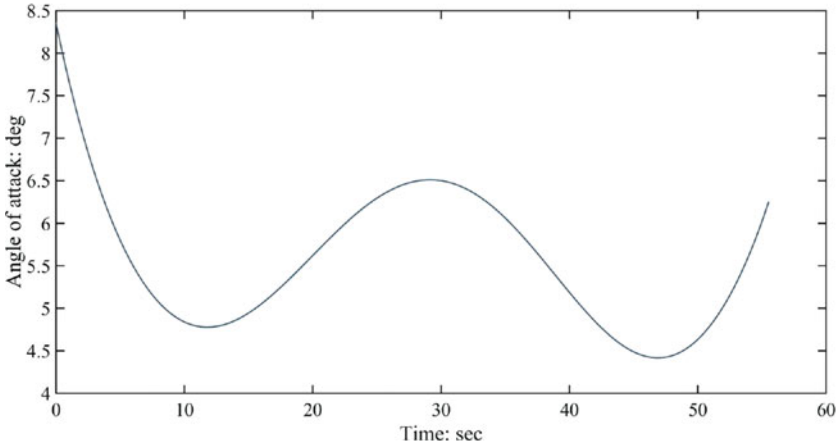


Fig. 7 727 aircraft: optimal control time history

4.3 Application of VTD-NOG

The neighboring optimal guidance algorithm proposed in this work is applied to the minimum-time-to-climb path of the Boeing 727 aircraft. Perturbations from the nominal situation are considered, in order to simulate a realistic scenario. In particular, perturbations on the initial state, thrust magnitude, and atmospheric density are taken into account. Several Monte Carlo campaigns are run, with the intent of obtaining some useful statistical information on the accuracy of the algorithm at hand, in the presence of the previously mentioned deviations, which are simulated stochastically. Monte Carlo campaigns test the guidance algorithm by running a significant number of numerical simulations. Each perturbed quantity in the initial state is associated with a Gaussian distribution, with mean value equal to the respective nominal one and with a specific σ-value, which is related to the statistical dispersion about the mean value. The σ-values are

$$\Delta z^{(\sigma)} = 5\,\mathrm{m}, \quad \Delta v^{(\sigma)} = 5\,\mathrm{m/s}, \quad \Delta \gamma^{(\sigma)} = 5\,\mathrm{deg} \tag{64}$$

In the numerical simulations the deviations from the nominal values are constrained to the intervals $\left[0, 2\Delta\chi^{(\sigma)}\right]$ ($\chi = z$ or γ) and $\left[-2\Delta v^{(\sigma)}, 2\Delta v^{(\sigma)}\right]$. A different approach was chosen for the perturbation of the thrust magnitude. In fact, usually the thrust magnitude exhibits small fluctuations. This time-varying behavior is modeled through a trigonometric series with random coefficients,

$$T_{pert} = T\left\{1 + \sum_{k=1}^{5} a_k \sin(2k\pi\tau) + \sum_{k=1}^{5} a_{k+5} \cos(2k\pi\tau)\right\} \tag{65}$$

where T_{pert} denotes the perturbed thrust, whereas a_k represents a random number with Gaussian distribution, zero mean, and standard deviation equal to 0.01. The atmospheric density fluctuations are modeled through a similar trigonometric series,

$$\rho_{pert} = \rho\left\{1 + \sum_{k=1}^{5} b_k \sin(2k\pi\tau) + \sum_{k=1}^{5} b_{k+5} \cos(2k\pi\tau)\right\} \tag{66}$$

where ρ_{pert} denotes the perturbed density, whereas b_k represents a random number with the same statistic properties as a_k.

At the end of the algorithmic process described in Sect. 3.3, two statistical quantities are evaluated, i.e. the mean value and the standard deviation for all of the outputs of interest. The symbols $\bar{\chi}$ and $\chi^{(\sigma)}$ will denote the mean value and standard deviation of χ ($\chi = z$ or v or γ or t_f) henceforth. Five campaigns are performed, each including 100 runs. The first four campaigns (MC1 through MC4) use a sampling time $\Delta t_S = 2\,\mathrm{s}$, whereas MC5 adopts a sampling time $\Delta t_S = 1\,\mathrm{s}$. MC1 assumes only perturbations of the initial state. The second campaign (MC2) considers the thrust fluctuations, whereas MC3 takes into account only the atmospheric density perturbations. MC4 and MC5 include all the deviations from the nominal flight conditions (with different sampling times). Table 1 summarizes the results for the five Monte Carlo campaigns and reports the related statistics. Application of VTD-NOG to the problem of interest leads to excellent results, with modest errors on the desired final conditions. More specifically, inspection of Table 1 points out that errors on the initial conditions are corrected very effectively; however, also the remaining results exhibit modest deviations at the final time. The latter is extremely close to the optimal value, and this is an intrinsic characteristic of VTD-NOG, which employs first order expansions of the state and costate equations in the proximity of the optimal solution. As a final remark, in the presence of all of

Table 1 Statistics on the time of flight and the errors on altitude, velocity, and flight path angle

Statistics	MC1	MC2	MC3	MC4	MC5
$\Delta\bar{z}_f$ (cm)	1.2	−11.5	8.8	−5.3	1.1
$\Delta z_f^{(\sigma)}$ (cm)	6.3	39.7	14.9	57.3	22.4
$\Delta\bar{v}_f$ (cm/s)	−6.1	−5.9	−6.7	−14.7	−9.7
$\Delta v_f^{(\sigma)}$ (cm/s)	1.7	6.9	6.6	41.6	13.2
$\Delta\bar{\gamma}_f$ (deg)	−0.225	−0.232	−0.268	−0.487	−0.304
$\Delta\gamma_f^{(\sigma)}$ (deg)	0.099	0.166	0.123	1.205	0.356
$\bar{t}_f$ (s)	53.79	53.40	55.34	53.77	53.76
$t_f^{(\sigma)}$ (s)	4.97	0.13	0.33	4.95	4.92

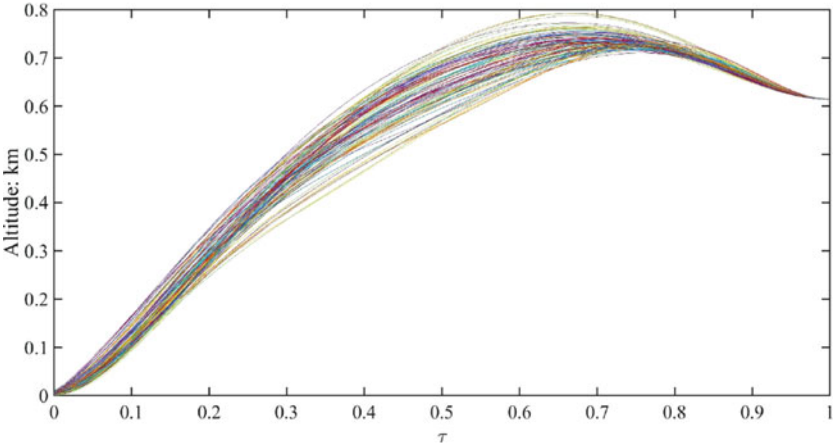


Fig. 8 727 aircraft: altitude time histories obtained in MC5

the perturbations (MC4 and MC5), decreasing the sampling time implies a reduction of the final errors. Figures 8, 9, 10, and 11 depict the perturbed state components and control angle, obtained in MC5.

5 Interception

As a second application of VTD-NOG, this section considers the interception of a target by a maneuvering vehicle in exoatmospheric flight or in the presence of negligible aerodynamic forces, e.g. an intercepting rocket operating at high altitudes. Both the pursuing vehicle and the target are modeled as point masses, in the context of a three-degree-of-freedom problem.

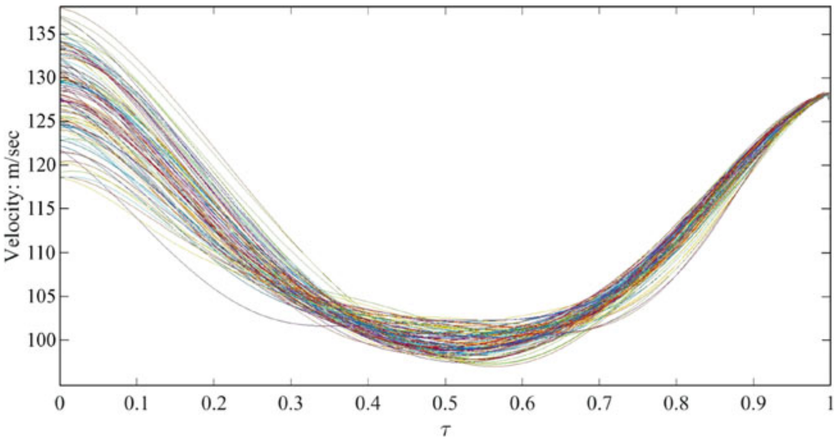


Fig. 9 727 aircraft: velocity time histories obtained in MC5

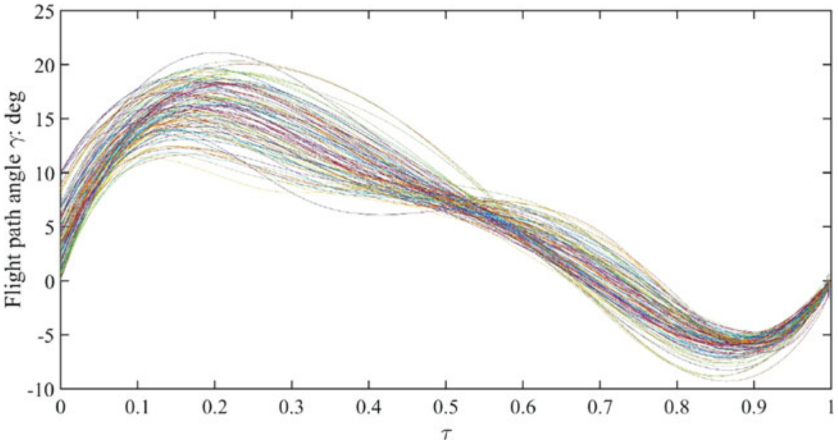


Fig. 10 727 aircraft: time histories of the flight path angle obtained in MC5

5.1 Problem Definition

Under the assumption that interception occurs in a sufficiently short time interval, the flat Earth approximation can be adopted again. This means that the Cartesian reference frame can be defined as follows: the x_1-axis is aligned with the local upward direction, the x_2-axis is directed eastward, and the x_3-axis is aligned with the local North direction. As a result, the Cartesian equations of motion for the intercepting rocket are

$$\dot{x}_1 = t_f x_4 \tag{67}$$

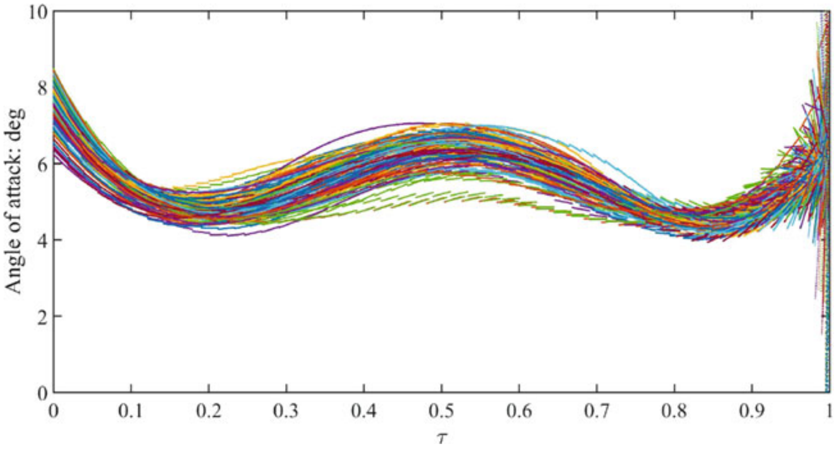


Fig. 11 727 aircraft: time histories of the angle of attack obtained in MC5

$$\dot{x}_2 = t_f x_5 \tag{68}$$

$$\dot{x}_3 = t_f x_6 \tag{69}$$

$$\dot{x}_4 = t_f(-g + a_T \cos u_2 \sin u_1) \tag{70}$$

$$\dot{x}_5 = t_f a_T \cos u_2 \cos u_1 \tag{71}$$

$$\dot{x}_6 = t_f a_T \sin u_2 \tag{72}$$

where the derivatives are written with respect to τ, $\{x_1, x_2, x_3\}$ are the three position coordinates, $\{x_4, x_5, x_6\}$ are the corresponding velocity components, and t_f represents the time of flight up to interception. The symbol g denotes the magnitude of the (constant) gravitational force at the reference altitude, whereas the thrust acceleration has magnitude a_T and direction identified through the two angles u_1 and u_2. The target position is assumed as known, and therefore it is expressed by three specified functions of the dimensionless time τ,

$$x_1^{(T)} = f_1(\tau), \qquad x_2^{(T)} = f_2(\tau), \qquad x_3^{(T)} = f_3(\tau) \tag{73}$$

While the state vector contains the position and velocity components $\{x_i\}_{i=1,\ldots,6}$, the control vector is $\mathbf{u} = [u_1 \quad u_2]^T$, and the parameter vector $\mathbf{a}$ includes only t_f (as in the previous application). As the time is to be minimized, the objective function is

$$J = t_f \tag{74}$$

5.2 Optimal Trajectory

The first-order conditions for optimality are obtained after introducing the Hamiltonian H and the function of the boundary conditions Φ, according to (7)

$$H = \lambda_1 t_f x_4 + \lambda_2 t_f x_5 + \lambda_3 t_f x_6 + \lambda_4 t_f (-g + a_T \cos u_2 \sin u_1) \\ +\lambda_5 t_f a_T \cos u_2 \cos u_1 + \lambda_6 t_f a_T \sin u_2 \tag{75}$$

$$\Phi = \upsilon_1 \left[x_{1f} - f_1(1)\right] + \upsilon_2 \left[x_{2f} - f_2(1)\right] + \upsilon_3 \left[x_{3f} - f_3(1)\right] \tag{76}$$

where the subscript f refers to the value of the respective variable at the final time. The adjoint equations (9) in conjunction with the respective boundary conditions (10) for λ_4 through λ_6 lead to

$$\dot{\lambda}_1 = 0 \quad \Rightarrow \quad \lambda_1 = \lambda_{1,0} \tag{77}$$

$$\dot{\lambda}_2 = 0 \quad \Rightarrow \quad \lambda_2 = \lambda_{2,0} \tag{78}$$

$$\dot{\lambda}_3 = 0 \quad \Rightarrow \quad \lambda_3 = \lambda_{3,0} \tag{79}$$

$$\dot{\lambda}_4 = -\lambda_{1,0} t_f \quad \Rightarrow \quad \lambda_4 = \lambda_{1,0} t_f (1-\tau) \tag{80}$$

$$\dot{\lambda}_5 = -\lambda_{2,0} t_f \quad \Rightarrow \quad \lambda_5 = \lambda_{2,0} t_f (1-\tau) \tag{81}$$

$$\dot{\lambda}_6 = -\lambda_{3,0} t_f \quad \Rightarrow \quad \lambda_6 = \lambda_{3,0} t_f (1-\tau) \tag{82}$$

where $\lambda_{i,0}$ denotes the (unknown) initial value of the adjoint variable λ_i. Then, the Pontryagin minimum principle yields

$$u_2 = -\arcsin \frac{\lambda_{3,0}}{\sqrt{\lambda_{1,0}^2 + \lambda_{2,0}^2 + \lambda_{3,0}^2}} \tag{83}$$

$$\sin u_1 = -\frac{\lambda_{1,0}}{\sqrt{\lambda_{1,0}^2 + \lambda_{2,0}^2}} \quad \text{and} \quad \cos u_1 = -\frac{\lambda_{2,0}}{\sqrt{\lambda_{1,0}^2 + \lambda_{2,0}^2}} \tag{84}$$

These relations imply that the optimal thrust direction is time-independent, regardless of the (known) target position. It is relatively straightforward to prove that for the present application the remaining necessary conditions coming from (10) are useless for the purpose of identifying the optimal solution, in the sense that they do not lead to establishing any new relation among the unknowns of the problem, i.e., $\left\{\lambda_{1,0}, \lambda_{2,0}, \lambda_{3,0}, t_f\right\}$. Also the parameter condition (11) can be proven to be ignorable, in a way similar to that used in Sect. 4.2. Moreover, as the two angles u_1 and u_2 are constant, they can be considered as the unknown quantities in place of $\{\lambda_{1,0}, \lambda_{2,0}, \lambda_{3,0}\}$. Under the assumption that a_T is constant, integration of (67)–(72) leads to obtaining the following explicit solution for x_1, x_2, and x_3:

$$x_1 = x_{1,0} + x_{4,0} t_f \tau + \frac{1}{2} a_T (t_f \tau)^2 \cos u_2 \sin u_1 - \frac{1}{2} g (t_f \tau)^2 \tag{85}$$

$$x_2 = x_{2,0} + x_{5,0} t_f \tau + \frac{1}{2} a_T (t_f \tau)^2 \cos u_2 \cos u_1 \tag{86}$$

$$x_3 = x_{3,0} + x_{6,0} t_f \tau + \frac{1}{2} a_T (t_f \tau)^2 \sin u_2 \tag{87}$$

These expressions are evaluated at $\tau = 1$, then they are set equal to the respective position coordinates of the target at $\tau = 1$. From (85)–(87) one obtains the following equations:

$$\begin{aligned}
&(a_T^2 - g^2) t_f^4 + 4 g x_{4,0} t_f^3 - 4 t_f^2 (x_{4,0}^2 + x_{5,0}^2 + x_{6,0}^2) - 4 g t_f^2 \left[f_1(1) - x_{1,0}\right] \\
&+ 8 x_{4,0} t_f \left[f_1(1) - x_{1,0}\right] + 8 x_{5,0} t_f \left[f_2(1) - x_{2,0}\right] + 8 x_{6,0} t_f \left[f_3(1) - x_{3,0}\right] \\
&- 4 \left[f_1(1) - x_{1,0}\right]^2 - 4 \left[f_2(1) - x_{2,0}\right]^2 - 4 \left[f_3(1) - x_{3,0}\right]^2 = 0
\end{aligned} \tag{88}$$

$$u_2 = \arcsin \frac{2 \left[f_3(1) - x_{3,0}\right] - 2 x_{6,0} t_f}{a_T t_f^2} \tag{89}$$

$$\cos u_1 = \frac{2 \left[f_2(1) - x_{2,0}\right] - 2 x_{5,0} t_f}{a_T t_f^2 \cos u_2} \quad \text{and} \quad \sin u_1 = \frac{2 \left[f_1(1) - x_{1,0}\right] - 2 x_{4,0} t_f + g t_f^2}{a_T t_f^2 \cos u_2} \tag{90}$$

Depending on the analytical form of f_1, f_2, and f_3, (88) can either represent a transcendental equation or simplify to a polynomial equation of fourth degree. Once (88) has been solved, calculation of the optimal thrust angles is straightforward, by means of (89) and (90).

5.3 Application of VTD-NOG

The guidance algorithm described in this work is applied to the interception problem in the presence of nonnominal flight conditions. In particular, perturbations on the initial state and oscillations of the thrust acceleration magnitude over time are modeled. Several Monte Carlo campaigns are run, with the intent of obtaining some useful statistical information on the accuracy of the algorithm at hand, in the presence of the previously mentioned deviations, which are simulated stochastically. The nominal initial position is perturbed by a random vector $\boldsymbol{\rho}$: its magnitude ρ is associated with a Gaussian distribution, with standard deviation equal to 5 m and maximal value never exceeding 10 m, whereas the corresponding unit vector $\hat{\rho}$ has direction uniformly distributed over a unit sphere. Similarly, the nominal initial velocity is perturbed by a random vector $\mathbf{w}$: its magnitude w is associated with a Gaussian distribution, with standard deviation equal to 5 m/s and maximal value never exceeding 10 m/s, whereas the corresponding unit vector $\hat{w}$ has direction uniformly distributed over a unit sphere. A different approach is adopted for the perturbation of the thrust acceleration. As the thrust magnitude (and the related acceleration a_T) exhibits fluctuations, the perturbed thrust acceleration is modeled through a trigonometric series,

$$a_{T,pert} = a_T \left\{ 1 + \sum_{k=1}^{5} c_k \sin(2k\pi\tau) + \sum_{k=1}^{5} c_{k+5} \cos(2k\pi\tau) \right\} \tag{91}$$

where a_T is the nominal thrust acceleration, whereas the corresponding (time-varying) perturbed value $a_{T,pert}$ is actually used in the MC simulations. The coefficients $\{c_k\}_{k=1,\ldots,10}$ have a random Gaussian distribution with zero mean and a standard deviation equal to 0.01. At the end of the algorithmic process described in Sect. 3.3, two statistical quantities are evaluated, i.e. the mean value and the standard deviation for all of the outputs of interest (with a notation similar to that adopted for the preceding application).

In this section, three different targets are taken into account. For each of them, four Monte Carlo campaigns have been performed, each including 100 runs. The first campaign (MC1) assumes only perturbations of the initial state. The second campaign (MC2) considers only oscillations of the thrust acceleration magnitude, while the third and fourth campaigns (MC3 and MC4) include both types of perturbations, with different sampling time intervals ($\Delta t_S = 1$ s and $\Delta t_S = 0.5$ s, respectively).

Fixed Target As a first special case, a fixed target is considered. This means that the three functions f_1, f_2, and f_3 equal three prescribed values $x_1^{(T)}$, $x_2^{(T)}$, and $x_3^{(T)}$. As a result, (88) assumes the form of a fourth degree equation; its smallest real root represents the minimum time to interception.

In the numerical example that follows, the reference altitude (needed for defining the value of g) is set to the initial altitude of the intercepting rocket, whereas $a_T = 2g$. The initial state of the pursuing vehicle is

$$x_{1,0} = 30\,\text{km} \quad x_{2,0} = 0\,\text{km} \quad x_{3,0} = 3\,\text{km} \quad x_{4,0} = x_{5,0} = x_{6,0} = 0.1\,\text{km/s} \tag{92}$$

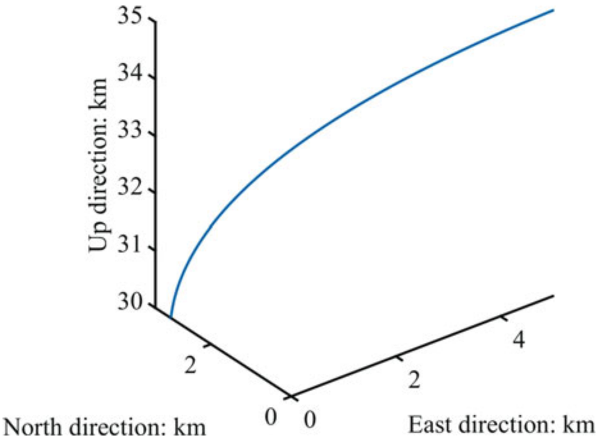


Fig. 12 Optimal interception trajectory of a fixed target

Table 2 Fixed target: statistics on the time of flight and the miss distance

	Δt_S (s)	$\bar{d}_f$ (m)	$d_f^{(\sigma)}$ (m)	$\bar{t}_f$ (s)	$t_f^{(\sigma)}$ (s)
MC1	1	1.71	1.34	30.58	0.22
MC2	1	1.24	0.36	30.57	0.10
MC3	1	1.74	1.04	30.60	0.25
MC4	0.5	1.05	1.50	30.60	0.27

whereas the target position is

$$x_1^{(T)} = 35\,\text{km} \quad x_2^{(T)} = 5\,\text{km} \quad x_3^{(T)} = 0\,\text{km} \tag{93}$$

The minimum time of flight up to interception turns out to equal 30.58 s, whereas the two optimal thrust angles are $u_1^* = 73.3$ deg and $u_2^* = -41.8$ deg. Figure 12 portrays the optimal intercepting trajectory.

Application of VTD-NOG yields results associated with the statistics summarized in Table 2. They regard the miss distance d_f (at the end of nonnominal paths) and the time of flight. Inspection of Table 2 reveals that VTD-NOG generates accurate results, with modest values of the miss distance, which decreases by 40 % from MC3 to MC4. Figure 13 portrays the time evolution of the corrected control angles, obtained in MC3.

Falling Target The second special case assumes a target in free fall (e.g., a ballistic missile at high altitudes), with given initial conditions, denoted with $\left\{x_{i,0}^{(T)}\right\}_{i=1,\ldots,6}$. This means that the three functions f_1, f_2, and f_3 are

$$f_1 = x_{1,0}^{(T)} + x_{4,0}^{(T)} t_f \tau - \frac{1}{2} g (t_f \tau)^2 \qquad f_2 = x_{2,0}^{(T)} + x_{5,0}^{(T)} t_f \tau \qquad f_3 = x_{3,0}^{(T)} + x_{6,0}^{(T)} t_f \tau \tag{94}$$

As a result, (88) assumes again the form of a fourth degree equation; its smallest real root represents the minimum time to interception.

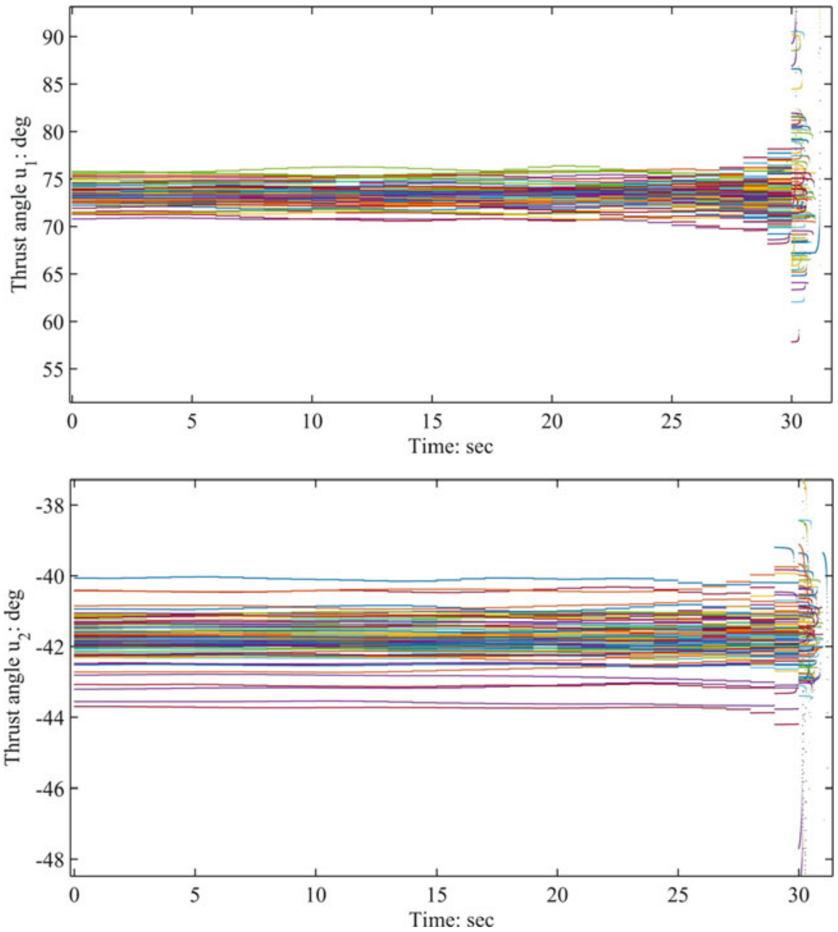


Fig. 13 Fixed target interception: time histories of the control angles obtained in MC3

In the numerical example that follows, the reference altitude is set to the initial altitude of the rocket, whereas $a_T = 3g$. The initial state of the pursuing vehicle is

$$x_{1,0} = 15\,\text{km} \quad x_{2,0} = x_{3,0} = 0\,\text{km} \quad x_{4,0} = 0\,\text{km/s} \quad x_{5,0} = x_{6,0} = 0.2\,\text{km/s} \tag{95}$$

whereas the initial state of the target is given by

$$\begin{aligned} x_{1,0}^{(T)} &= 30\,\text{km} \quad x_{2,0}^{(T)} = x_{3,0}^{(T)} = 1\,\text{km} \\ x_{4,0}^{(T)} &= -1\,\text{km/s} \quad x_{5,0}^{(T)} = x_{6,0}^{(T)} = 0.1\,\text{km/s} \end{aligned} \tag{96}$$

The minimum time of flight up to interception turns out to equal 12.68 s, whereas the two optimal thrust angles are $u_1^* = 96.6$ deg and $u_1^* = -6.5$ deg.

Table 3 Falling target: statistics on the time of flight and the miss distance

	Δt_S (s)	$\bar{d}_f$ (m)	$d_f^{(\sigma)}$ (m)	$\bar{t}_f$ (s)	$t_f^{(\sigma)}$ (s)
MC1	1	0.46	0.27	12.67	0.03
MC2	1	1.81	1.40	12.68	0.01
MC3	1	1.83	1.41	12.67	0.03
MC4	0.5	1.47	1.12	12.68	0.03

The guidance algorithm is applied again in the presence of nonnominal flight conditions. The same Monte Carlo campaigns performed for the previous case are repeated for the application at hand. Table 3 summarizes the results for the four Monte Carlo campaigns and reports the related statistics, with regard to the miss distance (at the end of nonnominal paths) and the time of flight. It is worth remarking that decreasing the sampling time (cf. Table 3) leads again to reducing the mean miss distance. As in the previous application, the actual times of flight are extremely close to the minimal value (12.68 s).

Moving Target The third special case assumes a target that describes a circular path at constant altitude (e.g., an unmanned aerial vehicle). This means that

$$f_1 = x_1^{(T)} \qquad f_2 = x_{2C}^{(T)} + R_T \cos(\omega t_f \tau + \varphi) \qquad f_3 = x_{3C}^{(T)} + R_T \sin(\omega t_f \tau + \varphi) \quad (97)$$

where $x_1^{(T)}$ denotes the target constant altitude, $\left(x_{2C}^{(T)}, x_{3C}^{(T)}\right)$ identify the center of the circular path, whereas ω and φ define, respectively, the angular rate of rotation and the initial angular position. In this case, (88) assumes the form of a transcendental equation, to be solved numerically (for instance, through the Matlab native function *fsolve*). However, numerical solvers need a suitable approximate guess solution to converge to a refined result. For the application at hand, this guess can be easily supplied. In fact, if the radius R_T is sufficiently small, one can assume that the target is located at $\left(x_1^{(T)}, x_{2C}^{(T)}, x_{3C}^{(T)}\right)$. In the presence of a fixed target, an analytical solution exists, and derives from solving a fourth degree equation, as already explained in Sect. 5.2. This solution is used as a guess for the moving target.

In the numerical example that follows, the reference altitude is set to the initial altitude of the rocket, whereas $a_T = 3g$. The initial state of the pursuing vehicle is

$$\begin{aligned} &x_{1,0} = 9\,\text{km} \quad x_{2,0} = x_{3,0} = 0\,\text{km} \\ &x_{4,0} = 0.01\,\text{km/s} \quad x_{5,0} = 0.1\,\text{km/s} \quad x_{6,0} = 0\,\text{km/s} \end{aligned} \quad (98)$$

whereas the fundamental parameters of the target are

$$x_1^{(T)} = 10\,\text{km} \quad x_{2C}^{(T)} = x_{3C}^{(T)} = 2\,\text{km} \quad 2\pi\omega^{-1} = 60\,\text{s} \quad R_T = 0.5\,\text{km} \quad \varphi = 0 \quad (99)$$

The minimum time of flight up to interception turns out to equal 14.76 s, whereas the two optimal thrust angles are $u_1^* = 74.4$ deg and $u_2^* = 51.5$ deg.

Table 4 Moving target: statistics on the time of flight and the miss distance

	Δt_S (s)	$\bar{d}_f$ (m)	$d_f^{(\sigma)}$ (m)	$\bar{t}_f$ (s)	$t_f^{(\sigma)}$ (s)
MC1	1	0.49	0.34	14.73	0.22
MC2	1	2.03	1.31	14.76	0.10
MC3	1	2.15	1.47	14.74	0.25
MC4	0.5	1.90	1.09	14.76	0.11

The guidance algorithm is applied again in the presence of nonnominal flight conditions. The same Monte Carlo campaigns performed for the previous case are repeated for the application at hand. Table 4 summarizes the results for the four Monte Carlo campaigns and reports the related statistics, with regard to the miss distance (at the end of nonnominal paths) and the time of flight. It is worth remarking that decreasing the sampling time (cf. Table 4) leads again to reducing the mean miss distance.

6 Concluding Remarks

This work describes and applies the recently introduced, general-purpose variable-time-domain neighboring optimal guidance algorithm. Usually, all the neighboring optimal guidance schemes require the preliminary determination of an optimal trajectory. Moreover, complex analytical developments accompany the implementation of this kind of perturbative guidance. However, the main difficulties encountered in former formulations of neighboring optimal guidance are the occurrence of singularities for the gain matrices and the challenging implementation of the updating law for the time-to-go. A fundamental original feature of the variable-time-domain neighboring optimal guidance is the use of a normalized time scale as the domain in which the nominal trajectory and the related vectors and matrices are defined. As a favorable consequence, the gain matrices remains finite for the entire time of flight and no extension of their domain is needed. Moreover, the updating formula for the time-to-go derives analytically from the natural extension of the accessory optimization problem associated with the original optimal control problem. This extension leads also to obtaining new equations for the sweep method, which provides all the time-varying gain matrices, computed offline and stored in the onboard computer. In this mathematical framework, the guidance termination criterion finds a logical, consistent definition, and corresponds to the upper bound of the interval to which the normalized time is constrained. Two applications are considered in the paper: (a) minimum-time-to-climb path of a Boeing 727 aircraft and (b) minimum-time exoatmospheric interception of fixed or moving targets. In both cases (especially for (a)), as well as in alternative applications already reported in the scientific literature [16, 17], the variable-time-domain neighboring optimal guidance yields very satisfactory results, with runtime (per simulation) never exceeding the time of flight. This means that VTD-NOG actually represents an effective, general algorithm for the real-time determination of the corrective actions aimed at maintaining an aerospace vehicle in the proximity of its optimal path.

References

1. Afshari, H.H., Novinzadeh, A.B., Roshanian, J.: Determination of nonlinear optimal feedback law for satellite injection problem using neighboring optimal control. Am. J. Appl. Sci. **6**(3), 430–438 (2009)
2. Bryson, A.E.: Dynamic Optimization. Addison Wesley Longman, Boston (1999)
3. Calise, A.J., Melamed, N., Lee, S.: Design and evaluation of a three-dimensional optimal ascent guidance algorithm. J. Guid. Control. Dyn. **21**(6), 867–875 (1998)
4. Charalambous, C.B., Naidu, S.N., Hibey, J.L.: Neighboring optimal trajectories for aeroassisted orbital transfer under uncertainties. J. Guid. Control. Dyn. **18**(3), 478–485 (1995)
5. Chuang, C.-H.: Theory and computation of optimal low- and medium-thrust orbit transfers. NASA-CR-202202, NASA Marshall Space Flight Center, Huntsville (1996)
6. Hull, D.G.: Robust neighboring optimal guidance for the advanced launch system. NASA-CR-192087, Austin (1993)
7. Hull, D.G.: Optimal Control Theory for Applications. Springer, New York (2003)
8. Jo, J.-W., Prussing, J.E.: Procedure for applying second-order conditions in optimal control problems. J. Guid. Control. Dyn. **23**(2), 241–250 (2000)
9. Kugelmann, B., Pesch, H.J.: New general guidance method in constrained optimal control, part 1: numerical method. J. Optim. Theory Appl. **67**(3), 421–435 (1990)
10. Pontani, M., Conway, B.A.: Particle swarm optimization applied to space trajectories. J. Guid. Control. Dyn. **33**(5), 1429–1441 (2010)
11. Pontani, M., Conway, B.A.: Particle swarm optimization applied to impulsive orbital transfers. Acta Astronaut. **74**, 141–155 (2012)
12. Pontani, M., Conway, B.A.: Optimal finite-thrust rendezvous trajectories found via particle Swarm algorithm. J. Spacecr. Rocket. **50**(6), 1222–1234 (2013)
13. Pontani, M., Conway, B.A.: Optimal low-thrust orbital maneuvers via indirect swarming method. J. Optim. Theory Appl. **162**(1), 272–292 (2014)
14. Pontani, M., Ghosh, P., Conway, B.A.: Particle swarm optimization of multiple-burn rendezvous trajectories. J. Guid. Control. Dyn. **35**(4), 1192–1207 (2012)
15. Pontani, M., Cecchetti, G., Teofilatto, P.: Variable-time-domain neighboring optimal guidance, part 1: algorithm structure. J. Optim. Theory Appl. **166**(1), 76–92 (2015)
16. Pontani, M., Cecchetti, G., Teofilatto, P.: Variable-time-domain neighboring optimal guidance, part 2 application to lunar descent and soft landing. J. Optim. Theory Appl. **166**(1), 93–114 (2015)
17. Pontani, M., Cecchetti, G., Teofilatto, P.: Variable-time-domain neighboring optimal guidance applied to space trajectories. Acta Astronaut. **115**, 102–120 (2015)
18. Seywald, H., Cliff, E.M.: Neighboring optimal control based feedback law for the advanced launch system. J. Guid. Control. Dyn. **17**(3), 1154–1162 (1994)
19. Teofilatto, P., De Pasquale, E.: A non-linear adaptive guidance algorithm for last-stage launcher control. J. Aerosp. Eng. **213**, 45–55 (1999)
20. Vinh, N.X.: Optimal Trajectories in Atmospheric Flight. Elsevier, New York (1981)
21. Yan, H., Fahroo, F., Ross, I.: Real-time computation of neighboring optimal control laws. AIAA Guidance, Navigation and Control Conference, Monterey. AIAA Paper 2002–465 (2002)

State-of-the-Art of Optimum Shape Design Methods for Industrial Applications and Beyond

Bruno Stoufflet

Abstract This presentation gives an overview of the capacities of optimal shape design methods as actual engineering tools utilized for industrial applications, mostly for aerodynamic shape design. Numerical formulation and implementation are recalled and illustrations of applications are discussed.

1 Contents

Among all the challenges related to future aircraft concepts and multidisciplinary approaches, optimum shape design is called to play a crucial role to create new configurations of aircraft aiming at exhibiting significant performance improvements.

Some can be quoted:

- Revisited propulsion integration to reduce the installation drag,
- Wings with extended laminar regions,
- Masking effects of empennage to reduce noise footprint.

Taking advantage of the industrial maturity of computational fluid dynamics codes which currently deliver precise simulations of flow, one can seriously consider an industrial utilization of automatic shape design.

The optimization problem whose variables are usually shape parameters is based upon state equations which are the equations of the fluid motion and constraints functions which are either geometric constraints or aerodynamics ones in order to minimize a cost function.

Gradient of the cost function is computed via an adjoint equation of the state equation embarking a mesh deformation equation relating mesh points to shape parameters.

From an engineering standpoint, the implementation of such an approach makes use of adjoint solvers including mesh deformation, parametrization through a CAD description of the shape and extensive library of cost functions.

B. Stoufflet (✉)
Dassault Aviation, 92250 Saint-Cloud, France
e-mail: bruno.stoufflet@dassault-aviation.com

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_17

Automatic differentiation software such as Tapenade (developed at INRIA) turns out to be a key ingredient.

Optimization strategy based on Feasible Sequential Quadratic Programming or Feasible Arc Interior point Algorithm has been implemented.

Several recent computations exhibit the large diversity of applications which can be handled by shape optimization:

- Design of a fuselage shape based on full 3D Navier–Stokes solver to reduce transonic interaction with nacelles,
- Laminar wing optimization to increase laminar portion of the wing close to the fuselage,
- Afterbody optimization of innovative configuration with a cost function based on the boundary layer shape parameter,
- Multipoint optimization at low and high speed of a wing tip involving geometrical parameters as twist, sweep angle, dihedral angle, thickness and span,
- Control and optimization of separated flows in order to optimize the shape and locations of mechanical or fluidic vortex generators. Some investigations of reliability-based design (i.e. find a shape associated with a probability to not realize a target less than a given acceptable value) have also been carried out.

2 Concluding Remarks

In conclusion, it is now assessed that automatic shape optimization is currently daily used for design and accelerates the elementary design cycle and gives access to an enlargement of potential solutions.

Variational Analysis and Euler Equation of the Optimum Propeller Problem

Francesco Torrigiani, Aldo Frediani, and Antonio Dipace

Abstract The problem of the optimum propeller with straight blades was first solved by Goldstein; in this paper, a variational formulation is proposed in order to extend the solution to non-planar blades. First, we find a class of functions (the circulation along the blade axis) for which the thrust and the aerodynamic drag moment are well defined. In this class, the objective functional is proved to be strictly convex and then the global minimum exists and is unique. Then we determine the Euler equation in the case of a general blade and show that the numerical results are consistent with the Goldstein's solution. Finally, some numerical results with the Ritz method are presented for optimum propeller blades.

1 Introduction

The design of optimum propellers is a historical challenge. The first theory, the actuator disk model, was formulated by Rankine [15] in 1865 and Froude [6] in 1878 for naval propeller. The rotor vortex theory was independently developed by Betz and Prandtl [1] in 1919, Pistolesi [13] in 1921 and Joukowsky [9] in 1929. Following this theory, Goldstein in 1929 published a benchmark work [8]; he found the optimal circulation distribution along the blade. Later, researchers extended and organized these results. Glauert [7] in 1935, exploiting all the previous models, founded the blade element theory. Theodorsen [16] in 1948 studied the case of highly loaded propellers. Larrabee [10] in 1979 developed a design procedure to determine the optimal chord and twist distribution along the blade. Other authors kept fixed the chord and studied only the twist distribution. Dorfling [4] in 2015 used this approach and obtained the optimal twist distribution for different flight conditions. Another approach is shown by Chattot [2] in 2000. Firstly, he analysed only the circulation distribution, then, when he evaluated also the viscous effect,

F. Torrigiani (✉) • A. Frediani • A. Dipace
Civil and Industrial Engineering Department, University of Pisa, Largo Lucio Lazzarino, 56122, Pisa, Italy
e-mail: torrigianifrancesco@gmail.com; a.frediani@dia.unipi.it; antoniodipax@gmail.com

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_18

chord and twist distribution were determined. We follow this method but the main focus is on non-classical blade shapes, hence, for the moment, viscous effects are not included in the analysis.

Nowadays, according to the main engine manufacturers, big counter-rotating open rotor engines could be the most efficient solution; an evolution of this concept could be a propeller with non-planar blades.

Airscrew propellers and airscrew turbines are unconventional turbomachines. They have rotating parts and exert force and momentum on the fluid as turbomachines but, contrary to turbomachines, airscrews' blades are not bounded at their tips by a rigid wall and the aerodynamic phenomena occurring at blade tip results in a loss of performance. Moreover, turbomachines do work starting from a difference of pressure or absorb work to produce a difference of pressure, while airscrews do (or absorb) work starting from (or to produce) a difference of kinetic energy on the fluid passed through the blades. Nevertheless, very close to the rotor, a difference of pressure between a section before and a section after the rotor is found and, thus, airscrew can be regarded as the turbomachine with the minimum compression (or expansion) ratio.

The aim of an airscrew propeller is to convert the power provided by the engine (*shaft brake power* P_b) into the available power P_a; if M and Ω indicate the moment and the angular velocity, respectively, we have

$$P_b := M\Omega,$$

and

$$P_a := TV_\infty,$$

where T and V_∞ are the thrust and the asymptotic velocity, respectively. The efficiency of a propeller is defined as follows:

$$E := \frac{P_a}{P_b} = \frac{TV_\infty}{M\Omega} \tag{1}$$

In this paper airscrews in axial flow (see axes in Fig. 1) are analysed. The rotation is opposite to unit vector $\mathbf{i}$. The fluid motion around the airscrew is a viscous, compressible and non-stationary flow. However, it is possible to use simplified theories, with different levels of approximation as: actuator disk model, blade element theory, vortex theory. All these simplified theories are based on the hypotheses of stationary motion, non-viscous fluid, incompressible fluid and asymptotic velocity field, V_∞, parallel to the airscrew axis.

The actuator disk model provides only global information without considering the blades.

The blade element theory provides the aerodynamic action on each section of blade, according to the Kutta-Joukowski law, where, according to reference system

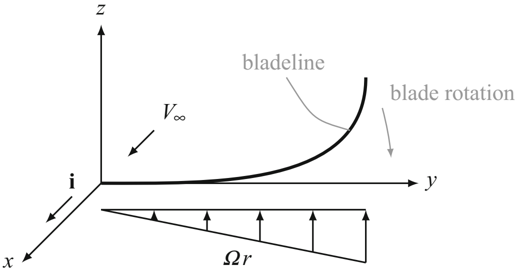


Fig. 1 Reference system

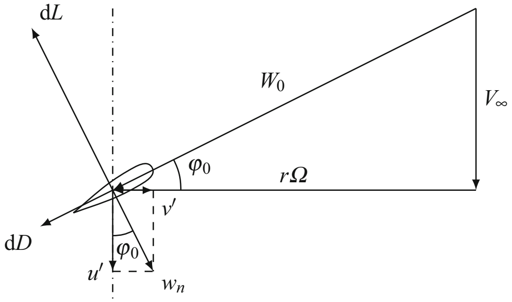


Fig. 2 Velocities and forces on blade section

defined in Fig. 1, the velocity field is composed of the asymptotic, the induced and the rotational velocities.

2 Airscrew Vortex Theory

Rotor vortex theory was the first and the simplest theory to include the induced velocity (using the same concepts of the lifting line theory) to the rotating wing. Figure 2 shows velocities and forces in a plane perpendicular to the blade axis. The components of the forces on a strip dr are

$$\mathrm{d}L = \rho \Gamma W_0 \mathrm{d}r$$

$$\mathrm{d}D = \rho \Gamma w_n \mathrm{d}r$$

where Γ is the circulation around the blade section, W_0 is the asymptotic ($V_\infty \mathbf{i}$) and rotational ($\Omega r \mathbf{e}_\theta$) composition ($\mathbf{e}_\theta$ is the rotational unit vector), w_n is the component of the induced velocity perpendicular to W_0 and to the blade axis. The thrust is given by the component of dL along the rotor axis if the effect of the drag is negligible:

$$\mathrm{d}T = \rho \Gamma W_0 \cos \varphi_0 \mathrm{d}r\ .$$

The lost power, P_l, is the work of the drag per unit time, i.e.:

$$\mathrm{d}P_l = \rho \Gamma w_n W_0 \mathrm{d}r \; .$$

The optimum problem can be formulated as follows:
to find the distribution of circulation Γ along the blade that minimize the lost power P_l for a given thrust T_{target}.
The Lagrangian, $\mathscr{L}$, of the problem is

$$\mathscr{L}(\Gamma) = N \int_0^R \rho \Gamma w_n W_0 \, \mathrm{d}r + \lambda \left(N \int_0^R \rho \Gamma W_0 \cos\varphi_0 \, \mathrm{d}r - T_{target} \right) \; , \tag{2}$$

where N is the number of the blades. Now, we assume (according to [14]) that the induced velocity w_n is not dependent on a perturbation of Γ; thus, the Euler equation becomes

$$N\rho w_n W_0 + \lambda N \rho W_0 \cos\varphi_0 = 0 \; , \tag{3}$$

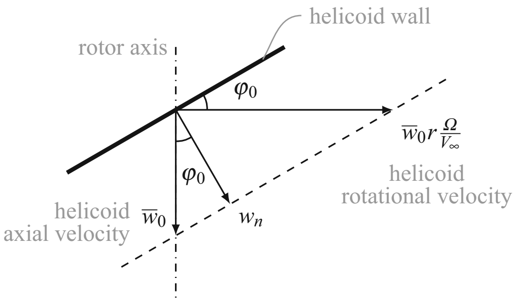
and

$$w_n = -\lambda \cos\varphi_0 \; , \tag{4}$$

where λ is a constant depending on T_{target}. Now, if we assume the wake helicoids as rigid walls moving in the axial direction with velocity $\overline{w}_0 := -\lambda$, we obtain exactly the same expression[1] (4) of w_n (Fig. 3). If we consider the distribution of circulation Γ along the blade and the intensity, γ, of a single vortex filament in the wake, we obtain, by means of the Stokes' theorem, the following relationship:

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}\eta} = -\gamma \tag{5}$$

Fig. 3 Velocity of the rigid helicoids near the rotor



[1] Note that, from Fig. 2, $\cos\varphi_0/\sin\varphi_0 = \Omega r/V_\infty$, the same expression of w_n is obtained also if the helicoids rotate with angular velocity $\overline{w}_0 \Omega/V_\infty$.

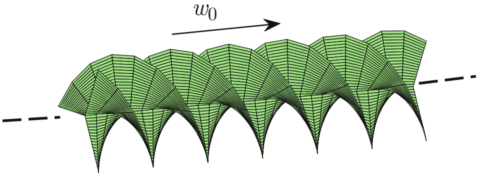


Fig. 4 Geometry of the potential problem

where η is a curvilinear coordinate along the blade. From the Stokes theorem the intensity γ of a free filament is constant along the wake and the distribution of circulation at any station of the wake is the same as the one on the blade. Thus, we can study the flow in the far wake instead of the one near the rotor and therefore the potential problem becomes

Find the flow around an infinite helicoid moving in the axial direction with velocity[2] w_0 (Fig. 4):

$$w_0 := 2\overline{w}_0 \ .$$

In order to obtain the value of the constant w_0, we need to introduce the constraint on thrust:

$$N \int_0^R \rho \Gamma W_0 \cos \varphi_0 \, \mathrm{d}r = T_{target} \ .$$

Prandtl [1] proposed the following approximated solution of problem (2):

$$\frac{N\Gamma\Omega}{2\pi w_0 V_\infty} = \frac{\mu^2}{1+\mu^2}\left[\frac{2}{\pi}\arccos\left(\mathrm{e}^{-\frac{\pi}{h}(R-r)}\right)\right] , \tag{6}$$

where h is the distance between the spirals of the wake helicoid, and $\mu = r\Omega/V_\infty$ is the dimensionless coordinate along the blade. Goldstein [8] gave an exact solution of the problem by means of an expansion of Bessel functions. Figure 5 shows the exact optimal dimensionless circulations, $N\Gamma\Omega/2\pi V_\infty w_0$, and the Prandtl solutions in the case of a two bladed propeller, versus μ. The different curves in Fig. 5 are relevant to different values of the non-dimensional tip speed, $\mu_0 := R\Omega/V_\infty = 2, 3, \ldots 10$.

[2] The induced velocity in the far wake is the double of the one near the rotor; this result is usually known as the Froude theorem.

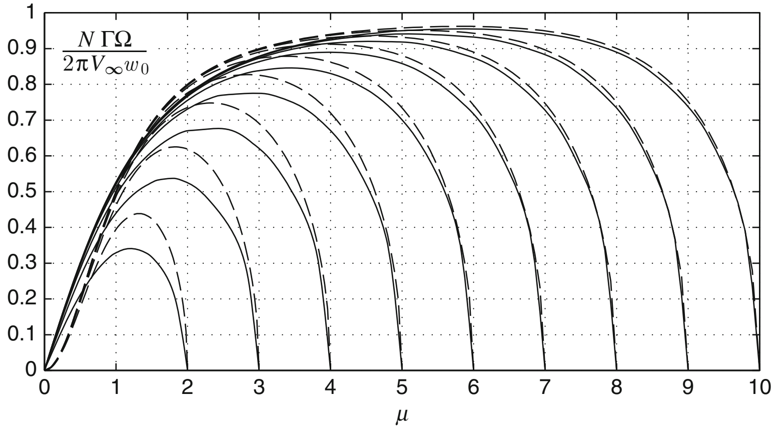


Fig. 5 Goldstein (*solid line*) and Prandtl (*dashed line*) optimal circulation distribution for a two blade propeller

3 Variational Formulation

The force exerted by the fluid on the blade section is given by the Kutta-Joukowski law:

$$\mathrm{d}\mathbf{F}(\mathbf{r}) = \rho\mathbf{U_{tot}}(\mathbf{r}) \times \Gamma(\mathbf{r})\mathrm{d}\mathbf{r}\ ,$$

where $\mathbf{r}$ is the position vector according to Fig. 6 and $\mathbf{U_{tot}}$ is the flow velocity, namely:

$$\mathbf{U_{tot}} = V_\infty\mathbf{i} + \Omega r\mathbf{e}_\theta + \mathbf{u}_{ind}\ , \tag{7}$$

where Ω is the airscrew angular velocity, r is the distance from the rotor axis[3] and $\mathbf{e}_\theta$ is the rotational unit vector. We refer just to right propeller, that means: $\mathbf{e}_\theta$ follows the right-hand rule with respect to the x axis of rotation.

3.1 Induced Velocity

According to the vortex theory, airscrew and wake are replaced by bounded and free vortex filaments, respectively. Blade line is described by a curvilinear coordinate ξ, defined on [0; 1]. Induced velocity $\mathbf{u}_{ind}$ is obtained by Biot-Savart law by assuming

[3] r corresponds to the module of position vector only for plain blades in the plane (y, z); in this case, we have $\mathbf{r} \times \mathbf{e}_\theta = r\mathbf{i}$.

Fig. 6 Reference system

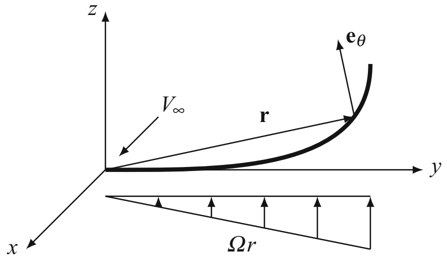
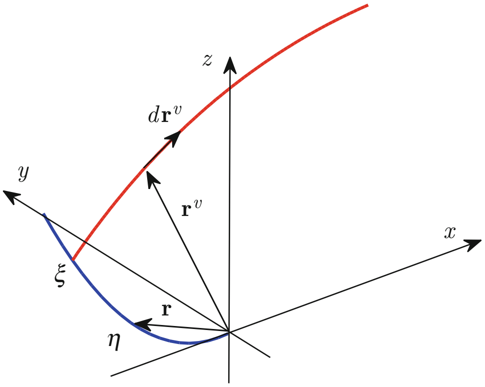


Fig. 7 Vector definition



that the effect of the bounded filaments can be neglected (according to [3]); more details are given in Appendix 5. The contribution of a single free filament originated from a point ξ on the blade line is

$$\mathrm{d}\mathbf{u}_{ind}(\mathbf{r},\xi) = \gamma(\xi)\mathrm{d}\xi \left(\frac{1}{4\pi} \int_{\gamma_v} \mathrm{d}\mathbf{r}^v(t) \times \frac{\mathbf{r}(\eta) - \mathbf{r}^v(t,\xi)}{\|\mathbf{r}(\eta) - \mathbf{r}^v(t,\xi)\|^3} \right) . \tag{8}$$

The superscript ^{v} indicates the inducing vortex filament (Fig. 7). Note that the vortex filament $\gamma_v(\xi)$ is completely defined by the origin on the blade ξ and a point on this filament is identified by $\mathbf{r}^v = \mathbf{r}^v(t,\xi)$, where $t \in [0;\infty)$ is a curvilinear coordinate defined on the filament:

$$\mathbf{r}^v : [0;\infty) \times [0;1] \to \mathbb{R}^3 ,$$

$$\mathbf{r}^v(t,\xi) = \begin{cases} x^v = d(\xi) + V_\infty t \\ y^v = m(\xi)\cos(\Omega t + \theta(\xi)) \\ z^v = m(\xi)\sin(\Omega t + \theta(\xi)) \end{cases} , \tag{9}$$

$m(\xi)$ and $\theta(\xi)$ are the module and the phase of the projection of $\mathbf{r}(\xi)$ on the plane of rotation (y, z), $d(\xi)$ is the distance of $\mathbf{r}(\xi)$ from the plane of rotation (y, z) and

$$\mathbf{r}(\xi) = \begin{cases} x = d(\xi) \\ y = m(\xi)\cos\theta(\xi) \\ z = m(\xi)\sin\theta(\xi) \end{cases}$$

is the position vector of the origin of the filament on the blade line. From (5), Eq. (8) becomes

$$\mathrm{d}\mathbf{u}_{ind}(\eta, \xi) = -\frac{\mathrm{d}\Gamma(\xi)}{4\pi}\int_{\gamma_v} \mathrm{d}\mathbf{r}^v(t) \times \frac{\mathbf{r}(\eta) - \mathbf{r}^v(t, \xi)}{\|\mathbf{r}(\eta) - \mathbf{r}^v(t, \xi)\|^3}\,, \tag{10}$$

where η indicates the point of the blade line where the induced velocity is calculated. Then, all the contributions of free filaments are summed up along the blade span:

$$\mathbf{u}_{ind}(\eta) = \int_{\gamma_b} \mathrm{d}\mathbf{u}_{ind}(\eta, \xi) = -\frac{1}{4\pi}\int_{\gamma_b} \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi}\int_{\gamma_v} \mathrm{d}\mathbf{r}^v(t) \times \frac{\mathbf{r}(\eta) - \mathbf{r}^v(t, \xi)}{\|\mathbf{r}(\eta) - \mathbf{r}^v(t, \xi)\|^3}\mathrm{d}\xi, \tag{11}$$

where γ_b is the blade line. Equation (11) represents the velocity induced by the i^{th} helicoid ($i \in [0; N - 1]$ where N is the number of blades) and (11) is modified as follows:

$$\mathbf{u}^i_{ind}(\eta) = -\frac{1}{4\pi}\int_{\gamma^i_b} \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi}\int_{\gamma^i_v} \mathrm{d}\mathbf{r}^v_i(t) \times \frac{\mathbf{r}(\eta) - \mathbf{r}^v_i(t, \xi)}{\left\|\mathbf{r}(\eta) - \mathbf{r}^v_i(t, \xi)\right\|^3}\mathrm{d}\xi \tag{12}$$

The superscript ^{0} refers to the reference blade, which is the one where the induced velocity is calculated. The resultant induced velocity is the sum of N blades contributions:

$$\mathbf{u}_{ind}(\eta) = \sum_{i=0}^{N-1} \mathbf{u}^i_{ind}(\eta)\,.$$

By integrating a second time along the reference blade we obtain the resultant force and momentum exerted by the flow:

$$\mathbf{F} = \int_{\gamma^0_b} \rho\Gamma(\mathbf{r})\,(\mathbf{U_{tot}}(\mathbf{r}) \times \mathrm{d}\mathbf{r})\ ,$$

$$\mathbf{M} = \int_{\gamma^0_b} \rho\Gamma(\mathbf{r})\mathbf{r} \times (\mathbf{U_{tot}}(\mathbf{r}) \times \mathrm{d}\mathbf{r})\ .$$

Note that the total force on the propeller is given by the one on the reference blade times the number of blades N.

3.2 Dimensionless Formulation

We define the following non-dimensional quantities:

$$\begin{cases} \tilde{\mathbf{r}} = \dfrac{\mathbf{r}}{R} & \mu_0 = \dfrac{\varOmega R}{V_\infty} \\ \tilde{\mathbf{u}} = \dfrac{\mathbf{u}}{V_\infty} & \tilde{\varGamma} = \dfrac{\varGamma}{RV_\infty} \\ \tilde{t} = \dfrac{V_\infty t}{R} & \end{cases}$$

$$\mathrm{d}\tilde{\mathbf{F}} = \frac{\mathrm{d}\mathbf{F}}{\frac{1}{2}\rho V_\infty^2 R^2} = 2\,(\mathbf{i} + \mu_0 \tilde{r}\mathbf{e}_\theta + \tilde{\mathbf{u}}_{ind}) \times \tilde{\varGamma}\mathrm{d}\tilde{\mathbf{r}}\ , \tag{13}$$

where R is the external rotor radius and $\mu_0 = \varOmega R/V_\infty$ the *tip speed ratio*. From now on, we will handle only these non-dimensional variables and therefore, for the sake of brevity, we omit $\tilde{}$. Thrust coefficient C_T and momentum coefficient C_M are defined as follows:

$$C_T := -N\int_{\gamma_b^0} \mathrm{d}\mathbf{F}(\mathbf{r})\cdot\mathbf{i} = -2N\int_{\gamma_b^0} \varGamma(\mathbf{r})\,(\mathbf{i} + \mu_0 r\mathbf{e}_\theta + \mathbf{u}_{ind}) \times \mathrm{d}\mathbf{r}\cdot\mathbf{i}\ , \tag{14}$$

$$C_M := N\int_{\gamma_b^0} \mathbf{r}\times\mathrm{d}\mathbf{F}(\mathbf{r})\cdot\mathbf{i} = 2N\int_{\gamma_b^0} \varGamma(\mathbf{r})\mathbf{r}\times[(\mathbf{i} + \mu_0 r\mathbf{e}_\theta + \mathbf{u}_{ind}) \times \mathrm{d}\mathbf{r}]\cdot\mathbf{i}\ . \tag{15}$$

These equations show that thrust and momentum, for blade line γ_b, depend on circulation distribution $\varGamma(\eta)$. In stationary condition $C_T = C_{T\,target}$, and we can formulate the problem of the *optimum propeller*.

Find the circulation function, $\varGamma$, belonging to a function set, χ, so that the moment of drag force, C_M, is minimum, with the constraint $C_T = C_{T\,target}$.

That is:

$$\min_{\varGamma}\left\{2N\int_{\gamma_b^0} \varGamma(\mathbf{r})\mathbf{r}\times[(\mathbf{i} + \mu_0 r\mathbf{e}_\theta + \mathbf{u}_{ind}) \times \mathrm{d}\mathbf{r}]\cdot\mathbf{i}\right\} \tag{16}$$

subjected to

$$2N\int_{\gamma_b^0} \varGamma(\mathbf{r})\,(\mathbf{i} + \mu_0 r\mathbf{e}_\theta + \mathbf{u}_{ind}) \times \mathrm{d}\mathbf{r}\cdot\mathbf{i} + C_{T\,target} = 0\ , \tag{17}$$

with

$$\varGamma \in \chi \ , \quad \chi = \{f : [0;1] \to \mathbb{R} : f(0) = f(1) = 0\} \ , \tag{18}$$

and where

$$\mathbf{u}_{ind}(\mathbf{r}) = -\frac{1}{4\pi} \sum_{i=0}^{N-1} \int_{\gamma_b^i} \frac{\mathrm{d}\varGamma(\xi)}{\mathrm{d}\xi} \int_{\gamma_v^i} \mathrm{d}\mathbf{r}_i^v(t) \times \frac{\mathbf{r} - \mathbf{r}_i^v(t,\xi)}{\left\|\mathbf{r} - \mathbf{r}_i^v(t,\xi)\right\|^3} \mathrm{d}\xi \ . \tag{19}$$

Note that the solution of the problem depends on the blade line and, also on three scalar quantities: number of blades N, tip speed ratio μ_0, necessary thrust $C_{T\,target}$.

The function class, $\mathscr{X}$, where the solution is sought must have some features. Firstly, it must be a subset of χ, a set of real value functions defined on [0; 1] with $\varGamma(0) = \varGamma(1) = 0$. Moreover, for $\varGamma \in \mathscr{X}$, we define the integrals of C_T and C_M on $T \times T \times [0;\infty)$, where for $(t, \xi, \eta) = (0, \eta, \eta)$ we have the value of the limit of the integrand; in this way, the extended functionals are continuous. After having restricted χ in order to guarantee the existence and continuity of the functionals, we must assure the existence of the constrained minimum, that is the solution of the problem (16), (17) and (18). This can be achieved in several ways. To this end we restrict χ in order to make C_M convex; then, we seek for an extremal function for C_M which also respects the constraint on C_T. Extremal means that makes null the first variation of the functional. If we find this function, then this is the minimum function and it is unique. In Appendix 2 and 3 it is proved that a class of functions, having all these features, is the following

$$\mathscr{X} = \left\{\varGamma \in AC[0;1],\ \varGamma,\ \varGamma' \in L^{1+\varepsilon}(0;1) \text{ with } \varepsilon > 0,\ \varGamma(0) = \varGamma(1) = 0,\ C_{M_{ind}}(\varGamma) > 0\right\} \ , \tag{20}$$

where $AC[0;1]$ are absolute continuous functions defined on $[0,1], L^p$ is the Lebesgue space, $C_{M_{ind}}$ is the momentum associated to the induced velocity $\mathbf{u}_{ind}$:

$$C_{M_{ind}} = 2N \int_{\gamma_b^0} \varGamma(\mathbf{r})\mathbf{r} \times (\mathbf{u}_{ind} \times \mathrm{d}\mathbf{r}) \cdot \mathbf{i} \ .$$

We use the method of Lagrangian multipliers. This is an isoperimetrical problem, hence, the multipliers are scalar. The Lagrangian is

$$\mathscr{L}(\varGamma, \lambda) = C_M(\varGamma) + \lambda \left(C_T(\varGamma) - C_{T\,target}\right) \ . \tag{21}$$

In order to calculate the extremal function of $\mathscr{L}(\varGamma, \lambda)$ we use a direct method, in particular a Ritz method [5]. A different approach is given in Sect. 4 where the Lagrangian is differentiated and the explicit Euler equation is obtained.

4 Euler Equation

According to (14), (15) and (19) we define

$$C_M(\Gamma) = C_{M_1}(\Gamma) + C_{M_2}(\Gamma)$$
$$\begin{cases} C_{M_1}(\Gamma) := 2N \int_0^1 \Gamma(\eta)\mathbf{r}(\eta) \times [(\mathbf{i} + \mu_0 r(\eta)\mathbf{e}_\theta(\eta)) \times \boldsymbol{\tau}(\eta)] \cdot \mathbf{i}\, \mathrm{d}\eta \\ C_{M_2}(\Gamma) := -\frac{1}{2\pi} \int_0^1 \int_0^1 \Gamma(\eta)\Gamma'(\xi)\mathbf{r}(\eta) \times \left(\sum_{i=0}^{N-1} \int_0^\infty \mathbf{g}_i(t,\xi,\eta)\mathrm{d}t \times \boldsymbol{\tau}(\eta) \right) \cdot \mathbf{i}\, \mathrm{d}\xi \mathrm{d}\eta \end{cases}$$

$$C_T(\Gamma) = C_{T_1}(\Gamma) + C_{T_2}(\Gamma)$$
$$\begin{cases} C_{T_1}(\Gamma) := -2N \int_0^1 \Gamma(\eta)\mu_0 r(\eta)\mathbf{e}_\theta(\eta) \times \boldsymbol{\tau}(\eta) \cdot \mathbf{i}\, \mathrm{d}\eta \\ C_{T_2}(\Gamma) := \frac{1}{2\pi} \int_0^1 \int_0^1 \Gamma(\eta)\Gamma'(\xi) \sum_{i=0}^{N-1} \int_0^\infty \mathbf{g}_i(t,\xi,\eta)\mathrm{d}t \times \boldsymbol{\tau}(\eta) \cdot \mathbf{i}\, \mathrm{d}\xi \mathrm{d}\eta \end{cases}$$

where $\boldsymbol{\tau}(\eta) := \mathrm{d}\mathbf{r}/\mathrm{d}(\eta)$ and

$$\mathbf{g}_i(t,\xi,\eta) := \frac{\mathrm{d}\mathbf{r}_i^v(t)}{dt} \times \frac{\mathbf{r}(\eta) - \mathbf{r}_i^v(t,\xi)}{\left\| \mathbf{r}(\eta) - \mathbf{r}_i^v(t,\xi) \right\|^3} .$$

We define

$$M_1(\eta) := 2N\, \mathbf{r}(\eta) \times [(\mathbf{i} + \mu_0 r(\eta)\mathbf{e}_\theta(\eta)) \times \boldsymbol{\tau}(\eta)] \cdot \mathbf{i}\,, \tag{22}$$

$$M_2(\xi,\eta) := -\frac{1}{2\pi}\mathbf{r}(\eta) \times \left(\sum_{i=0}^{N-1} \int_0^\infty \mathbf{g}_i(t,\xi,\eta)\mathrm{d}t \times \boldsymbol{\tau}(\eta) \right) \cdot \mathbf{i}\,, \tag{23}$$

$$T_1(\eta) := -2N\, \mu_0 r(\eta)\mathbf{e}_\theta(\eta) \times \boldsymbol{\tau}(\eta) \cdot \mathbf{i}\,, \tag{24}$$

$$T_2(\xi,\eta) := \frac{1}{2\pi} \sum_{i=0}^{N-1} \int_0^\infty \mathbf{g}_i(t,\xi,\eta)\mathrm{d}t \times \boldsymbol{\tau}(\eta) \cdot \mathbf{i}\,. \tag{25}$$

The Lagrangian associated with the problem is

$$\begin{aligned} \mathscr{L}(\Gamma,\lambda) := &\int_0^1 \Gamma(\eta)M_1(\eta)\mathrm{d}\eta + \int_0^1 \int_0^1 \Gamma(\eta)\Gamma'(\xi)M_2(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \\ &+ \lambda \left[\int_0^1 \Gamma(\eta)T_1(\eta)\mathrm{d}\eta + \int_0^1 \int_0^1 \Gamma(\eta)\Gamma'(\xi)T_2(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta - C_{T\,target} \right] . \end{aligned}$$

The shown integrals are all between 0 and 1, hence

$$\mathscr{L}(\Gamma,\lambda) = \iint_D [\Gamma(\eta)\,(M_1(\eta)+\lambda T_1(\eta)) \\ +\Gamma(\eta)\Gamma'(\xi)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta)) - \lambda C_{T\,target}\big]\,\mathrm{d}\xi\mathrm{d}\eta\ ,$$

where $D=[0,1]\times[0,1]$. Let us calculate (as done in [12]) the variation of $\mathscr{L}(\Gamma,\lambda)$ due to Γ:

$$\mathscr{L}(\alpha,\lambda) := \iint_D [(\Gamma(\eta)+\alpha\delta\Gamma(\eta))\,(M_1(\eta)+\lambda T_1(\eta)) \\ +\,(\Gamma(\eta)+\alpha\delta\Gamma(\eta))\left(\Gamma'(\xi)+\alpha\delta\Gamma'(\xi)\right)(M_2(\xi,\eta)+\lambda T_2(\xi,\eta)) \\ -\lambda C_{T\,target}\big]\,\mathrm{d}\xi\mathrm{d}\eta\ .$$

When the derivative of $\mathscr{L}(\alpha,\lambda)$ with respect to α is evaluated in zero, we find

$$\mathscr{L}'(0,\lambda) = \iint_D [\delta\Gamma(\eta)\,(M_1(\eta)+\lambda T_1(\eta)) \\ +\,\left(\Gamma'(\xi)\delta\Gamma(\eta)+\Gamma(\eta)\delta\Gamma'(\xi)\right)(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\big]\,\mathrm{d}\xi\mathrm{d}\eta\ .$$

Note that

$$\int_0^1\int_0^1 \Gamma'(\xi)\delta\Gamma(\eta)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,\mathrm{d}\xi\mathrm{d}\eta = \\ = \int_0^1 \Big\{\,[\Gamma(\xi)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,\delta\Gamma(\eta)]_0^1 \\ -\int_0^1 \Gamma(\xi)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,_{,\xi}\,\delta\Gamma(\eta)\mathrm{d}\xi\Big\}\,\mathrm{d}\eta = \\ = -\int_0^1\int_0^1 \Gamma(\xi)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,_{,\xi}\,\delta\Gamma(\eta)\mathrm{d}\xi\mathrm{d}\eta\ ,$$

while

$$\int_0^1\int_0^1 \Gamma(\eta)\delta\Gamma'(\xi)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,\mathrm{d}\xi\mathrm{d}\eta = \\ = \int_0^1 \Big\{\,[\Gamma(\eta)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,\delta\Gamma(\xi)]_0^1 \\ -\int_0^1 \Gamma(\eta)\,(M_2(\xi,\eta)+\lambda T_2(\xi,\eta))\,_{,\xi}\,\delta\Gamma(\xi)\mathrm{d}\xi\Big\}\,\mathrm{d}\eta =$$

$$= -\int_0^1 \int_0^1 \Gamma(\eta)\left(M_2(\xi,\eta) + \lambda T_2(\xi,\eta)\right)_{,\xi} \delta\Gamma(\xi)\mathrm{d}\xi\mathrm{d}\eta =$$

$$= -\int_0^1 \int_0^1 \Gamma(\xi)\left(M_2(\eta,\xi) + \lambda T_2(\eta,\xi)\right)_{,\eta} \delta\Gamma(\eta)\mathrm{d}\xi\mathrm{d}\eta \,,$$

because, by definition, $\delta\Gamma(0) = \delta\Gamma(1) = 0$. So we have

$$\mathscr{L}'(0,\lambda) = \iint_D \big[(M_1(\eta) + \lambda T_1(\eta)) - \Gamma(\xi)\left(M_2(\xi,\eta) + \lambda T_2(\xi,\eta)\right)_{,\xi} \\ -\Gamma(\xi)\left(M_2(\eta,\xi) + \lambda T_2(\eta,\xi)\right)_{,\eta}\big]\,\delta\Gamma(\eta)\mathrm{d}\xi\mathrm{d}\eta \,.$$

The constraint on the thrust gives the following equation, that is the variation of $\mathscr{L}(\Gamma,\lambda)$ due to λ,

$$\int_0^1 \int_0^1 \left[\Gamma(\eta)T_1(\eta) + \Gamma(\eta)\Gamma'(\xi)T_2(\xi,\eta) - C_{T\,target}\right]\mathrm{d}\xi\mathrm{d}\eta =$$

$$\int_0^1 \int_0^1 \left[\Gamma(\eta)T_1(\eta) - \Gamma(\eta)\Gamma(\xi)T_{2,\xi}(\xi,\eta) - C_{T\,target}\right]\mathrm{d}\xi\mathrm{d}\eta = 0 \,.$$

The system of Euler equations is

$$\begin{cases} \displaystyle\int_0^1 \big[M_1(\eta) + \lambda T_1(\eta) - \Gamma(\xi)\left(M_2(\xi,\eta) + \lambda T_2(\xi,\eta)\right)_{,\xi} \\ \qquad\qquad - \Gamma(\xi)\left(M_2(\eta,\xi) + \lambda T_2(\eta,\xi)\right)_{,\eta}\big]\mathrm{d}\xi = 0 \,, \\ \displaystyle\int_0^1 \Gamma(\eta)T_1(\eta)\mathrm{d}\eta - \int_0^1 \int_0^1 \Gamma(\eta)\Gamma(\xi)T_{2,\xi}(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta = C_{T\,target} \end{cases} \tag{26}$$

This system is valid for a general blade line, also non-planar. Note that this is a system of non-linear integral equations.

5 Numerical Method

An approximation of the solution is obtained as a linear combination of base functions. Moreover, we need to evaluate the integrals for each of the base functions. For the singular integrals we adopt a numerical method proposed in [11].

5.1 Ritz Method

The approximated solution is

$$\Gamma_B(\eta) := \sum_{k=1}^{B} b_k f_k(\eta)\,, \quad B \in \mathbb{N}\,. \tag{27}$$

The base functions $f_k(\eta)$ must be continuous, with a continuous derivative in $[0, 1]$ and with $f_k(0) = f_k(1) = 0$. The following polynomials satisfy these conditions:

$$f_k(\eta) = \left(1 - (2\eta - 1)^2\right)(2\eta - 1)^{k-1} \tag{28}$$

$$f'_k(\eta) = \left(i - 1 - (i + 1)(2\eta - 1)^2\right)(2\eta - 1)^{k-2}\,, \tag{29}$$

Figure 8 shows a representation of (28). The Lagrangian is the following function of Γ_B:

$$\mathscr{L}(\Gamma_B, \lambda) = C_M(\Gamma_B) + \lambda\left(C_T(\Gamma_B) - C_{T\,target}\right)$$

and we obtain the following algebraic system:

$$\mathscr{L}(\mathbf{b}, \lambda) = \mathbf{b}^T\mathbf{m} + \mathbf{b}^T\mathbf{M}\mathbf{b} + \lambda\left(\mathbf{b}^T\mathbf{t} + \mathbf{b}^T\mathbf{T}\mathbf{b} - c\right)\,,$$

where $\mathbf{b}$ is the vector of the coefficients of the linear combination, c is the assigned thrust coefficient $C_{T\,target}$ and the other matrices are defined as follows:

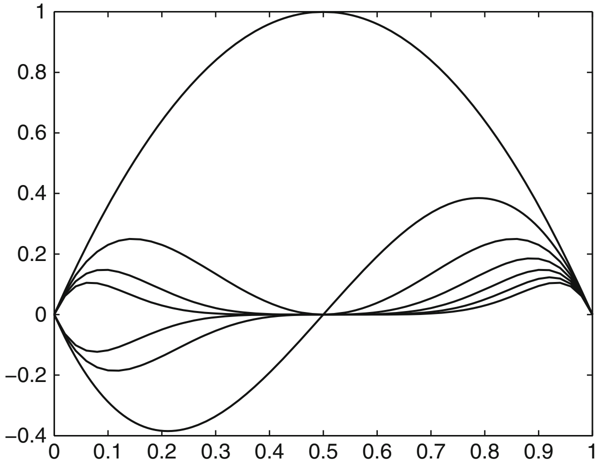


Fig. 8 Base functions for $k \in [1, 7]$

$$\begin{cases} m_k = 2N\int_0^1 f_k(\eta)\mathbf{r}(\eta)\times[(\mathbf{i}+\mu_0 r(\eta)\mathbf{e}_\theta(\eta))\times \mathrm{d}\mathbf{r}(\eta)]\cdot\mathbf{i} \\ M_{kh} = -\dfrac{1}{2\pi}\displaystyle\sum_{i=0}^{N-1}\int_0^1 f_k(\eta)\mathbf{r}(\eta)\times\left(\int_0^1 f_h'(\xi)\int_0^\infty \mathbf{g}_i(t,\xi,\eta)\mathrm{d}t\mathrm{d}\xi\times\mathrm{d}\mathbf{r}(\eta)\right)\cdot\mathbf{i} \\ t_k = -2N\int_0^1 f_k(\eta)\mu_0 r(\eta)\mathbf{e}_\theta(\eta)\times\mathrm{d}\mathbf{r}(\eta)\cdot\mathbf{i} \\ T_{kh} = \dfrac{1}{2\pi}\displaystyle\sum_{i=0}^{N-1}\int_0^1 f_k(\eta)\int_0^1 f_h'(\xi)\int_0^\infty \mathbf{g}_i(t,\xi,\eta)\mathrm{d}t\mathrm{d}\xi\times\mathrm{d}\mathbf{r}(\eta)\cdot\mathbf{i} \end{cases}.$$

In the above equations we define the *kernel*, $\mathbf{g}_i(t,\xi,\eta)$; it is the geometrical induction of the helicoidal wake vortex associated with the blade i

$$\mathbf{g}_i(t,\xi,\eta) := \frac{\mathrm{d}\mathbf{r}_i^v(t)}{dt}\times\frac{\mathbf{r}(\eta)-\mathbf{r}_i^v(t,\xi)}{\left\|\mathbf{r}(\eta)-\mathbf{r}_i^v(t,\xi)\right\|^3}. \tag{30}$$

From the null variation of $\mathscr{L}(\mathbf{b},\lambda)$, we obtain the following algebraic system

$$\begin{cases} \mathbf{m}+2\mathbf{Mb}+\lambda\,(\mathbf{t}+2\mathbf{Tb})=\mathbf{0} \\ \mathbf{b}^T\mathbf{t}+\mathbf{b}^T\mathbf{Tb}=c \end{cases}, \tag{31}$$

which is non-linear, because the first equation contains the product $\lambda\mathbf{b}$. An approximated solution of the system (31) is obtained by means of Newton's iterative method.

5.2 Evaluation of the Singular Integral

In both M_{kh} and T_{kh}, the first term of the summation (index $i=0$) is singular (see Appendix 1); in particular, the singular part are the inner two-dimensional integrals (the explicit expression of the components of vector $\mathbf{g}_0$ are given in Appendix 1):

$$\begin{aligned} I_j(\Gamma;\eta) &:= \int_0^1\int_0^\infty \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi}\left(\mathbf{g}_0(t,\xi,\eta)\right)_j \mathrm{d}t\mathrm{d}\xi \\ &= \int_0^1\int_0^\infty \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi}\left(\frac{\mathrm{d}\mathbf{r}_0^v(t)}{\mathrm{d}t}\times\frac{\mathbf{r}(\eta)-\mathbf{r}_0^v(t,\xi)}{\left\|\mathbf{r}(\eta)-\mathbf{r}_0^v(t,\xi)\right\|^3}\right)_j \mathrm{d}t\mathrm{d}\xi \\ &\text{for}\quad j=1,2,3\,. \end{aligned}$$

The other integrals are regular and we use the standard quadrature rule of Gauss-Legendre to solve them.

We split I_j into two parts:

$$I_j(\Gamma;\eta) = I_j^S(\Gamma;\eta) + I_j^R(\Gamma;\eta)$$
$$= \int_0^1 \int_0^{t^*} \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi} \left(\mathbf{g}_0(t,\xi,\eta)\right)_j \,\mathrm{d}t\mathrm{d}\xi + \int_0^1 \int_{t^*}^{\infty} \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi} \left(\mathbf{g}_0(t,\xi,\eta)\right)_j \,\mathrm{d}t\mathrm{d}\xi \ ,$$

where $I_j^S(\Gamma;\eta)$ is singular with a bounded domain, while $I_j^R(\Gamma;\eta)$ is regular with an unbounded domain. We can evaluate I_j^R as two consecutive one-dimensional integrals. The only difficulty is the unbounded domain of the inner integral. It can be estimated with the classical Gauss-Legendre quadrature rule, after the following substitution:

$$t = \frac{2t^*}{1-q}$$
$$\mathrm{d}t = \frac{2t^*}{(1-q)^2}\mathrm{d}q \ .$$

Instead, for the integral I_j^S, we need an *ad hoc* quadrature rule. The one proposed by Monegato in [11] can be applied only if the singularity is a second order pole. In Appendix 1 we prove that I_j^S respects this condition. With the substitution[4]

$$\begin{cases} t = \dfrac{t^*}{2} r\cos\theta \\ \xi = \eta + r\sin\theta \end{cases}$$

we obtain

$$I_j^S(\Gamma;\eta) = \frac{t^*}{2} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{R(\theta)} \Gamma'(r,\theta) \left(\frac{h_j(\theta)}{r^2} + \mathscr{O}\left(\frac{1}{r}\right)\right) r\mathrm{d}r\mathrm{d}\theta \ ,$$

where $h_j(\theta)$ is a regular function of θ. The above expression shows, indeed, that the integrand has exactly a second order pole in $r = 0$. Monegato proposed a product of one-dimensional quadrature rules of Gaussian type. For the inner integral, in r, we need the quadrature rule of singular integrals of the type

$$\int_a^b w(x) \frac{f(x)}{x-a} \,\mathrm{d}x \ .$$

[4] Note that the factor $\frac{t^*}{2}$ is needed just to make dimensionless the integration set on t.

This rule[5] uses nodes and weights coming from the correspondent Gauss-Jacobi rule. For us the weight function is $w(x) = 1$, so we adopt Gauss-Legendre. For the outer integral, in θ, we can use a Gauss-Legendre quadrature rule or a Gauss-Legendre-Lobatto, as suggested by Monegato for rectangular domains.

6 Straight Blade

The described procedure is applied to the case of a straight blade propeller. Figure 9 shows the Goldstein circulation and our result, in the case of a two blade propeller, versus the dimensionless coordinate along the blade $\mu = r\Omega/V_\infty$. The different curves in Fig. 9 are relevant to different values of tip speed ratio μ_0. Figure 10 represents the same comparison for different number of blades[6] N. Another validation of our method comes from the comparison with the results of momentum theory. By increasing the number of blades, the distribution obtained must approach the circulation for an infinite number of blades Γ_∞:

$$\frac{N\Gamma_\infty\Omega}{2\pi V_\infty w_0} = \frac{\mu^2}{1+\mu^2} .$$

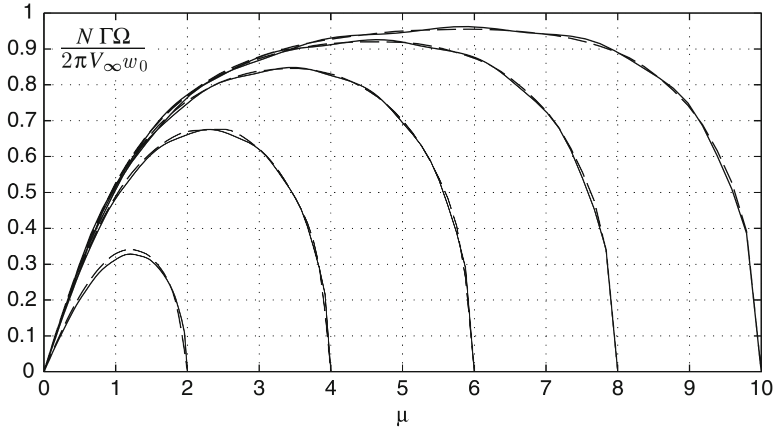


Fig. 9 Obtained (*solid line*) and Goldstein (*dashed line*) circulation distribution for a two-bladed rotor

[5] See p.768 of [11].

[6] Goldstein shows the results just for a two- and a four-bladed propeller; for higher number of blades we refer to [17].

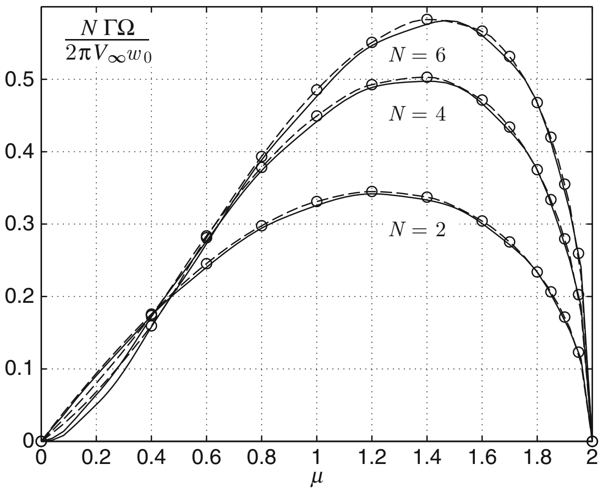


Fig. 10 Present (*solid line*) and Goldstein (*dashed line*) circulation distribution for different number of blades and $\mu_0 = 2$

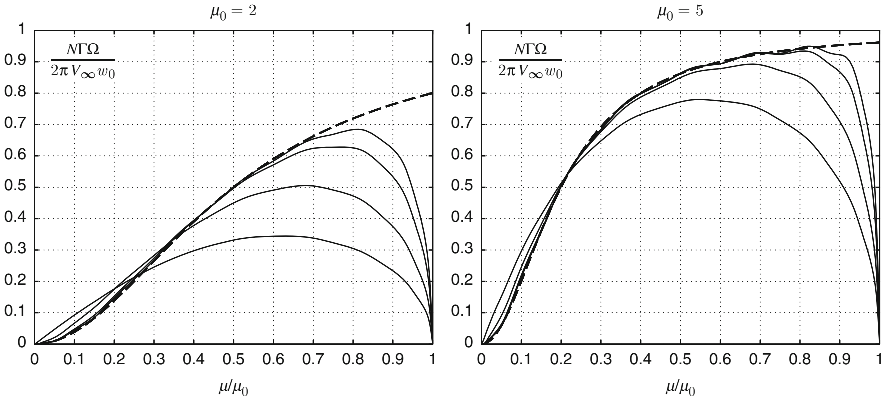


Fig. 11 Obtained circulation (*solid line*), respectively, for propeller with $N = 2,\ 4,\ 8,\ 12$ blades, and momentum theory circulation (*dashed line*). For $\mu_0 = 2, 5$ and $C_{T\,target} = 0.01$

Figure 11 shows the behaviour of our results when the number of blades increases, for two different values of tip speed ratio μ_0 and $C_{T\,target} = 0.01$. Note that for high tip speed ratio, the circulation distribution is near Γ_∞ also for a lower number of blades. This can be explained by observing the wake structure. As shown by Prandtl's tip loss factor, Γ_∞ is obtained when the distance between helicoids is null; therefore, every time we decrease this distance we get closer to Γ_∞. There are two ways to decrease this distance, keeping fixed the external radius: by increasing the number of blades or increasing the tip speed ratio.

6.1 Shifted Blade

We consider a blade with a lifting part shifted outward. The external radius of the blade is kept constant (in a dimensionless formulation is unit) while the start of the lifting point is at a distance s from the rotor axis. This is a more realistic model to represent, for example, the effect of the hub. The equation of a generic blade line shifted by s is

$$\begin{cases} x(\eta) = (1-s)f_x(\eta) \\ y(\eta) = (1-s)f_y(\eta) + s \qquad \text{with } \eta \in [0,1] \ , \\ z(\eta) = (1-s)f_z(\eta) \end{cases}$$

where $f_x(\eta)$,$f_y(\eta)$ and $f_z(\eta)$ are the functions representing the original non-shifted blade. Note that the blade line is shifted and scaled in order to keep the external radius equal to 1, hence the aspect ratio is constant. We define the *relative efficiency* ε, which is the ratio between momentum coefficient of the straight blade and of the analysed blade line.

$$\varepsilon := \frac{C_{M\,straight}}{C_M} \tag{32}$$

Therefore a blade line is more efficient than straight blade when $\varepsilon > 1$. Note that for a 60 % reduction of the blade lifting part we have a reduction in efficiency of just 30 %, see Fig. 12. Figure 13 shows circulation distribution for two different shifted blades.

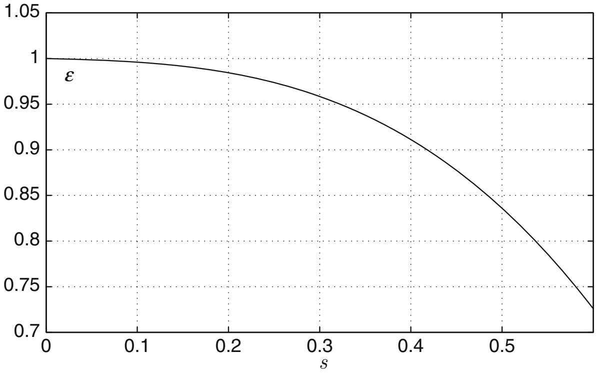


Fig. 12 Relative efficiency of shifted straight blades for $N = 2, \Omega = 2$ and $C_{T\,target} = 0.2$

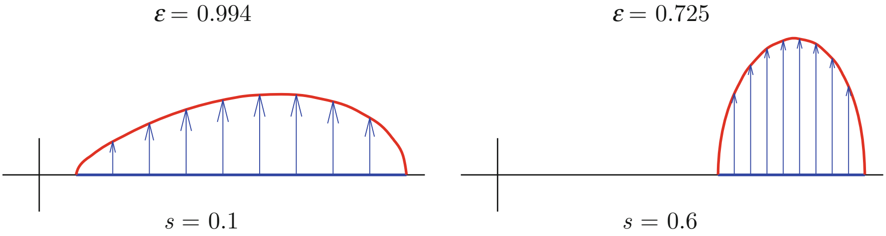
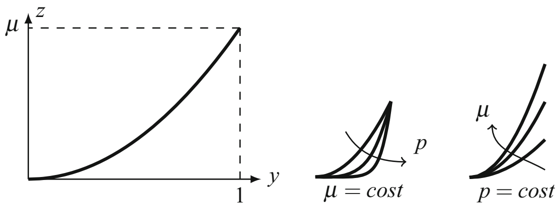


Fig. 13 Circulation distribution for shifted straight blades with $N = 2$, $\mu_0 = 2$ and $C_{T\,target} = 0.2$

Fig. 14 Swept blade



7 Swept Blade

"Swept" means with a parabolic blade line. We can sweep the blade in the plane of rotation or in axial direction; however, the meaning of the parameter defining the blade line is the same. The first kind of swept blade has the following equation

$$\begin{cases} x(\eta) = 0 \\ y(\eta) = \eta \\ z(\eta) = \mu \eta^p \end{cases} \qquad \text{with} \qquad \begin{array}{l} \eta \in [0, 1] \\ p \in \mathbb{N}\backslash\{0\} \end{array},$$

where μ is the aspect ratio and p represents the curvature (Fig. 14). We investigate the effects of these two geometrical parameters on the performance of the blade. Figure 15 shows the dependence of the relative efficiency on the aspect ratio for different values of the curvature (Fig. 16). We consider also blades swept in the axial direction. Figure 17 shows the dependence of the relative efficiency on the aspect ratio, μ, for different values of the curvature, p, while in Fig. 18 we have two examples of circulation obtained for this kind of swept blade.

8 Conclusions

A variational optimization procedure is presented and used to study non-straight blade propellers. The problem of the optimal propeller is inserted in a well-stated

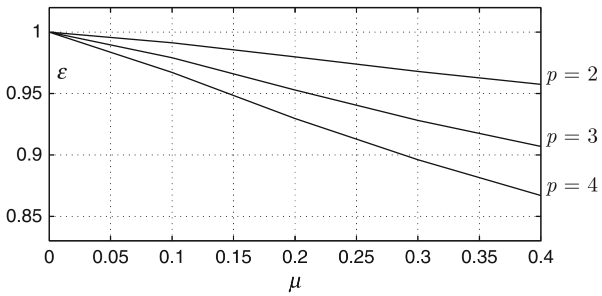


Fig. 15 Relative efficiency of blades swept in rotation plane for $N = 1$, $\Omega = 2$ and $C_{T\,target} = 0.1$

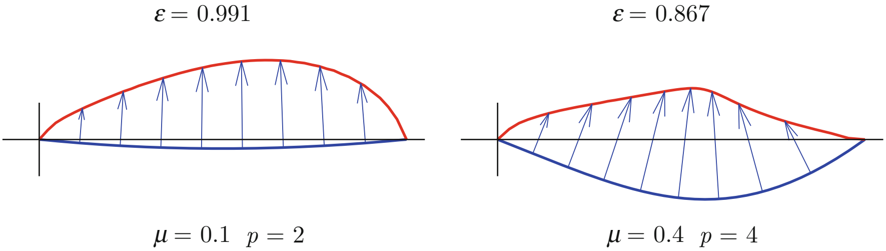


Fig. 16 Circulation distribution for blades swept in rotation plane for $N = 1, \Omega = 2$ and $C_{T\,target} = 0.1$

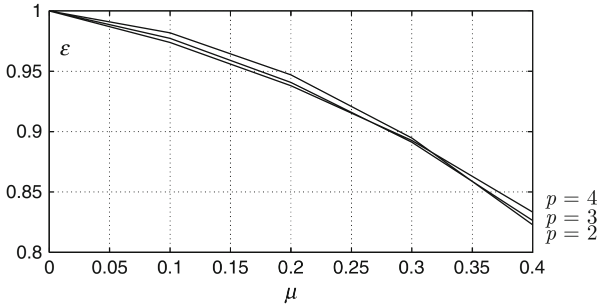


Fig. 17 Relative efficiency of blades swept in axial direction for $N = 1$, $\Omega = 2$ and $C_{T\,target} = 0.1$

mathematical environment, existence and uniqueness of the solution are proved for a generic blade line. The Euler equation, associated with this optimum problem, is found. A numerical procedure, based on a direct approach, is presented in order to determine the optimal circulation for a generic blade line.

Due to the low number of works on this argument, the method is validated just for the case of straight blade. However the method is suitable also for non-straight blades, and the example of the swept blades is shown. Finally we investigate the effect of the velocity induced by the bounded vorticity.

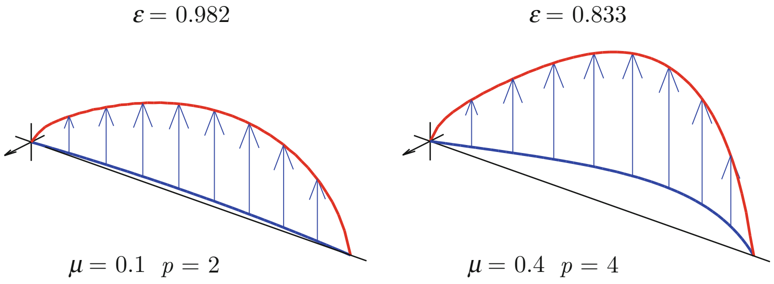


Fig. 18 Circulation distribution for blades swept in axial direction for $N = 1, \Omega = 2$ and $C_{T\,target} = 0.1$

The present paper lays on the theoretical foundation of the new technique. Applications to specific case and validation for non-classical case will be presented in subsequent papers.

Appendix 1: Singular Part of the Kernel

To obtain the singular part of the kernel is important for two reasons. First, the complete expression of the kernel is too complex to allow an analysis of existence, we study just the singular part (see Appendix 2). Furthermore, we have to demonstrate that the singularity is a second order pole in order to use the Monegato's quadrature rule (see Sect. 5). The expression of the kernel, as defined in Sect. 5.1 by Eq. (30), is

$$\mathbf{g}_i(t,\xi,\eta) := \frac{\mathrm{d}\mathbf{r}_i^v(t)}{\mathrm{d}t} \times \frac{\mathbf{r}(\eta) - \mathbf{r}_i^v(t,\xi)}{\left\| \mathbf{r}(\eta) - \mathbf{r}_i^v(t,\xi) \right\|^3} ,$$

where i is the origin blade of the helicoidal wake. For $i = 0$ the blade is the same where we calculate the induced velocity. Therefore we have a singularity in $(t,\xi) = (0,\eta)$, because

$$\lim_{(t,\xi)\to(0,\eta)} \mathbf{r}_0^v(t,\xi) = \mathbf{r}(\eta) .$$

We take into account just the kernel $\mathbf{g}_0(t,\xi,\eta)$ (from now on the subscript $_0$ is dropped for simplicity), its components are

$$\begin{cases} g_x(t,\xi,\eta) = \dfrac{(z(\eta)-z^v(t,\xi))\,y_{,t}^{v}\,(t,\xi)-(y(\eta)-y^v(t,\xi))\,z_{,t}^{v}\,(t,\xi)}{\left|(x(\eta)-x^v(t,\xi))^2+(y(\eta)-y^v(t,\xi))^2+(z(\eta)-z^v(t,\xi))^2\right|^{3/2}} \\ g_y(t,\xi,\eta) = \dfrac{(x(\eta)-x^v(t,\xi))\,z_{,t}^{v}\,(t,\xi)-(z(\eta)-z^v(t,\xi))\,x_{,t}^{v}\,(t,\xi)}{\left|(x(\eta)-x^v(t,\xi))^2+(y(\eta)-y^v(t,\xi))^2+(z(\eta)-z^v(t,\xi))^2\right|^{3/2}} \\ g_z(t,\xi,\eta) = \dfrac{(y(\eta)-y^v(t,\xi))\,x_{,t}^{v}\,(t,\xi)-(x(\eta)-x^v(t,\xi))\,y_{,t}^{v}\,(t,\xi)}{\left|(x(\eta)-x^v(t,\xi))^2+(y(\eta)-y^v(t,\xi))^2+(z(\eta)-z^v(t,\xi))^2\right|^{3/2}} \end{cases}.$$

We analyse just the z component, for the other is nearly the same. We need the following Taylor expansions around the singularity.

$$\begin{cases} x^v(t,\xi) = x(\eta) + t\,x_{,t}^{v}\,\big|_{(0,\eta)} + (\xi-\eta)\,x_{,\xi}^{v}\Big|_{(0,\eta)} + \mathcal{O}(t^2+(\xi-\eta)^2) \\ y^v(t,\xi) = y(\eta) + t\,y_{,t}^{v}\,\big|_{(0,\eta)} + (\xi-\eta)\,y_{,\xi}^{v}\Big|_{(0,\eta)} + \mathcal{O}(t^2+(\xi-\eta)^2) \end{cases}$$

$$\begin{cases} x_{,t}^{v}\,(t,\xi) = x_{,t}^{v}\,\big|_{(0,\eta)} + t\,x_{,tt}^{v}\,\big|_{(0,\eta)} + (\xi-\eta)\,x_{,\xi t}^{v}\Big|_{(0,\eta)} + \mathcal{O}(t^2+(\xi-\eta)^2) \\ y_{,t}^{v}\,(t,\xi) = y_{,t}^{v}\,\big|_{(0,\eta)} + t\,y_{,tt}^{v}\,\big|_{(0,\eta)} + (\xi-\eta)\,y_{,\xi t}^{v}\Big|_{(0,\eta)} + \mathcal{O}(t^2+(\xi-\eta)^2) \end{cases}$$

We move to polar coordinates around the singularity[7]

$$\begin{cases} t = r\cos\theta \\ \xi-\eta = r\sin\theta \end{cases}$$

and the Taylor expansions become

$$\begin{cases} x^v(r,\theta) = x(\eta) + x_{,t}^{v}\,\big|_{(0,\eta)}\,r\cos\theta + x_{,\xi}^{v}\Big|_{(0,\eta)}\,r\sin\theta + \mathcal{O}(r^2) \\ y^v(r,\theta) = y(\eta) + y_{,t}^{v}\,\big|_{(0,\eta)}\,r\cos\theta + y_{,\xi}^{v}\Big|_{(0,\eta)}\,r\sin\theta + \mathcal{O}(r^2) \end{cases},$$

$$\begin{cases} x_{,t}^{v}\,(r,\theta) = x_{,t}^{v}\,\big|_{(0,\eta)} + x_{,tt}^{v}\,\big|_{(0,\eta)}\,r\cos\theta + x_{,\xi t}^{v}\Big|_{(0,\eta)}\,r\cos\theta + \mathcal{O}(r^2) \\ y_{,t}^{v}\,(r,\theta) = y_{,t}^{v}\,\big|_{(0,\eta)} + y_{,tt}^{v}\,\big|_{(0,\eta)}\,r\cos\theta + y_{,\xi t}^{v}\Big|_{(0,\eta)}\,r\sin\theta + \mathcal{O}(r^2) \end{cases}.$$

[7] As done in [18].

In order to clarify the calculation we define the functions at the numerator and at the denominator of g_z:

$$g_z(r,\theta,\eta) = \frac{\mathfrak{N}(t,\xi,\eta)}{\mathfrak{D}(t,\xi,\eta)} \ ,$$

where

$$\begin{aligned}
\mathfrak{N}(r,\theta,\eta) &= (y(\eta) - y^v(t,\xi))\, x^v_{,t}\,(t,\xi) - (x(\eta) - x^v(t,\xi))\, y^v_{,t}\,(t,\xi), \\
\mathfrak{D}(r,\theta,\eta) &= \left|(x(\eta) - x^v(t,\xi))^2 + (y(\eta) - y^v(t,\xi))^2 + (z(\eta) - z^v(t,\xi))^2\right|^{3/2} \ .
\end{aligned}$$

For the terms of $\mathfrak{N}$ we have

$$\begin{aligned}
&(x(\eta) - x^v(t,\xi))\, y^v_{,t}\,(t,\xi) = \\
&= \left(- x^v_{,t}\big|_{(0,\eta)}\, r\cos\theta - x^v_{,\xi}\Big|_{(0,\eta)}\, r\sin\theta + \mathcal{O}(r^2)\right) \\
&\left(y^v_{,t}\big|_{(0,\eta)} + y^v_{,tt}\big|_{(0,\eta)}\, r\cos\theta + y^v_{,\xi t}\Big|_{(0,\eta)}\, r\sin\theta + \mathcal{O}(r^2)\right) = \\
&= -\left(x^v_{,t}\big|_{(0,\eta)}\, r\cos\theta + x^v_{,\xi}\Big|_{(0,\eta)}\, r\sin\theta\right) y^v_{,t}\big|_{(0,\eta)} + \mathcal{O}(r^2)
\end{aligned}$$

and in the same way:

$$\begin{aligned}
&(y(\eta) - y^v(t,\xi))\, x^v_{,t}\,(t,\xi) = \\
&= -\left(y^v_{,t}\big|_{(0,\eta)}\, r\cos\theta + y^v_{,\xi}\Big|_{(0,\eta)}\, r\sin\theta\right) x^v_{,t}\big|_{(0,\eta)} + \mathcal{O}(r^2) \ .
\end{aligned}$$

Hence the expression of $\mathfrak{N}$:

$$\begin{aligned}
\mathfrak{N}(r,\theta,\eta) &= \left(x^v_{,t}\big|_{(0,\eta)}\, r\cos\theta + x^v_{,\xi}\Big|_{(0,\eta)}\, r\sin\theta\right) y^v_{,t}\big|_{(0,\eta)} \\
&\quad - \left(y^v_{,t}\big|_{(0,\eta)}\, r\cos\theta + y^v_{,\xi}\Big|_{(0,\eta)}\, r\sin\theta\right) x^v_{,t}\big|_{(0,\eta)} + \mathcal{O}(r^2) = \\
&= \left(x^v_{,t}\big|_{(0,\eta)}\, y^v_{,t}\big|_{(0,\eta)} - y^v_{,t}\big|_{(0,\eta)}\, x^v_{,t}\big|_{(0,\eta)}\right) r\sin\theta + \mathcal{O}(r^2) = \\
&= N r\sin\theta + \mathcal{O}(r^2) \ .
\end{aligned}$$

where

$$N = x^v_{,t}\big|_{(0,\eta)}\, y^v_{,t}\big|_{(0,\eta)} - y^v_{,t}\big|_{(0,\eta)}\, x^v_{,t}\big|_{(0,\eta)} \ .$$

Instead for the terms of $\mathfrak{D}$:

$$
\begin{aligned}
&(x(\eta) - x^v(t,\xi))^2 = \\
&= \left(- x,_t^v \big|_{(0,\eta)} r\cos\theta - x,_\xi^v \Big|_{(0,\eta)} r\sin\theta + \mathscr{O}(r^2) \right)^2 = \\
&= \left(x,_t^v \big|_{(0,\eta)} r\cos\theta + x,_\xi^v \Big|_{(0,\eta)} r\sin\theta \right)^2 + \mathscr{O}(r^3)
\end{aligned}
$$

and nearly the same for the other components. The expression of $\mathfrak{D}$ becomes

$$
\begin{aligned}
\mathfrak{D}(r,\theta,\eta) = \Bigg| &\left(x,_t^v \big|_{(0,\eta)} r\cos\theta + x,_\xi^v \Big|_{(0,\eta)} r\sin\theta \right)^2 \\
&+ \left(y,_t^v \big|_{(0,\eta)} r\cos\theta + y,_\xi^v \Big|_{(0,\eta)} r\sin\theta \right)^2 \\
&+ \left(z,_t^v \big|_{(0,\eta)} r\cos\theta + z,_\xi^v \Big|_{(0,\eta)} r\sin\theta \right)^2 + \mathscr{O}(r^3) \Bigg|^{3/2} = \\
&= r^2 \left| D(\theta) + \mathscr{O}(r) \right|^{3/2} = r^2 D^{\frac{3}{2}}(\theta) \left(1 + \mathscr{O}(r)\right)^{3/2} ,
\end{aligned}
$$

where

$$
\begin{aligned}
D(\theta) = &\left(x,_t^v \big|_{(0,\eta)} \cos\theta + x,_\xi^v \Big|_{(0,\eta)} \sin\theta \right)^2 \\
&+ \left(y,_t^v \big|_{(0,\eta)} \cos\theta + y,_\xi^v \Big|_{(0,\eta)} \sin\theta \right)^2 \\
&+ \left(z,_t^v \big|_{(0,\eta)} \cos\theta + z,_\xi^v \Big|_{(0,\eta)} \sin\theta \right)^2 .
\end{aligned} \tag{33}
$$

Therefore g_z becomes

$$
\begin{aligned}
g_z(r,\theta,\eta) &= \frac{Nr\cos\theta + \mathscr{O}(r^2)}{r^2 D^{\frac{3}{2}}(\theta) \left(1 + \mathscr{O}(r)\right)^{3/2}} \\
&= \frac{\left(N\sin\theta + \mathscr{O}(r)\right)\left(1 + \mathscr{O}(r)\right)^{-3/2}}{r^2 D^{\frac{3}{2}}(\theta)} \\
&= \frac{\left(N\sin\theta + \mathscr{O}(r)\right)\left(1 - \frac{3}{2}\mathscr{O}(r) + \mathscr{O}(r^2)\right)}{r^2 D^{\frac{3}{2}}(\theta)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{N\sin\theta + \mathscr{O}(r)}{r^2 D(\theta)} \\
&= \frac{N\sin\theta}{r^2 D^{\frac{3}{2}}(\theta)} + \mathscr{O}\left(\frac{1}{r}\right) .
\end{aligned}
$$

Finally we obtain

$$
\begin{cases}
g_x(r,\theta,\eta) = \dfrac{1}{r^2}\,\dfrac{\left(y,_{\xi}^{v}\, z,_{t}^{v} - z,_{\xi}^{v}\, y,_{t}^{v}\right)\sin\theta}{[D(\theta)]^{3/2}} + \mathscr{O}\left(\dfrac{1}{r}\right) \\[2ex]
g_y(r,\theta,\eta) = \dfrac{1}{r^2}\,\dfrac{\left(z,_{\xi}^{v}\, x,_{t}^{v} - x,_{\xi}^{v}\, z,_{t}^{v}\right)\sin\theta}{[D(\theta)]^{3/2}} + \mathscr{O}\left(\dfrac{1}{r}\right) \\[2ex]
g_z(r,\theta,\eta) = \dfrac{1}{r^2}\,\dfrac{\left(x,_{\xi}^{v}\, y,_{t}^{v} - y,_{\xi}^{v}\, x,_{t}^{v}\right)\sin\theta}{[D(\theta)]^{3/2}} + \mathscr{O}\left(\dfrac{1}{r}\right)
\end{cases} , \tag{34}
$$

where all the partial derivatives are calculated in $(0, \eta)$, and $D(\theta)$ is defined by (33). These expressions proved that the singularity is a second order pole. The partial derivative in (34), according to Eq. (9) are:

$$
\begin{cases}
x,_{t}^{v} = 1 \\
y,_{t}^{v} = -\mu_0 z(\eta) \\
z,_{t}^{v} = \mu_0 y(\eta)
\end{cases}
\qquad
\begin{cases}
x,_{\xi}^{v} = x_{,\eta}(\eta) \\
y,_{\xi}^{v} = y_{,\eta}(\eta) \\
z,_{\xi}^{v} = z_{,\eta}(\eta)
\end{cases} ,
$$

thus Eq. (34) becomes

$$
\begin{cases}
g_x(t,\xi,\eta) = \mu_0 \dfrac{yy,_{\eta} + zz,_{\eta}}{\left[\overline{D}(t,\xi,\eta)\right]^{3/2}}(\xi-\eta) + \mathscr{O}\left(\dfrac{1}{\sqrt{t^2 + (\xi-\eta)^2}}\right) \\[2ex]
g_y(t,\xi,\eta) = \dfrac{z,_{\eta} - \mu_0 yx,_{\eta}}{\left[\overline{D}(t,\xi,\eta)\right]^{3/2}}(\xi-\eta) + \mathscr{O}\left(\dfrac{1}{\sqrt{t^2 + (\xi-\eta)^2}}\right) \\[2ex]
g_z(t,\xi,\eta) = -\dfrac{y,_{\eta} + \mu_0 zx,_{\eta}}{\left[\overline{D}(t,\xi,\eta)\right]^{3/2}}(\xi-\eta) + \mathscr{O}\left(\dfrac{1}{\sqrt{t^2 + (\xi-\eta)^2}}\right)
\end{cases} , \tag{35}
$$

where

$$
\begin{aligned}
\overline{D}(t,\xi,\eta) = {} & \left[1 + \mu_0^2(y^2 + z^2)\right]t^2 + 2\left[x,_{\eta} + \mu_0(yz,_{\eta} - zy,_{\eta})\right](\xi-\eta)t + \\
& + \left(x,_{\eta}^2 + y,_{\eta}^2 + z,_{\eta}^2\right)(\xi-\eta)^2 .
\end{aligned} \tag{36}
$$

Appendix 2: Existence of Functionals

In order to prove existence and continuity of the functionals we have to prove that the integrals, defining C_T and C_M, exist. Just the term containing induced velocity is singular; this is obtained as a summation over the blades and only the first term of this sum is singular, the one indicated by the apex ^{0}. Therefore we analyse the following integrals:

$$C^0_{T_{ind}}(\Gamma) = \frac{1}{2\pi}\int_0^1 \Gamma(\eta)\int_0^1 \frac{\mathrm{d}\Gamma}{\mathrm{d}\xi}\int_0^\infty \mathbf{g}_0(t,\xi,\eta)\mathrm{d}t\mathrm{d}\xi \times \mathrm{d}\mathbf{r}(\eta)\cdot\mathbf{i} \tag{37}$$

$$C^0_{M_{ind}}(\Gamma) = -\frac{1}{2\pi}\int_0^1 \Gamma(\eta)\mathbf{r}(\eta)\times\left(\int_0^1 \frac{\mathrm{d}\Gamma}{\mathrm{d}\xi}\int_0^\infty \mathbf{g}_0(t,\xi,\eta)\mathrm{d}t\mathrm{d}\xi \times \mathrm{d}\mathbf{r}(\eta)\right)\cdot\mathbf{i}\,, \tag{38}$$

where $\mathbf{g}_0(t,\xi,\eta)$ is the singular kernel[8]:

$$\mathbf{g}_0(t,\xi,\eta) := \frac{\mathrm{d}\mathbf{r}^v_0(t)}{\mathrm{d}t}\times\frac{\mathbf{r}(\eta)-\mathbf{r}^v_0(t,\xi)}{\left\|\mathbf{r}(\eta)-\mathbf{r}^v_0(t,\xi)\right\|^3}\,.$$

Performing the operations between vectors we obtain the same kind of functional for both, momentum and thrust:

$$C(\Gamma) := \int_0^1 \Gamma(\eta)\int_0^1 \Gamma'(\xi)\int_0^\infty \sum_{j=1}^3 f_j(\eta)\left(\mathbf{g}_0(t,\xi,\eta)\right)_j \mathrm{d}t\mathrm{d}\xi\mathrm{d}\eta\,,$$

where $f_j(\eta)$ is a regular function depending on the blade line. We employ Eq. (35) and then we neglect the regular part of $C(\Gamma)$ in order to obtain the following singular integral

$$\overline{C}(\Gamma) := \int_0^1 \Gamma(\eta)\int_0^1 \Gamma'(\xi)\int_0^\infty \frac{f(\eta)(\xi-\eta)}{\left[a(\eta)t^2+2b(\xi,\eta)t+c(\xi,\eta)\right]^{\frac{3}{2}}}\mathrm{d}t\mathrm{d}\xi\mathrm{d}\eta\,, \tag{39}$$

where

$$\begin{aligned} a(\eta) &= 1+\mu_0^2(y^2+z^2) \\ b(\xi,\eta) &= \left[x,_\eta + \mu_0(yz,_\eta - zy,_\eta)\right](\xi-\eta) \\ c(\xi,\eta) &= \left(x,_\eta^2 + y,_\eta^2 + z,_\eta^2\right)(\xi-\eta)^2 \end{aligned} \tag{40}$$

[8] The presence of an indefinite integral is not a problem. Note that for $t \to \infty$ we have $x^v \to \infty$, while y^v and z^v are limited. This means for $t \to \infty$ we have $g_x(t,\xi,\eta) \to 0$ as $1/(x^v)^3$, while $g_y(t,\xi,\eta)$ and $g_z(t,\xi,\eta) \to 0$ as $1/(x^v)^2$.

and $f(\eta)$ is a regular function depending on the blade line. This expression does not respect the definition of integral according to Riemann.[9] We have to adopt *Cauchy principal value*. That means associating with the integral the value of a correspondent limit, if this limit exists finite then the integral converges according to Cauchy. Before proving the existence of this integral, we recall some useful definitions.

Definition 1. A function $f : [a, b] \to \mathbb{R}$ is absolutely continuous in $[a, b]$, and we write $f \in AC[a, b]$ iff, for any $\varepsilon > 0$ it exists $\delta > 0$ such that for any finite collections of disjoint intervals $]\alpha_i, \beta_i[,\ i = 1, \ldots, k$, included in $[a, b]$ and with

$$\sum_{i=1}^{k}(\beta_i - \alpha_i) < \delta\ , \qquad \text{it results} \qquad \sum_{i=1}^{k}|f(\beta_i) - f(\alpha_i)| < \varepsilon\ .$$

Definition 2. Let $(Y, \mathscr{F}, \mu)$ be a measure space and $1 \le p \le \infty$. We put

$$L^p(Y) = \left\{f : Y \to \overline{\mathbb{R}} : f \text{ is measurable and } \int_Y |f|^p\, dy < \infty\right\}\ ,$$

$$\|f\|_{L^p(Y)} = \left[\int_Y |f|^p\, \mathrm{d}y\right]^{\frac{1}{p}}\ .$$

Hölder Inequality. *If q is conjugate exponent of p* (i.e. $1/p + 1/q = 1$, *by stipulation, the conjugate exponent of* 1 *is* ∞*), if $f \in L^p(Y)$ and $g \in L^q(Y)$, then*

$$fg \in L^1(Y) \quad \text{and} \quad \|fg\|_{L^1(Y)} \le \|f\|_{L^p(Y)}\|g\|_{L^q(Y)}\ .$$

Proposition 1. *Let $\Gamma \in AC[0; 1]$ be such that*

$$\Gamma(0) = \Gamma(1) = 0, \qquad \Gamma,\ \Gamma' \in L^{1+\varepsilon}(0; 1) \text{ with } \varepsilon > 0\ .$$

Then $\overline{C}(\Gamma)$, defined in (39)*, is convergent as a Cauchy improper integral.*

Proof. Let us set

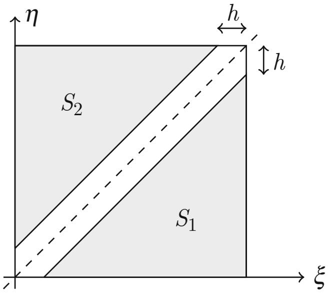
$$S_1(h) := \left\{(\xi, \eta) \in \mathbb{R}^2 : 0 \le \xi \le 1 - h,\ \xi + h \le \eta \le 1\right\}$$
$$S_2(h) := \left\{(\xi, \eta) \in \mathbb{R}^2 : h \le \xi \le 1,\ 0 \le \eta \le \xi - h\right\}\ ,$$

with $h \in [0; 1]$ (Fig. 19), and

$$G_{S_i(h)} := \iint_{S_i(h)} \Gamma'(\xi)\Gamma(\eta) \int_0^\infty \frac{f(\eta)(\xi - \eta)}{[a(\eta)t^2 + 2b(\xi, \eta)t + c(\xi, \eta)]^{\frac{3}{2}}}\mathrm{d}t\mathrm{d}\xi\mathrm{d}\eta$$

[9] Because for $(t, \xi) = (0, \eta)$ the denominator is null. The square root gives no problems of singularity, because its argument is a distance, that is always non-negative.

Fig. 19 Regular integration sets



for $i = 1, 2$. The standard integration rules can be adopted for these integrals because they are regular. We solve the inner integral, in t:

$$T(\xi,\eta) := \int_0^\infty \frac{\mathrm{d}t}{(at^2+2bt+c)^{\frac{3}{2}}} = \int_0^\infty \frac{\mathrm{d}t}{\left[\left(\sqrt{a}t + b/\sqrt{a}\right)^2 + c - b^2/a\right]^{3/2}} ,$$

with the change of variable

$$s = \sqrt{a}t + b/\sqrt{a} \quad \Rightarrow \quad \mathrm{d}t = \mathrm{d}s/\sqrt{a}$$
$$t \to \infty \quad , \quad s \to \infty$$
$$t = 0 \, , \, s = b/\sqrt{a} \, ,$$

we have[10]

$$T(\xi,\eta) = \frac{1}{\sqrt{a}} \int_{\frac{b}{\sqrt{a}}}^\infty \frac{\mathrm{d}s}{[s^2+c-b^2/a]^{3/2}} = \frac{1}{\sqrt{a}} \left[\frac{s}{(c-b^2/a)\sqrt{c-b^2/a+s^2}} \right]_{\frac{b}{\sqrt{a}}}^\infty =$$
$$= \frac{1}{\sqrt{c}\left(\sqrt{ac}+b\right)} = \frac{g(\eta)}{(\xi-\eta)^2} ,$$

where, in the last passage, we used Eq. (40) and $g(\eta)$ is a regular function depending on the blade line. We obtain

$$G_{S_i(h)} = \iint_{S_i(h)} \frac{\Gamma'(\xi)\Gamma(\eta)\bar{f}(\eta)}{\eta-\xi} \mathrm{d}\xi \mathrm{d}\eta \, , \quad i = 1, 2 \, ,$$

[10] We use the standard integration rule $\int \frac{\mathrm{d}t}{(a^2+t^2)^{3/2}} = \frac{t}{a^2\sqrt{a^2+t^2}}$.

where $\overline{f}(\eta) = -f(\eta)g(\eta)$, that is a regular function depending on blade line.[11] Let us integrate by parts both $G_{S_1(h)}$ and $G_{S_2(h)}$. For the first integral we have

$$
\begin{aligned}
G_{S_1(h)} &= \int_0^{1-h}\int_{\xi+h}^{1} \Gamma'(\xi)\Gamma(\eta)\overline{f}(\eta)\frac{1}{\eta-\xi}\mathrm{d}\eta\mathrm{d}\xi \\
&= \int_0^{1-h}\Bigg\{\left[\Gamma'(\xi)\Gamma(\eta)\overline{f}(\eta)\ln(\eta-\xi)\right]_{\xi+h}^{1} \\
&\quad -\int_{\xi+h}^{1}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\eta-\xi)\mathrm{d}\eta\Bigg\}\,\mathrm{d}\xi\ ,
\end{aligned}
$$

because $\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\ln(\eta-\xi)\right] = \frac{1}{\eta-\xi}$. Then, by using the boundary condition $\Gamma(1) = 0$,

$$
\begin{aligned}
G_{S_1(h)} &= \int_0^{1-h}\Bigg\{-\Gamma'(\xi)\Gamma(\xi+h)\overline{f}(\xi+h)\ln h \\
&\quad -\int_{\xi+h}^{1}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\eta-\xi)\mathrm{d}\eta\Bigg\}\,\mathrm{d}\xi\ .
\end{aligned}
$$

We proceed in the same way for the second integral:

$$
\begin{aligned}
G_{S_2(h)} &= \int_h^{1}\int_0^{\xi-h} \Gamma'(\xi)\Gamma(\eta)\overline{f}(\eta)\frac{1}{\eta-\xi}\mathrm{d}\eta\mathrm{d}\xi \\
&= \int_h^{1}\Bigg\{\left[\Gamma'(\xi)\Gamma(\eta)\overline{f}(\eta)\ln(\xi-\eta)\right]_0^{\xi-h} \\
&\quad -\int_0^{\xi-h}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\eta-\xi)\mathrm{d}\eta\Bigg\}\,\mathrm{d}\xi\ ,
\end{aligned}
$$

because $\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\ln(\xi-\eta)\right] = -\frac{1}{\xi-\eta} = \frac{1}{\eta-\xi}$. Then, by using the other boundary condition $\Gamma(0) = 0$,

$$
\begin{aligned}
G_{S_2(h)} &= \int_h^{1}\Bigg\{\Gamma'(\xi)\Gamma(\xi-h)\overline{f}(\xi-h)\ln h \\
&\quad -\int_0^{\xi-h}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\xi-\eta)\mathrm{d}\eta\Bigg\}\,\mathrm{d}\xi\ .
\end{aligned}
$$

[11] From now on the procedure is very similar to the one used in [12] Appendix 1.

It results that

$$\begin{aligned}\overline{C}(\Gamma) &= \lim_{h\to 0}\left[G_{S_1(h)} + G_{S_2(h)}\right]\\ &= \lim_{h\to 0}\Bigg[-\int_0^{1-h}\int_{\xi+h}^{1}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\eta-\xi)\mathrm{d}\eta\mathrm{d}\xi\\ &+ \ln h\left(\int_h^1 \Gamma'(\xi)\Gamma(\xi-h)\overline{f}(\xi-h)\mathrm{d}\xi - \int_0^{1-h}\Gamma'(\xi)\Gamma(\xi+h)\overline{f}(\xi+h)\mathrm{d}\xi\right)\\ &-\int_h^1\int_0^{\xi-h}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\xi-\eta)\mathrm{d}\eta\mathrm{d}\xi\Bigg]\ .\end{aligned}$$

where

$$\lim_{h\to 0}\ \ln h\left(\int_h^1 \Gamma'(\xi)\Gamma(\xi-h)\overline{f}(\xi-h)\mathrm{d}\xi \right.\\ \left. -\int_0^{1-h}\Gamma'(\xi)\Gamma(\xi+h)\overline{f}(\xi+h)\mathrm{d}\xi\right) = 0\ ,$$

thus

$$\begin{aligned}\overline{C}(\Gamma) = \lim_{h\to 0}\Bigg[&-\int_0^{1-h}\int_{\xi+h}^{1}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\eta-\xi)\mathrm{d}\eta\mathrm{d}\xi\\ &-\int_h^1\int_0^{\xi-h}\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln(\xi-\eta)\mathrm{d}\eta\mathrm{d}\xi\Bigg] =\\ &= -\int_0^1\int_0^1\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln|\xi-\eta|\,\mathrm{d}\eta\mathrm{d}\xi\ .\end{aligned}$$

The absolute value of sum is less than or equal to the sum of absolute values:

$$\left|\int_0^1\int_0^1\Gamma'(\xi)\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\ln|\xi-\eta|\,\mathrm{d}\eta\mathrm{d}\xi\right| \le \\ \int_0^1\left|\Gamma'(\xi)\right|\int_0^1\left|\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\right|\left|\ln|\xi-\eta|\right|\mathrm{d}\eta\mathrm{d}\xi. \quad (41)$$

From Hölder inequality we obtain

$$\int_0^1\left|\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\right|\left|\ln|\xi-\eta|\right|\mathrm{d}\eta\mathrm{d}\xi \le \\ \left\|\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\Gamma(\eta)\overline{f}(\eta)\right]\right\|_{L^{1+\varepsilon}(0,1)}\left\|\ln|\xi-\eta|\right\|_{L^{\frac{1+\varepsilon}{\varepsilon}}(0,1)}\ .$$

Observe that[12]

$$\|\ln|\xi-\eta|\|_{L^{\frac{1+\varepsilon}{\varepsilon}}(0,1)} < c \quad \text{with} c \in \mathbb{R}$$

and

$$\left\| \frac{\mathrm{d}}{\mathrm{d}\eta} \left[\Gamma(\eta) \overline{f}(\eta) \right] \right\|_{L^{1+\varepsilon}(0,1)} \leq \left\| \Gamma'(\eta) \overline{f}(\eta) \right\|_{L^{1+\varepsilon}(0,1)} + \left\| \Gamma(\eta) \overline{f}'(\eta) \right\|_{L^{1+\varepsilon}(0,1)} \leq$$
$$\leq P \left\| \Gamma'(\eta) \right\|_{L^{1+\varepsilon}(0,1)} + Q \left\| \Gamma(\eta) \right\|_{L^{1+\varepsilon}(0,1)} ,$$

because, given the regularity of $\overline{f}(\eta)$, we have that

$$\exists P, Q \in \mathbb{R} : \quad \left| \overline{f}(\eta) \right| \leq P, \ \left| \overline{f}'(\eta) \right| \leq Q \ \forall \eta \in [0; 1] \ .$$

From (41) we have, finally:

$$\left| \overline{C}(\Gamma) \right| \leq cP \left\| \Gamma(\eta) \right\|_{L^1(0,1)} \left\| \Gamma'(\eta) \right\|_{L^{1+\varepsilon}(0,1)} +$$
$$+ cQ \left\| \Gamma(\eta) \right\|_{L^1(0,1)} \left\| \Gamma(\eta) \right\|_{L^{1+\varepsilon}(0,1)} < \infty \quad (42)$$

as required. □

Appendix 3: Convexity of Momentum Functional

According to (15), (19) and (30) we define $C_{M_1}(\Gamma)$ and $C_{M_2}(\Gamma)$ such that

$$C_{M_1}(\Gamma) := 2N \int_{\gamma_b^0} \Gamma(\mathbf{r}) \mathbf{r} \times [(\mathbf{i} + \mu_0 r \mathbf{e}_\theta) \times \mathrm{d}\mathbf{r}] \cdot \mathbf{i} \ ,$$

$$C_{M_2}(\Gamma) := -\frac{N}{2\pi} \sum_{i=0}^{N-1} \int_{\gamma_b^0} \Gamma(\mathbf{r}) \mathbf{r} \times \left(\int_{\gamma_b^i} \frac{\mathrm{d}\Gamma(\xi)}{\mathrm{d}\xi} \int_{\gamma_v^i} \mathbf{g}_i(t, \xi, \eta) \mathrm{d}t \mathrm{d}\xi \times \mathrm{d}\mathbf{r} \right) \cdot \mathbf{i} \ ,$$

that means

$$C_M(\Gamma) = C_{M_1}(\Gamma) + C_{M_2}(\Gamma) \ . \quad (43)$$

Before proving the convexity of the functional C_M, we recall the following:

[12] See Appendix 1 of [12].

Definition 3. Let K be a vector space. A function $f : K \to \mathbb{R}$ is called convex, if and only if

$$(1-\alpha)f(x) + \alpha f(x) \geq f((1-\alpha)x + \alpha y), \quad \forall x, y \in K, \quad \forall \alpha \in [0,1]\,. \tag{44}$$

We say that function f is strictly convex, if and only if the inequality (44) holds strictly.

Theorem 1. *Let K be a vector space and $f : K \to \mathbb{R}$ be a function whatever. Then f is strictly convex on K, if and only if $\forall x, y \in K$ the quotient ratio*

$$t \to R_y(t) = \frac{f(x+ty) - f(x)}{t}, \quad t \in \mathbb{R}_+ \setminus \{0\}$$

is an increasing function.

Proposition 2. *Let be:*

$$I := \left\{\Gamma \in AC[0;1],\ \Gamma,\ \Gamma' \in L^{1+\varepsilon}(0;1) \text{ with } \varepsilon > 0,\ \Gamma(0) = \Gamma(1) = 0\right\}$$

and $C_M(\Gamma)$ defined by (43), so we have

(a) the functional C_M is not strictly convex in I;
(b) the functional C_M is strictly convex in $I^+ := \{\Gamma \in I\ :\ C_{M_2}(\Gamma) > 0\}$.

Proof. We use a more compact expression for momentum functional.

$$\begin{cases} C_{M_1} = \displaystyle\int_{\gamma_b^0} \Gamma(\eta) H_1(\eta) \mathrm{d}\eta \\ C_{M_2} = \displaystyle\sum_{i=0}^{N-1} \int_{\gamma_b^0} \int_{\gamma_b^i} \Gamma(\eta)\Gamma'(\xi) H_2^i(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \end{cases}$$

Given $\Gamma,\ g \in I$ and $t > 0$ we have

$$\begin{aligned} R_h(t) = \quad & \frac{C_M(\Gamma + tg) - C_M(\Gamma)}{t} \\ = \quad & \frac{C_{M_1}(\Gamma + tg) - C_{M_1}(\Gamma)}{t} + \frac{C_{M_2}(\Gamma + tg) - C_{M_2}(\Gamma)}{t} \\ = \quad & \frac{1}{t}\Bigg[\int (\Gamma(\eta) + tg(\eta)) H_1(\eta)\mathrm{d}\eta - \int \Gamma(\eta) H_1(\eta)\mathrm{d}\eta \\ & + \sum_{i=0}^{N-1} \iint_i (\Gamma(\eta) + tg(\eta)) \left(\Gamma'(\xi) + tg'(\xi)\right) H_2^i(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \end{aligned}$$

$$- \sum_{i=0}^{N-1} \iint_i \Gamma(\eta)\Gamma'(\xi)H_2^i(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \Bigg] =$$

$$= \int g(\eta)H_1(\eta)\mathrm{d}\eta + \sum_{i=0}^{N-1} \iint_i \left(\Gamma(\eta)g'(\xi) + g(\eta)\Gamma'(\xi)\right) H_2^i(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta$$

$$+ t \sum_{i=0}^{N-1} \iint_i g(\eta)g'(\xi)\, H_2^i(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \,.$$

Hence the first derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[R_h(t)\right] = \int\int g(\eta)\, g'(\xi)\, H_2(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta = C_{M_2}(g) \,.$$

Note that the first derivative of quotient ratio, for $g \in I$ is not, in general, positive. Whereas for $g \in I^+$ we do have $C_{M_2}(g) > 0$, then the function $R_h(t)$ is strictly increasing. □

Appendix 4: Velocity Induced by the Bounded Vorticity

For wing the induction of the bounded vortex filament is in direction of the asymptotic velocity. That is the reason why this contribution is neglected in [3]. For airscrew the velocity seen by the blade is the composition of asymptotic and rotational velocity, thus, the velocity induced by the bounded filament does not have, in general, the same direction of the velocity seen by the blade. For straight blade propeller the velocity induced by the bounded vorticity, $\mathbf{u}_{B\,ind}(\eta)$, is null for reason of symmetry, but in general this contribution is not zero. According to the Biot-Savart law we have

$$\mathbf{u}_{B\,ind}(\eta) = \frac{1}{4\pi}\sum_{i=0}^{N-1}\int_{\gamma_b^i} \Gamma(\xi)\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}\xi} \times \frac{\mathbf{r}(\eta) - \mathbf{r}_i(\xi)}{\|\mathbf{r}(\eta) - \mathbf{r}_i(\xi)\|^3}\mathrm{d}\xi \,, \tag{45}$$

where $\mathbf{r}(\eta)$ is the position vector on the induced blade and $\mathbf{r}_i(\xi)$ on the inducing blade. Note that for $i = 0$ the integrand is singular. We find that

$$\frac{\mathrm{d}\mathbf{r}_0}{\mathrm{d}\xi} \times \frac{\mathbf{r}(\eta) - \mathbf{r}_0(\xi)}{\|\mathbf{r}(\eta) - \mathbf{r}_0(\xi)\|^3}\mathrm{d}\xi = \frac{f(\eta)}{(\xi - \eta)} + \frac{\mathscr{O}(\xi - \eta)}{(\xi - \eta)} \,,$$

where $f(\eta)$ is a regular function, thus there are not problem of existence in our class of functions $\mathscr{X}$, defined by (20). The following figures show the results obtained for a swept blade when the contribution of bounded vorticity is not neglected (Figs. 20, 21, 22, and 23). The comparison with the correspondent figures in Sect. 7, where bounded vorticity is neglected, is interesting because it shows the importance of the velocity induced by the bounded vorticity.

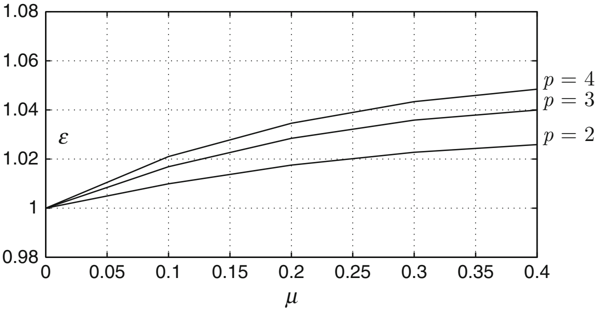


Fig. 20 Relative efficiency of blades swept in rotation plane for $N = 1$, $\Omega = 2$ and $C_{T\,target} = 0.1$

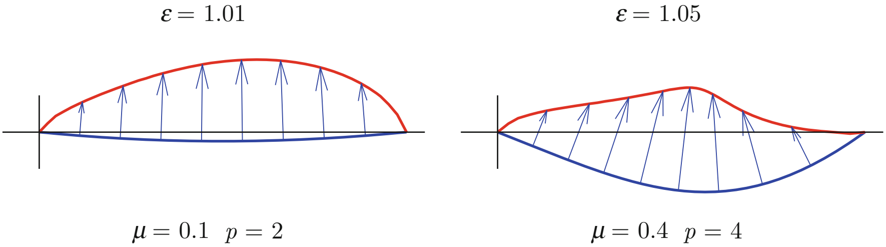


Fig. 21 Circulation distribution for blades swept in rotation plane for $N = 1$, $\Omega = 2$ and $C_{T\,target} = 0.1$

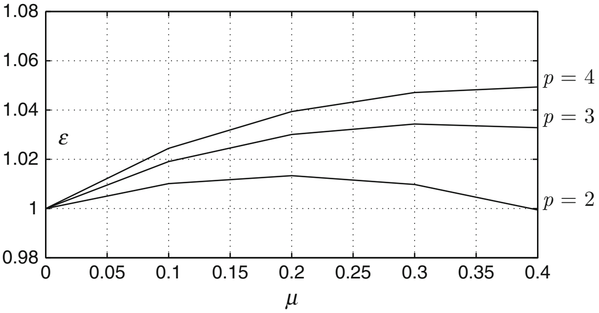


Fig. 22 Relative efficiency of blades swept in axial direction for $N = 1$, $\Omega = 2$ and $C_{T\,target} = 0.1$

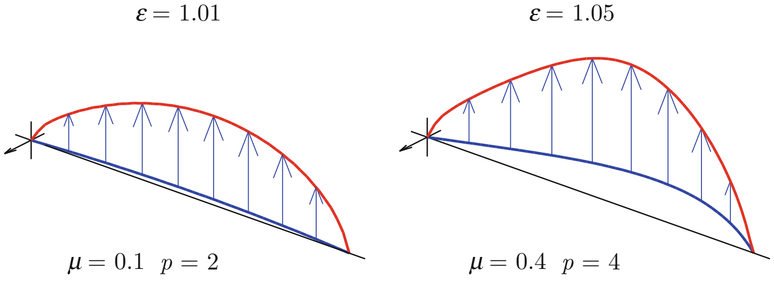


Fig. 23 Circulation distribution for blades swept in axial direction for $N = 1, \Omega = 2$ and $C_{T\,target} = 0.1$

References

1. Betz, A., Prandtl, L.: Schraubenpropeller mit geringstem Energieverlust. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse 193–217 (1919)
2. Chattot, J.J.: Optimization of propellers using helicoidal vortex model. Comput. Fluid Dyn. J. **9**, (2000)
3. Demasi. L., Dipace, A., Monegato, G., Cavallaro, R.: An Invariant formulation for the minimum induced drag conditions of non-planar wing systems. AIAA J **52**, 2223–2240 (2014)
4. Dorfling, J., Rokhsaz, K.: Constrained and unconstrained propeller blade optimization. AIAA J 52 (2015)
5. Elsgolc, L.E.: Calculus of Variation. Pergamon Press, Oxford (1961)
6. Froude, W.: On the elementary relation between pitch, slip and propulsive efficiency. Trans. Ist. Nav. Archit. **19**, 47 (1878)
7. Glauert, H.: Airplane propellers. In: Durand, W.F. (ed.) Aerodynamic Theory. Julius Springer, Berlin (1935)
8. Goldstein, S.: On the vortex theory of screw propellers. Proc. R Soc. Lond. (1929). Doi: 10.1098/rspa.1929.0078
9. Joukowsky, N.Y.: Theorie Turbillonnaire de l'hélice Propulsive. Guathier-Villars, Paris (1929)
10. Larrabea, E.: The screw propeller. Sci. Am. **243**, 134–148 (1980)
11. Monegato, G.: The numerical evaluation of a 2-D Cauchy principal value integral arising in boundary integral equation methods. Math. Comp. (1994). doi: 10.2307/2153536
12. Panaro, M.T., Frediani, A., Giannessi, F., Rizzo, E.: Variational approach to the problem of the minimum induced drug of wings. In: Buttazzo, G., Frediani, A. (eds.) Variational Analysis and Aerospace Engineering. Springer, New York (2009)
13. Pistolesi, E.: La Teoria dei vortici in aerodinamica. L'Aeronautica IV (1921)
14. Pistolesi, E.: Aerodinamica. UTET, Torino (1932)
15. Rankine, W.J.M.: On the mechanical principles of the action of propellers. Trans. Ist. Nav. Archit. **6**, 13 (1865)
16. Theodorsen, T.: Theory of Propellers. McGraw-Hill, New York (1948)
17. Tibery, C.L., Wrench, J.W. Jr.: Tables of Goldstein Factor. David Taylor Model Basin, Washington DC (1964)
18. Tricomi, F.G.: Equazioni integrali contenenti il valor principale di un integrale doppio. In: Lichtenstein L (ed) Mathematische Zeitschrift. Springer, Berlin (1928)

Another View on Planar Anisotropy: The Polar Formalism

Paolo Vannucci

Abstract The polar formalism is a mathematical method, based upon a complex variable transformation, proposed in 1979 by G. Verchery for representing plane tensors of any rank using invariants and angles. As such, it is particularly suited for representing anisotropic properties, in particular elasticity.

In this paper, we give a brief account of the fundamentals of the polar formalism, stressing in particular the role played by the polar invariants on the characterization of elastic symmetries, that leads to a new classification of them, based upon an algebraic criterion and that has allowed for the discovery of two special orthotropies.

Then, we focus on some special theoretical subjects: anisotropy of complex or rari-constant layers, some strange cases of interaction between geometry and anisotropy, the anisotropy of damaged layers initially isotropic.

1 Introduction: Why the Polar Formalism?

In 1979 G. Verchery presented a memory about the invariants of an elasticity-type tensor [41]. This short paper marks the birth of the *polar formalism* or *method*.

For anisotropic materials the Cartesian components of a tensor describing a given property all depend upon the direction in a rather cumbersome way: *none of these components are an intrinsic quantity: all of them are frame-dependent parameters*. In addition, if a privileged direction linked to the anisotropic property exists, *it does not appear explicitly*.

The polar formalism is an *algebraic technique to represent a plane tensor using only tensor invariants and angles* (that is why the method is called *polar*). Hence, the intrinsic quantities describing a given anisotropic property and the direction directly and explicitly appear in the equations. It is exactly the use of *invariants and angles that makes the polar method interesting for analyzing anisotropic phenomena*: on one side, the invariants are not linked to the particular choice of

P. Vannucci (✉)
LMV - Laboratoire de Mathmatiques de Versailles - UMR8100, CNRS - UVSQ, University Paris-Saclay, 45, Avenue des Etats-Unis, 78035 Versailles, France
e-mail: paolo.vannucci@uvsq.fr

A. Frediani et al. (eds.), *Variational Analysis and Aerospace Engineering*,
Springer Optimization and Its Applications 116, DOI 10.1007/978-3-319-45680-5_19

the axes, so they give an intrinsic representation of the property. On the other side, the explicit use of angles makes appear directly one of the fundamental aspects of anisotropy: the direction.

Moreover, the invariants used in the polar formalism are linked to the elastic symmetries: *they represent in an invariant way the symmetries*.

Because the polar invariants represent intrinsically the symmetries, the polar formalism opens the way to a new approach to the analysis of the material symmetries. While in a traditional approach the analysis of the symmetries is essentially *geometric*, in the polar formalism it is intrinsically *algebraic*.

In addition, the polar method allows for obtaining *much simpler formulae for the rotation of the axes* than the classical Cartesian ones.

The entire method is based upon the use of a special complex variable transformation, that is why it can be used only for representing plane tensors.

The subject treated in this chapter is, first, a short recall of the fundamentals of the polar formalism (a deeper insight on this subject can be found in [32, p. 65] and [33], besides the original memory from Verchery), then a presentation of different recent results.

2 The Transformation of Verchery

As said in the introduction, the polar formalism is an algebraic technique based upon the use of a complex variable change. Verchery introduces a special transformation that allows for obtaining particularly simple transformation matrices, namely diagonal matrices for the rotations and anti-diagonal matrices for mirror symmetries.

The transformed of Verchery introduces a complex variable change, interpreted as a change of frame: let us consider a vector $\mathbf{x} = (x_1, x_2)$, and the transformation

$$\mathsf{X}^1 = \frac{1}{\sqrt{2}}\bar{k}z, \quad \mathsf{X}^2 = \overline{\mathsf{X}}^1, \quad k = e^{i\frac{\pi}{4}}, \tag{1}$$

giving the contravariant components of $\mathsf{X}^{cont} = (X^1, X^2)$, the transformed of $\mathbf{x}$ (the transformation is not orthogonal). Equation (1) is the *transformation of Verchery*; z is the complex variable

$$z = x_1 + ix_2. \tag{2}$$

Matrix $\mathbf{m}_1$ operates the transformation of rank-1 tensors, and it has some remarkable algebraic properties, that can be readily found, and that are shared by all the matrices $\mathbf{m}_j$ that operates the transformation for rank-j tensors.

We skip, for the sake of conciseness, all the rather technical passages leading to obtain the transformation for an elasticity-like fourth-rank tensor (the reader interested in the matter is addressed to [32] and [33]), and we go directly to the result:

$$
\begin{aligned}
&E_{1111}(\theta)=T_0+2T_1+R_0\cos 4\,(\Phi_0-\theta)\,+4R_1\cos 2\,(\Phi_1-\theta),\\
&E_{1112}(\theta)=R_0\sin 4\,(\Phi_0-\theta)\,+2R_1\sin 2\,(\Phi_1-\theta),\\
&E_{1122}(\theta)=-T_0+2T_1-R_0\cos 4\,(\Phi_0-\theta),\\
&E_{1212}(\theta)=T_0-R_0\cos 4\,(\Phi_0-\theta),\\
&E_{1222}(\theta)=-R_0\sin 4\,(\Phi_0-\theta)\,+2R_1\sin 2\,(\Phi_1-\theta),\\
&E_{2222}(\theta)=T_0+2T_1+R_0\cos 4\,(\Phi_0-\theta)\,-4R_1\cos 2\,(\Phi_1-\theta).
\end{aligned}
\tag{3}
$$

The above equations give the Cartesian components of an elastic tensor $\mathbb{E}$, in a frame rotated counterclockwise through an angle θ with respect to the x_1 axis, as functions of θ and of four *polar moduli*, T_0, T_1, R_0, R_1, and two *polar angles*, Φ_0 and Φ_1. The four polar moduli and the difference of the polar angles, $\Phi_0-\Phi_1$, constitute a complete set of independent tensor invariants: they are intrinsic quantities, i.e. they are frame independent. Fixing one of the two polar angles corresponds to fix a frame (the choice usually done is $\Phi_1=0$). The reverse of the above equations are

$$
\begin{aligned}
8T_0 &= E_{1111}(\theta)-2E_{1122}(\theta)+4E_{1212}(\theta)+E_{2222}(\theta),\\
8T_1 &= E_{1111}(\theta)+2E_{1122}(\theta)+E_{2222}(\theta),\\
8R_0e^{4i(\Phi_0-\theta)} &= E_{1111}(\theta)-2E_{1122}(\theta)-4E_{1212}(\theta)+E_{2222}(\theta)+\\
&\quad+4i\left[E_{1112}(\theta)-E_{1222}(\theta)\right],\\
8R_1e^{2i(\Phi_1-\theta)} &= E_{1111}(\theta)-E_{2222}(\theta)+2i\left[E_{1112}(\theta)+E_{1222}(\theta)\right].
\end{aligned}
\tag{4}
$$

Equation (3) show one of the greatest advantages of the polar formalism: *the Cartesian components in the new frame are obtained simply subtracting the angle θ from the polar angles*.

Looking at Eq. (3) we can see hence that each Cartesian component of an elastic tensor is the sum of different terms, and in the most general case, that of $E_{1111}(\theta)$ and $E_{2222}(\theta)$, we have

- an invariant term, T_0+2T_1, which represents the mean value of the components and its isotropic part; T_0 and T_1 are hence the *isotropic polar invariants*;
- a term which is a circular function of 2θ whose amplitude is proportional to R_1;
- a term which is a circular function of 4θ whose amplitude is proportional to R_0;
- these two terms are shifted of an angle $\Phi_0-\Phi_1$; this term is an invariant, so we have its physical meaning;
- R_0, R_1, and $\Phi_0-\Phi_1$ are hence the *anisotropic polar invariants*;
- R_0 and R_1 represent, to within a factor, the amplitude of the anisotropic phases, that are directional fluctuations around the isotropic average.

We have hence a new interpretation of anisotropic elasticity in $\mathbb{R}^2$: the anisotropic elastic behavior can be regarded as a finite sum of harmonics: a constant term, the isotropic phase, and two anisotropic phases, one varying with 2θ, the other one with 4θ. The amplitude of all of these phases and the phase offset of the anisotropic phases are intrinsic properties of the material, i.e. they are tensor invariants.

Equation (3) is valid for any elastic tensor, hence also for $\mathbb{E}^{-1}$, that we will indicate by $\mathbb{S}$. We denote the polar components of $\mathbb{S}$ by lowercase letters: t_0, t_1, r_0, r_1 and $\varphi_0 - \varphi_1$. These can be found expressing the Cartesian components of $\mathbb{S}$ as functions of those of $\mathbb{E}$, and these last by their polar components, Eq. (3). Comparing the result so found with Eq. (3) written for $\mathbb{S}$ gives $t_0, t_1, r_0, r_1, \varphi_0$, and φ_1. The calculations are rather heavy and only the final result is given here:

$$\begin{aligned}
t_0 &= \frac{2}{\Delta}\left(T_0T_1 - R_1^2\right),\\
t_1 &= \frac{1}{2\Delta}\left(T_0^2 - R_0^2\right),\\
r_0e^{4i\varphi_0} &= \frac{2}{\Delta}\left(R_1^2e^{4i\varPhi_1} - T_1R_0e^{4i\varPhi_0}\right),\\
r_1e^{2i\varphi_1} &= -\frac{R_1e^{2i\varPhi_1}}{\Delta}\left[T_0 - R_0e^{4i(\varPhi_0-\varPhi_1)}\right].
\end{aligned} \tag{5}$$

Equation (5) being symmetric, i.e. we can switch $\mathbb{E}$ and $\mathbb{S}$, we notice that

$$R_1 = 0 \Leftrightarrow r_1 = 0, \quad R_0 = 0 \nLeftrightarrow r_0 = 0. \tag{6}$$

Equation (6) has a considerable importance in the determination of all the elastic symmetries, analyzed in Sect. 3.

The positiveness of the strain energy V gives the bounds on the components of $\mathbb{E}$, so also on its polar invariants. It can be proved [37] that these bounds reduce to only four inequalities:

$$\begin{aligned}
&T_0 - R_0 > 0,\\
&T_1(T_0^2 - R_0^2) - 2R_1^2\left[T_0 - R_0\cos 4(\varPhi_0 - \varPhi_1)\right] > 0,\\
&R_0 \geq 0,\\
&R_1 \geq 0.
\end{aligned} \tag{7}$$

Conditions (7) imply that the isotropic part of $\mathbb{E}$ is strictly positive:

$$T_0 > 0, \;\; T_1 > 0. \tag{8}$$

The above *four intrinsic conditions* (7) *are valid for a completely anisotropic planar material.*

3 Elastic Symmetries

We consider in this section the existence of elastic symmetries for tensor $\mathbb{E}$. The polar formalism leads to a *general algebraic relation characterizing all the types of elastic symmetry in* $\mathbb{R}^2$:

$$R_0 R_1^2 \sin 4(\Phi_0 - \Phi_1) = 0. \tag{9}$$

Such condition depends upon three invariants, $R_0, R_1, \Phi_0 - \Phi_1$, and can be satisfied when these invariants take some special values. Each root of Eq. (9) corresponds to a different case of elastic symmetry in $\mathbb{R}^2$. To remark that condition (9) is an *intrinsic characterization of elastic symmetries in* $\mathbb{R}^2$, because it makes use of only tensor invariants. So, all the following special cases are also intrinsic conditions of orthotropy and so on. We consider all of them separately.

3.1 Ordinary Orthotropy

The first solution to (9) that we consider is

$$\begin{aligned}
&\sin 4(\Phi_0 - \Phi_1) = 0 \ \Rightarrow \ \ \Phi_0 - \Phi_1 = K\frac{\pi}{4}, \ K \in \{0, 1\} \ \Rightarrow \\
&(E_{1112} - E_{1222}) \left[(E_{1111} - E_{2222})^2 - 4\,(E_{1112} + E_{1222})^2 \right] - \\
&(E_{1112} + E_{1222})\,(E_{1111} - E_{2222})\,(E_{1111} - 2E_{1122} - 4E_{1212} + E_{2222}) = 0.
\end{aligned} \tag{10}$$

Condition (10) depends upon a cubic invariant. It characterizes intrinsically *ordinary orthotropy*, i.e. common orthotropy, as the particular anisotropic situation where *the shift angle between the two anisotropy phases is a multiple of* $\pi/4$; clearly, due to the periodicity of the functions, only two cases are meaningful: 0 or $\pi/4$.

This result shows that, generally speaking, for the same set of invariants T_0, T_1, R_0, and R_1 *two possible and distinct orthotropic materials can exist*: one with $K = 0$ and the other one with $K = 1$.

If for an ordinarily orthotropic material a frame rotation of Φ_1 is operated, Eq. (3) can be written as

$$\begin{aligned}
E_{1111}(\theta) &= T_0 + 2T_1 + (-1)^K R_0 \cos 4\theta + 4R_1 \cos 2\theta,\\
E_{1112}(\theta) &= -\,(-1)^K R_0 \sin 4\theta - 2R_1 \sin 2\theta,\\
E_{1122}(\theta) &= -T_0 + 2T_1 - (-1)^K R_0 \cos 4\theta,\\
E_{1212}(\theta) &= T_0 - (-1)^K R_0 \cos 4\theta,\\
E_{1222}(\theta) &= (-1)^K R_0 \sin 4\theta - 2R_1 \sin 2\theta,\\
E_{2222}(\theta) &= T_0 + 2T_1 + (-1)^K R_0 \cos 4\theta - 4R_1 \cos 2\theta.
\end{aligned} \tag{11}$$

This is the form of the polar representation normally used for orthotropic layers; of course, it corresponds to choose the frame where $\Phi_1 = 0$.

The parameter K, that is an invariant, characterizes ordinary orthotropy; its importance has been observed in different studies [34]. In particular K plays a fundamental role in several optimization problems: an optimal solution to a given problem becomes the *anti-optimal*, i.e. the worst one, when K switches from 0 to 1 and vice versa: if a solution is optimal for a material with $K = 0$ (or $K = 1$) it is *anti-optimal*, i.e. the worst one, for a material with $K = 1 (K = 0)$.

Two questions concern $\mathbb{S}$, the inverse of $\mathbb{E}$: how is it oriented the orthotropy of $\mathbb{S}$ and of which type is it?

It can be proved that

$$\varphi_1 = \Phi_1 + \frac{\pi}{2} \tag{12}$$

i.e., $\mathbb{S}$ is *always turned of* $\pi/2$ *with respect to* $\mathbb{E}$, and that

$$\left.\begin{array}{l} K = 0 \text{ and } R_1^2 > T_1 R_0 \\ \text{or} \\ K = 1 \end{array}\right\} \Rightarrow \ k = 0, \\ K = 0 \text{ and } R_1^2 < T_1 R_0 \ \Rightarrow \ k = 1. \tag{13}$$

So, an elasticity tensor and its inverse, when ordinarily orthotropic, *can be of a different type*; in particular, the possible combinations are three: $(K = 0, k = 0)$, $(K = 0, k = 1)$, $(K = 1, k = 0)$.

3.2 R_0-orthotropy

The general equation of elastic symmetries in $\mathbb{R}^2$, Eq. (9), can be satisfied also by other conditions than root (10). Algebraically speaking, unlike in the case of ordinary orthotropy, detected by a cubic invariant, all the other solutions are linked to special values get by *quadratic* invariants and they are characterized by the vanishing of at least one of the two anisotropic phases. For these reasons, such cases of elastic symmetry are called *special orthotropies*, besides the last case, that of isotropy. To analyze these cases, it is, however, necessary to choose $\mathbb{E}$, i.e. to decide wether it is a stiffness or a compliance tensor. Conventionally, R_0-orthotropy concerns stiffness. We will see further why this choice is necessary.

First, we consider the case of a material for which

$$R_0 = 0. \tag{14}$$

Of course, this is a root of Eq. (9), so the above condition identifies a special orthotropy, the so-called R_0-orthotropy [31]. The discovery of this type of special orthotropy has been done thanks to the polar formalism and it constitutes a rather strange case of elastic behavior.

It is easily recognized that

$$R_0 = 0 \;\Rightarrow\; (E_{1111} - 2E_{1122} - 4E_{1212} + E_{2222})^2 + 16(E_{1112} - E_{1222})^2 = 0. \quad (15)$$

Though this case of elastic symmetry presents two orthogonal axes of mirror symmetry, just like in ordinary orthotropy, it has some peculiar characteristics. First of all, the Cartesian components of an R_0-orthotropic material are (we have put $\Phi_1 = 0$ for fixing the frame)

$$\begin{aligned} E_{1111}(\theta) &= T_0 + 2T_1 + 4R_1 \cos 2\theta, \\ E_{1112}(\theta) &= -2R_1 \sin 2\theta, \\ E_{1122}(\theta) &= -T_0 + 2T_1, \\ E_{1212}(\theta) &= T_0, \\ E_{1222}(\theta) &= -2R_1 \sin 2\theta, \\ E_{2222}(\theta) &= T_0 + 2T_1 - 4R_1 \cos 2\theta. \end{aligned} \quad (16)$$

For a material R_0-orthotropic, the anisotropic phase depending on R_0 is absent. By consequence, some of the components, E_{1122} and E_{1212}, are *isotropic*, while the other ones, depending upon the circular functions of 2θ, *change like the components of a second-rank tensor*. We are hence faced to a very strange case, that of a fourth-rank tensor whose components do not vary according to the tensor law and, in addition, with some of them frame independent. Moreover, unlike what happens in all the other cases of anisotropy, $E_{1112}(\theta) = E_{1222}(\theta)\ \forall\theta$.

Let us now consider what happens for the compliance tensor $\mathbb{S}$: when $R_0 = 0$, Eq. (5) becomes

$$\begin{aligned} t_0 &= \frac{T_0 T_1 - R_1^2}{4T_0(T_0 T_1 - 2R_1^2)}, \\ t_1 &= \frac{T_0}{16(T_0 T_1 - 2R_1^2)}, \\ r_0 e^{4i\varphi_0} &= \frac{R_1^2 e^{4i\Phi_1}}{4T_0(T_0 T_1 - 2R_1^2)}, \\ r_1 e^{2i\varphi_1} &= -\frac{R_1 e^{2i\Phi_1}}{8(T_0 T_1 - 2R_1^2)}, \end{aligned} \quad (17)$$

and by consequence

$$\begin{aligned} r_0 &= \frac{R_1}{4T_0(T_0T_1-2R_1^2)}, \qquad \varphi_0=\Phi_1,\\ r_1 &= \frac{R_1}{8(T_0T_1-2R_1^2)}, \qquad \varphi_1=\Phi_1+\frac{\pi}{2}. \end{aligned} \tag{18}$$

In obtaining the results in Eq. (18), it has been considered that the denominator of both the terms is positive, cf. further in this section.

As remarked in Eq. (6), $R_0 = 0 \nRightarrow r0 = 0$: the compliance tensor $\mathbb{S}$ depends on both the anisotropic phases, that is, its components preserve a higher degree of symmetry than those of $\mathbb{E}$. This is a rather unusual case, where stiffness and compliance of the same material do not have the same kind of variation, the same morphology. In addition, tensor $\mathbb{S}$ has always $k = 0$.

Nevertheless, just like $\mathbb{E}$, also $\mathbb{S}$ depends upon only three independent nonzero invariants, because it is easily recognized from Eq. (17) that

$$r_0=\frac{r_1^2}{t_1}. \tag{19}$$

Hence, once a frame chosen fixing Φ_1, φ_0 and φ_1 are fixed too, and the only polar moduli t_0, t_1 and r_1 are sufficient to completely determine $\mathbb{S}$. If we put $\Phi_1 = 0$, we obtain

$$\begin{aligned} S_{1111} &= t_0+2t_1+\frac{r_1^2}{t_1}\cos 4\theta-4r_1\cos 2\theta,\\ S_{1112} &= -\frac{r_1^2}{t_1}\sin 4\theta+2r_1\sin 2\theta,\\ S_{1122} &= -t_0+2t_1-\frac{r_1^2}{t_1}\cos 4\theta,\\ S_{1212} &= t_0-\frac{r_1^2}{t_1}\cos 4\theta,\\ S_{1222} &= \frac{r_1^2}{t_1}\sin 4\theta+2r_1\sin 2\theta,\\ S_{2222} &= t_0+2t_1+\frac{r_1^2}{t_1}\cos 4\theta+4r_1\cos 2\theta. \end{aligned} \tag{20}$$

Finally, one can wonder if R_0-orthotropic materials do really exist. Actually, they do; in fact, it is rather simple, using the polar formalism and the classical lamination theory, to see that an R_0-orthotropic lamina can be fabricated reinforcing an isotropic matrix by unidirectional fibers arranged in equal quantity along two directions tilted of 45, [31].

3.3 r_0-Orthotropy

It has already been noticed that relations (5) are perfectly symmetric, i.e., they can be rewritten swapping the polar compliance constants with the polar stiffness constants, i.e., putting uppercase letters at the left-hand side and lowercase letters at the right-hand side of relations (5). This circumstance, together with Eq. (6), i.e. the fact that whenever $R_0 = 0$, then $r_0 \neq 0$, implies the existence of another special orthotropy, an analog of R_0-orthotropy, but concerning compliance, not stiffness: it will be indicated in the following as r_0-orthotropy [31]. So, we can see that a R_0-orthotropic layer is not also r_0-orthotropic, and vice versa. In this sense, special orthotropies of the type R_0 are *more a symmetry of a tensor than that of a material*, in the sense that a material, e.g., R_0-orthotropic, has a compliance tensor that, at least apparently,[1] has a common orthotropic behavior: the orthotropy axes do not change from stiffness to compliance, but the mechanical behavior is different in the two cases.

Of course, all the remarks done and results found in the previous section for R_0-orthotropy are still valid for r_0-orthotropy, it is sufficient to change the lowercase letters with capital letters to all the polar components and the word *stiffness* with the word *compliance*.

Putting $\varphi_1 = 0$, the compliance tensor $\mathbb{S}$ looks like

$$\begin{aligned}
S_{1111}(\theta) &= t_0 + 2t_1 + 4r_1 \cos 2\theta,\\
S_{1112}(\theta) &= -2r_1 \sin 2\theta,\\
S_{1122}(\theta) &= -t_0 + 2t_1,\\
S_{1212}(\theta) &= t_0,\\
S_{1222}(\theta) &= -2r_1 \sin 2\theta,\\
S_{2222}(\theta) &= t_0 + 2t_1 - 4r_1 \cos 2\theta,
\end{aligned} \tag{21}$$

which gives

$$G_{12}(\theta) = \frac{1}{4S_{1212}(\theta)} = \frac{1}{4t_0}: \tag{22}$$

the shear modulus $G_{12}(\theta)$ is isotropic. It was observed experimentally since 1951 that common paper is anisotropic but its shear modulus is independent from the direction [4, 14]; only recently an explanation of this fact in the framework of classical elasticity has been done, thanks to the polar formalism [35].

[1] Apparently because if one makes experimental tests on the components of $\mathbb{S}$ or traces the directional diagrams of its components, they look like those of an ordinarily orthotropic material with $k = 0$, the difference is in the special value get by r_0, Eq. (19).

Just like for R_0-orthotropy, only three nonzero independent invariants are sufficient to completely determine $\mathbb{S}$: t_0, t_1, r_1. The general bounds (7) become, for r_0-orthotropy,

$$t_0 > \frac{2r_1^2}{t_1}, \quad r_1 > 0, \tag{23}$$

like in the case of R_0-orthotropy, so also in this case, of course, only two intrinsic bounds are sufficient.

Finally, just like for the previous case of R_0-orthotropic materials, it is easy to see that for the stiffness tensor it is

$$R_0 = \frac{R_1^2}{T_1}, \quad K = 0. \tag{24}$$

3.4 Square Symmetry

Another root of the general equation of elastic symmetries in $\mathbb{R}^2$, Eq. (9), is

$$R_1 = 0. \tag{25}$$

Just like the case of R_0-orthotropy, also in this case an anisotropy phase, the one varying with 2θ, vanishes, so it is a special orthotropy, determined once more by a quadratic invariant:

$$R_1 = 0 \ \Rightarrow \ (E_{1111} - E_{2222})^2 + 4(E_{1112} + E_{1222})^2 = 0. \tag{26}$$

In this case, the polar angle Φ_1 is meaningless, so the frame can be fixed only fixing a value for Φ_0. Choosing $\Phi_0 = 0$, the Cartesian components of $\mathbb{E}$ are

$$\begin{aligned}
E_{1111}(\theta)&=T_0+2T_1+R_0\cos 4\theta,\\
E_{1112}(\theta)&=-R_0\sin 4\theta,\\
E_{1122}(\theta)&=-T_0+2T_1-R_0\cos 4\theta,\\
E_{1212}(\theta)&=T_0-R_0\cos 4\theta,\\
E_{1222}(\theta)&=R_0\sin 4\theta,\\
E_{2222}(\theta)&=T_0+2T_1+R_0\cos 4\theta.
\end{aligned} \tag{27}$$

We can remark that all the components are periodic of $\pi/2$:

$$E_{ijkl}\left(\theta + \frac{\pi}{2}\right) = E_{ijkl}(\theta) \ \ \forall\theta. \tag{28}$$

For this reason, this special orthotropy is known in the literature as *square symmetry* and actually, it is the corresponding, in $\mathbb{R}^2$, of the cubic syngony. These materials can be fabricated reinforcing an isotropic matrix with a balanced fabric, i.e. by a fabric having the same amount of fibers in warp and weft. Unlike the case of R_0-orthotropy, for Eq. (6)$_1$ when a material has $R_1 = 0$ it has also $r_1 = 0$: square symmetry is a property of both the stiffness and the compliance tensors. The general bound for the polar invariants (7) becomes now

$$T_1(T_0 - R_0) > 0, \quad R_0 \geq 0. \tag{29}$$

3.5 Isotropy

The last possible syngony for a planar material is isotropy; in this case, every angle α must determine the direction of a mirror symmetry or, which is the same, that Cartesian components of $\mathbb{E}$ are insensitive to the direction. This gives the condition

$$R_0 = R_1 = 0 \;\; \Rightarrow E_{1112} = E_{1222} = 0, \;\; E_{2222} = E_{1111}, \;\; E_{1111} = E_{1122} + 2E_{1212}. \tag{30}$$

Isotropy is hence characterized by the fact that the two anisotropy phases are null; it can be remarked also that a material is isotropic if and only if the conditions for the two special orthotropies are satisfied at the same time: algebraically, isotropy is determined by the vanishing of two quadratic invariants. Alternatively, isotropy can be determined by a unique condition in place of the two polar relations $R_0 = R_1 = 0$,

$$\begin{aligned} &R_0^2 + R_1^2 = 0 \;\Rightarrow \\ &\left[(E_{1111} - 2E_{1122} - 4E_{1212} + E_{2222})^2 + 16(E_{1112} - E_{1222})^2\right]^2 + \\ &\left[(E_{1111} - E_{2222})^2 + 4(E_{1112} + E_{1222})^2\right]^2 = 0 \end{aligned} \tag{31}$$

which makes use of a fourth degree invariant.

4 Special Plane Elastic Anisotropic Materials

The analysis of plane anisotropy made so far is tacitly based upon the assumption of *classical elastic body*. The mechanical response of such a body is described by an elastic tensor $\mathbb{E}$ characterized by having the minor and major symmetries:

$$\mathbb{E}_{ijkl} = \mathbb{E}_{jikl} = \mathbb{E}_{ijlk} = \mathbb{E}_{jilk}, \quad \mathbb{E}_{ijkl} = \mathbb{E}_{klij}. \tag{32}$$

A large part of existing materials belong to such a category, namely the most part of materials used for structural purposes, like metallic alloys, wood, composite materials, concrete, and so on.

Nevertheless, materials with different tensor symmetries can exist and we briefly consider them in this section. On one side, going towards an ancient, celebrated scientific diatribe in elasticity, we first consider the so-called *rari-constant materials*, having supplementary tensor symmetries adding to the minor and major ones of classical materials. Then, now looking at the most recent researches in mechanics of materials, we shortly analyze *complex materials*, calling with this name all the elastic materials that do not possess all of the minor or major symmetries.

There is a characteristic fact in all these cases: *the number of tensor symmetries is linked to the number of tensor invariants*. In particular, we will see that to any increase of the number of tensor symmetries corresponds a decrease of the number of tensor invariants and vice versa, when the tensor symmetries decrease, the number of tensor invariants increases.

Because the tensor invariants are linked to the material symmetries, it is to be expected that also the pattern of material symmetries changes with that of the tensor symmetries. Actually, this is not automatic: it is so for complex materials, where the number and types of special orthotropies are radically changed with respect to the case of classical materials, but it is not so for rari-constant materials, whose anisotropic part is not affected by the presence of supplementary tensor symmetries. This result puts in evidence an important fact: *there is a link between the possible material symmetries and the tensor symmetries*, i.e. the type of the mechanical response of the material. This fact shows once more that a mere analysis of anisotropy based exclusively upon considerations of *geometric* symmetries of the matter cannot be exhaustive.

In all the cases, however, the study is greatly facilitated by the use of the polar formalism: the different conditions of symmetry of the elastic response emerge directly and simply as purely algebraic conditions offered by the analysis of the polar invariants, while an analysis based upon considerations of symmetry of the matter should be rather cumbersome.

4.1 Rari-Constant Materials

4.1.1 A Brief Historical Background

Elastic materials whose behavior is described by a smaller number of parameters have been widely studied in the past and their existence has been the subject of one of the most famous diatribes in the theory of elasticity: that between what Pearson [29, p. 496] named *multi-constant* and *rari-constant* materials [1, v. 1, p. 227], [12, p. 398], [18, p. 6, p. 13].

The idea of rari-constant materials stems from the early works of Navier [20] and his model of matter, known as *molecular theory*, first presented at *Académie des Sciences* on May 14, 1821. Basically, the model proposed by Navier aims at explaining the behavior of elastic solids as that of a lattice of particles (*molecules*) interacting together via central forces proportional to their mutual distance. This is not a new idea: it has its last foundation in the works of Newton [21]. For what concerns the mechanics of solids, the true initiator of the molecular theory is considered to be Boscovich [3]; other works on this topic, before the *mémoire* of Navier, are those of Poisson [24, 25] on the equilibrium of bent plates, while subsequent fundamental contributions are due to [6, 7], still Poisson [26] and Saint Venant [10].

The basic idea in the classical molecular approach of Navier and Cauchy, the continuum as a limit of a discrete lattice of particles interacting together via central forces, has a direct consequence [11, 28]: 15 moduli describe the behavior of a completely anisotropic body in 3D, and only one modulus suffices to determine it for an isotropic material. These results were not confirmed by experimental tests, so doubts existed about its validity, until the molecular approach was completely bypassed by the theory proposed in 1837 by Green [13]: no underlying microscopic structure of the matter, considered as a continuum, is assumed, and the basic property defining the elastic behavior is *energetic*: in non-dissipative processes the internal forces derive from a quadratic potential.

The consequences of such an assumption lead to the multi-constant model: 21 independent moduli are necessary to describe the elastic response of a completely anisotropic body in 3D, which reduce to only 2 for an isotropic material. The results of the Green's theory were confirmed by experience which, together with its much simpler theoretical background, ensured the success of the multi-constant theory. Nonetheless, the diatribe between the molecular, rari-constant, and continuum, multi-constant theories lasted a long period: which is the right number of elastic constants and the correct model of elastic continuum?

The further developments of the molecular model by Voigt [42] and Poincaré [23] are refined models that, enriching in different ways the original model of Navier, obtain multi-constant theories starting from a molecular model, see [5, 22]. More recently, ideas inspired by the Navier–Cauchy approach have produced molecular dynamics models or models for explicating the behavior of complex bodies.

As an effect of this diatribe, the two models are usually considered as opposing and somewhat irreconcilable, though different researchers have made attempts to show that this is not the case [2], [18, Note B, p. 616], [19, p. 55].

The results presented below [38] concern the planar case and show some new results for an old problem: there exist two dual types of rari-constant materials and the classical Cauchy–Poisson conditions are not sufficient to characterize such a material: the only true necessary and sufficient condition is the number of independent linear tensor invariants, that must be of one.

4.1.2 The Polar Approach to the Study of Planar Rari-Constant Materials

Within the classical paradigm of elasticity, $\mathbb{E}$ possesses the minor and major symmetries of the indexes, Eq. (32), so describing a so-called *multi-constant material*: we know that for the complete anisotropic case, a whole of 21 independent components (18 tensor invariants and 3 frame dependent parameters) determine the material behavior; they reduce to only two for the isotropic case. In the plane case, there are five invariants plus a quantity taking into account for the frame orientation.

Let us ponder the consequences of the existence of six supplementary index symmetries, the so-called *Cauchy–Poisson symmetries*:

$$E_{ijkl} = E_{ikjl}, \tag{33}$$

that for the plane case reduce to the only supplementary condition

$$E_{1122} = E_{1212}. \tag{34}$$

It is immediate to recognize that in such a case the behavior is described by only 12 tensor invariants plus three quantities fixing the frame, for a whole of 15 independent components. In the plane case, we have five independent components, four of which are invariants, and isotropy is always described by a unique invariant quantity. Finally, the existence of supplementary index symmetries decreases the number of the material parameters needed to describe the material behavior; that is why, materials of such a type are called *rari-constant*.

Let us concentrate on the planar case; since now, we identify *rari-constant tensors* with those satisfying the Cauchy–Poisson conditions, and we show that identifying *rari-constant materials* is not so simple, because there are two possible and dual rari-constant materials, at least in $\mathbb{R}^2$.

We can easily state now the algebraic conditions for the *elastic tensor* $\mathbb{E}$ in $\mathbb{R}^2$ to be rari-constant:

Theorem 1. *$\mathbb{E}$ is a rari-constant elastic tensor in $\mathbb{R}^2$* $\iff$ $T_0 = T_1$.

Proof. The proof is immediate: if $\mathbb{E}$ is a rari-constant tensor, then $E_{1212}(\theta) = E_{1122}(\theta)\ \forall\theta$, and Eq. $(3)_{3,4}$ give $T_0 = T_1$. Conversely, if $T_0 = T_1$, then Eq. $(4)_{1,2}$ give $E_{1212}(\theta) = E_{1122}(\theta)\ \forall\theta$.

Let us consider all the consequences of such a result:

- the number of independent tensor invariants is linked to the number of index symmetries; in particular, a supplementary index symmetry corresponds to the identity of two invariants, so that the number of independent invariants is decreased by one;
- the rari-constant condition affects only the isotropic part of $\mathbb{E}$, i.e. only its linear invariants: the anisotropic part is not touched by the Cauchy–Poisson conditions, so that multi- and rari-constant materials share all the same types of elastic symmetries;

- the bounds on the polar parameters, Eq. (7), do not exclude the existence of the case $T_0 = T_1$: in the classical frame of continuum elastic bodies, materials with a rari-constant tensor $\mathbb{E}$ are *possible*;
- the existence of multi-constant materials with $T_0 = T_1$ *is not allowed*; this point is essential: apparently, just because Eq. (7) do not exclude the case $T_0 = T_1$ for multi-constant materials, then such materials could exist; nevertheless, this is not possible, because of Theorem 1; physically, this means that whenever $T_0 = T_1$, then tensor $\mathbb{E}$ is *necessarily* rari-constant: $E_{1212}(\theta) = E_{1122}(\theta)\ \forall\theta$: a particular value of the tensor invariants determines a change of the algebraic structure of the elastic tensor;

A fundamental remark can now be done: *all what has been said for* $\mathbb{E}$ *is equally valid for* $\mathbb{S}$: we can define a *dual class* of rari-constant materials, where the Cauchy–Poisson conditions are valid for the compliance tensor $\mathbb{S}$. We name in the following *direct*- and *inverse*- rari-constant materials those for which the Cauchy–Poisson condition (34) holds, respectively, for $\mathbb{E}$ or for $\mathbb{S}$. These two classes are *necessarily distinct*, i.e. it cannot exist a material being at the same time direct- and inverse-rari-constant: the Cauchy–Poisson conditions cannot be satisfied at the same time by $\mathbb{E}$ and $\mathbb{S}$. That is why the name rari-constant has been used to denote not only a class of materials, but also a type of elastic tensor: this distinction is necessary in the following.

For proving why a material cannot be at the same time direct- and rari-constant, we need first a preliminary result:

Theorem 2. *The value*

$$T_0 = \frac{4R_1^2 - R_0^2}{3} \tag{35}$$

is incompatible with the elastic bounds (7) on the polar invariants for direct- rari-constant materials, i.e. when $T_1 = T_0$.

Proof. Replacing Eq. (35) into Eq. $(7)_1$ and taking into account for Eq. $(7)_{3,4}$ gives

$$R_1 > R_0 > 0. \tag{36}$$

Now, injecting Eq. (35) into Eq. $(7)_2$ we get, after posing

$$\rho = \frac{R_0}{R_1},\ \ C = \cos 4(\Phi_0 - \Phi_1),\ \ 0 \le \rho < 1,\ -1 \le C \le 1, \tag{37}$$

$$\sqrt{\frac{4-\rho^2}{3}} < \frac{3\rho\, C}{1+2\rho^2}, \tag{38}$$

a condition that is satisfied if and only if

$$\begin{cases} \dfrac{4-\rho^2}{3} \geq 0, \\ \dfrac{3\rho\, C}{1+2\rho^2} \geq 0, \\ \dfrac{4-\rho^2}{3} < \dfrac{9\rho^2 C^2}{(1+2\rho^2)^2}. \end{cases} \tag{39}$$

Condition (39)$_1$ gives $\rho \leq 2$, which is redundant because of Eq. (37)$_3$, condition (39)$_2$ limits Eq. (37)$_4$ to $0 \leq C \leq 1$ while condition (39)$_3$ can be rewritten as

$$f = \frac{(4-\rho^2)(1+2\rho^2)^2}{27\rho^2} < C^2, \tag{40}$$

which is never satisfied because $f > 1 = \max C^2$ for $0 \leq \rho < 1$, as it can be easily recognized.

The isotropic case is trivial, for Eq. (35) should give $T_0 = 0$ which corresponds to a material with a null stiffness, hence it is impossible.

The two cases of special orthotropies are also impossible; in fact, the case of square symmetry, $R_1 = 0$, should imply a negative value for T_0^2, Eq. (35), while that of R_0-orthotropy, $R_0 = 0 \Rightarrow \rho = 0$, gives $f \to \infty$.

Theorem 3. *The Cauchy–Poisson condition (34) cannot be satisfied at the same time by* $\mathbb{E}$ *and* $\mathbb{S}$.

Proof. Be $\mathbb{E}$ rari-constant, i.e. $E_{1122} = E_{1212}$; then $T_0 = T_1$ by Theorem 1. The polar invariants of $\mathbb{S}$ can then be calculated through Eq. (5) that in this case become

$$\begin{aligned} t_0 &= \frac{2}{\Delta}\left(T_0^2 - R_1^2\right), \\ t_1 &= \frac{1}{2\Delta}\left(T_0^2 - R_0^2\right), \\ r_0 e^{4i\varphi_0} &= \frac{2}{\Delta}\left(R_1^2 e^{4i\Phi_1} - T_0 R_0 e^{4i\Phi_0}\right), \\ r_1 e^{2i\varphi_1} &= -\frac{R_1 e^{2i\Phi_1}}{\Delta}\left[T_0 - R_0 e^{4i(\Phi_0 - \Phi_1)}\right]. \end{aligned} \tag{41}$$

with

$$\Delta = 8T_0\left(T_0^2 - R_0^2\right) - 16R_1^2\left[T_0 - R_0 \cos 4\left(\Phi_0 - \Phi_1\right)\right]. \tag{42}$$

It is then apparent that

$$t_0 = t_1 \iff T_0^2 = \frac{4R_1^2 - R_0^2}{3}. \tag{43}$$

This value of T_0 is incompatible with the elastic bounds (7), as shown in Theorem 2, and hence, $t_0 \neq t_1$ when $T_0 = T_1$, so by Theorem 1 applied to $\mathbb{S}$, $S_{1212} \neq S_{1122}$.

The consequence is immediate: it is not correct to identify *automatically* rari-constant materials in $\mathbb{R}^2$ with the Cauchy–Poisson condition, because this concerns only one of the two elastic tensors of the material.

So, if $\mathbb{E}$ is rari-constant, it has only five distinct Cartesian components, but its inverse, $\mathbb{S}$ has six different components. Conversely, if $\mathbb{S}$ is rari-constant, it has five distinct Cartesian components, but they are 6 for $\mathbb{E}$. Nevertheless, in both the cases the number of independent tensor invariants is 4. In fact, if $\mathbb{E}$ is rari-constant, then $T_0 = T_1$ and by Eq. (41) we get

$$t_1 = \frac{T_0^2 - R_0^2}{4(T_0^2 - R_1^2)} t_0. \tag{44}$$

Hence, though $t_1 \neq t_0$, it is proportional to t_0. Of course, a similar relation exists for the dual case of $\mathbb{S}$ rari-constant, it is sufficient to swap lower- and upper-case letters.

Finally, there are two dual families of rari-constant materials:

- the *direct* rari-constant materials:

$$\begin{aligned}
&E_{1212}(\theta) = E_{1122}(\theta)\ \forall\theta,\\
&T_0 = T_1,\\
&S_{1212}(\theta) \neq S_{1122}(\theta),\\
&t_1 = \frac{T_0^2 - R_0^2}{4(T_0^2 - R_1^2)} t_0,
\end{aligned} \tag{45}$$

  and

$$\begin{aligned}
&E_{1111}(\theta) = 3T_0 + R_0\cos 4\,(\Phi_0 - \theta) + 4R_1 \cos 2\,(\Phi_1 - \theta),\\
&E_{1112}(\theta) = R_0 \sin 4\,(\Phi_0 - \theta) + 2R_1 \sin 2\,(\Phi_1 - \theta),\\
&E_{1122}(\theta) = E_{1212}(\theta) = T_0 - R_0 \cos 4\,(\Phi_0 - \theta),\\
&E_{1222}(\theta) = -R_0 \sin 4\,(\Phi_0 - \theta) + 2R_1 \sin 2\,(\Phi_1 - \theta),\\
&E_{2222}(\theta) = 3T_0 + R_0 \cos 4\,(\Phi_0 - \theta) - 4R_1 \cos 2\,(\Phi_1 - \theta).
\end{aligned} \tag{46}$$

- the *inverse* rari-constant materials:

$$\begin{aligned}
&S_{1212}(\theta) = S_{1122}(\theta)\ \forall\theta,\\
&t_0 = t_1,\\
&E_{1212}(\theta) \neq E_{1122}(\theta)\ \forall\theta,\\
&T_1 = \frac{t_0^2 - r_0^2}{4(t_0^2 - r_1^2)} T_0,
\end{aligned} \tag{47}$$

and

$$\begin{aligned}
&S_{1111}(\theta)=3t_0+r_0\cos 4\,(\varphi_0-\theta)+4r_1\cos 2\,(\varphi_1-\theta),\\
&S_{1112}(\theta)=r_0\sin 4\,(\varphi_0-\theta)+2r_1\sin 2\,(\varphi_1-\theta),\\
&S_{1122}(\theta)=S_{1212}(\theta)=t_0-r_0\cos 4\,(\varphi_0-\theta),\\
&S_{1222}(\theta)=-r_0\sin 4\,(\varphi_0-\theta)+2r_1\sin 2\,(\varphi_1-\theta),\\
&S_{2222}(\theta)=3t_0+r_0\cos 4\,(\varphi_0-\theta)-4r_1\cos 2\,(\varphi_1-\theta).
\end{aligned} \tag{48}$$

Finally, if we consider that special orthotropies are characterized by the vanishing of a tensor invariant, i.e. $R_0 = 0$ for the case of R_0-orthotropy, while $R_1 = 0$ for square-symmetry, or by being an invariant a function of the other ones, for the case of r_0-orthotropic materials, then it is clear that the only necessary and sufficient condition for identifying a rari-constant material, regardless of its type, i.e. independently of the number of distinct Cartesian components for $\mathbb{E}$ or $\mathbb{S}$, is that the number of *independent linear tensor invariants must be one*.

Two last remarks: first, while rari-constant materials can actually exist, multi-constant materials with $T_0 = T_1$ or $t_0 = t_1$ are not allowed. Then, condition $T_0 = T_1$ clearly indicates that the anisotropic part of a plane elastic tensor is necessarily rari-constant; in other words, in $\mathbb{R}^2$ only the isotropic part is responsible for the multi-constant behavior.

4.1.3 The Isotropic Case

Two isotropic rari-constant materials can exist, the direct and the inverse one. Let us briefly consider their properties.

The direct case first: isotropy is characterized by the vanishing of the anisotropic part, i.e. by

$$R_0 = R_1 = 0 \Rightarrow r_0 = r_1 = 0. \tag{49}$$

The stiffness behavior is uniquely determined by T_0:

$$E_{1111}(\theta) = E_{2222}(\theta) = 3T_0,\ E_{1122}(\theta) = T_0,\ E_{1112}(\theta) = E_{1222}(\theta) = 0\ \forall\theta. \tag{50}$$

For $\mathbb{S}$, it is

$$t_0 = \frac{1}{4T_0},\ t_1 = \frac{1}{16T_0} \Rightarrow t_0 = 4t_1 \tag{51}$$

and, $\forall\theta$,

$$\begin{aligned}
S_{1111}(\theta) &= S_{2222}(\theta) = t_0 + 2t_1 = \frac{3}{2}t_0 = \frac{3}{8T_0},\\
S_{1122}(\theta) &= -t_0 + 2t_1 = -\frac{t_0}{2} = -\frac{1}{8T_0},\\
S_{1212}(\theta) &= t_0 = \frac{1}{4T_0} \Rightarrow S_{1212}(\theta) = -2S_{1122}(\theta),\\
S_{1112}(\theta) &= S_{1222}(\theta) = 0.
\end{aligned} \tag{52}$$

We can also introduce the classic technical constants:

$$\begin{aligned}
E &:= \frac{1}{S_{1111}} = \frac{8}{3}T_0,\ \nu := -\frac{S_{1122}}{S_{1111}} = \frac{1}{3},\\
G &:= \frac{1}{4S_{1212}} = T_0,\ \kappa := \frac{1}{S_{1111} + 2S_{1122} + S_{2222}} = 2T_0.
\end{aligned} \tag{53}$$

It is then apparent the mechanical meaning of T_0: it is equal to the shear modulus G for the isotropic case; the result for the Poisson's coefficient is also classical, but it is worth to remark that it is only a *necessary* but *not sufficient* for a material to be direct- rari-constant: multi-constant materials with $\nu = 1/3$ do exist. Also, for these materials the bulk modulus κ is twice the shear modulus: they have a stiffness to spherical stress states that is the double of that to shear states. For the normal stiffness, this is $8/3$ times the shear one.

Finally, for what concerns the Lamé's constants, it is

$$\lambda := \kappa - G = T_0,\ \mu := G = T_0 \Rightarrow \lambda = \mu, \tag{54}$$

a classical result.

Let us now turn the attention to inverse- rari-constant materials; now, t_0 uniquely determines all the distinct components of $\mathbb{S}$:

$$S_{1111}(\theta) = S_{2222}(\theta) = 3t_0,\ S_{1122}(\theta) = t_0,\ S_{1112}(\theta) = S_{1222}(\theta) = 0\ \forall\theta. \tag{55}$$

For tensor $\mathbb{E}$, we get

$$T_0 = \frac{1}{4t_0},\ T_1 = \frac{1}{16t_0} \Rightarrow T_0 = 4T_1 \tag{56}$$

and, $\forall\theta$,

$$\begin{aligned}
&E_{1111}(\theta)=E_{2222}(\theta)=T_0+2T_1=\frac{3}{2}T_0=\frac{3}{8t_0},\\
&E_{1122}(\theta)=-T_0+2T_1=-\frac{T_0}{2}=-\frac{1}{8t_0},\\
&E_{1212}(\theta)=T_0=\frac{1}{4t_0}\Rightarrow E_{1212}(\theta)=-2E_{1122}(\theta),\\
&E_{1112}(\theta)=E_{1222}(\theta)=0.
\end{aligned} \tag{57}$$

Now, the technical constants are

$$E=\frac{4}{3}T_0,\ \nu=-\frac{1}{3},\ G=T_0,\ \kappa=\frac{T_0}{2}. \tag{58}$$

Inverse- rari-constant materials are hence necessarily materials with a negative Poisson's coefficient, whose value is exactly the opposite of the direct case; nevertheless, they can exist. Also, their normal stiffness is just half and their bulk modulus a fourth of the corresponding direct case ones. Now, the spherical stiffness is smaller than the shear one.

The Lamé's constants now are

$$\lambda=-\frac{T_0}{2},\ \mu=T_0\Rightarrow\lambda=-\frac{\mu}{2}, \tag{59}$$

i.e. λ is negative; nevertheless, thanks to Eq. (7), the bounds on the values of the Lamé's constants in $\mathbb{R}^2$ are satisfied:

$$\mu=T_0>0,\lambda+\mu=\frac{T_0}{2}>0. \tag{60}$$

4.2 Complex Materials

The case of complex materials, indicating here bodies whose elastic tensor has only a part of the minor and/or major symmetries, has been addressed in a theoretical work [40]. In particular, two cases have been examined: the first one, when $\mathbb{E}$ does not have the minor symmetries, and the second one when it has not the major ones. In both the cases, the way to handle the problem is that typical of the polar approach, of course modified by the different number of index symmetries. This is actually the key point: the influence that the index symmetries have on the anisotropic behavior. We give here only some details on both the cases considered in the cited paper, referring the reader to the same article for the theoretical developments, that are rather technical.

Let us first consider the case of a material whose elastic tensor does not have the minor symmetries; in such a case, the elastic tensor has 10 independent Cartesian components in the planar case, represented by an equal number of polar parameters, nine of them being tensor invariants:

$$
\begin{aligned}
E_{1111} &= T_0 + T_1 + T_2 + R_0 \cos 4\Phi_0 + 2R_1 \cos 2\Phi_1 + 2R_2 \cos 2\Phi_2,\\
E_{1112} &= -T_3 + R_0 \sin 4\Phi_0 + 2R_2 \sin 2\Phi_2,\\
E_{1121} &= T_3 + R_0 \sin 4\Phi_0 + 2R_1 \sin 2\Phi_1,\\
E_{1122} &= -T_0 + T_1 + T_2 - R_0 \cos 4\Phi_0,\\
E_{1212} &= T_0 + T_1 - T_2 - R_0 \cos 4\Phi_0 + 2R_1 \cos 2\Phi_1 - 2R_2 \cos 2\Phi_2,\\
E_{1221} &= T_0 - T_1 + T_2 - R_0 \cos 4\Phi_0,\\
E_{1222} &= -T_3 - R_0 \sin 4\Phi_0 + 2R_1 \sin 2\Phi_1,\\
E_{2121} &= T_0 + T_1 - T_2 - R_0 \cos 4\Phi_0 - 2R_1 \cos 2\Phi_1 + 2R_2 \cos 2\Phi_2,\\
E_{1112} &= T_3 - R_0 \sin 4\Phi_0 + 2R_2 \sin 2\Phi_2,\\
E_{2222} &= T_0 + T_1 + T_2 + R_0 \cos 4\Phi_0 - 2R_1 \cos 2\Phi_1 - 2R_2 \cos 2\Phi_2.
\end{aligned}
\tag{61}
$$

The invariants are all the polar moduli T_0, T_1, etc. and the differences of the polar angles $\Phi_0 - \Phi_2$ and $\Phi_1 - \Phi_2$. For this case, it can be proved that ordinary orthotropy corresponds to the conditions

$$
\Phi_0 - \Phi_1 = K_{01}\frac{\pi}{4}, \quad \Phi_0 - \Phi_2 = K_{02}\frac{\pi}{4}, \quad \Phi_1 - \Phi_2 = K_{12}\frac{\pi}{2}. \tag{62}
$$

As a consequence, there are four possible different ordinary orthotropic materials sharing the same polar moduli and determined by the combinations $K_{02} = K_{12} = 0,\ K_{02} = 1$ and $K_{12} = 0,\ K_{02} = 0$ and $K_{12} = 1, K_{02} = K_{12} = 1$.

Besides these four ordinary cases, there are six different special orthotropies, characterized by the following conditions:

$$
\begin{aligned}
&R_0 = 0,\ K_{12} = 0, \quad R_0 = 0,\ K_{12} = 1,\\
&R_1 = 0,\ K_{02} = 0, \quad R_1 = 0,\ K_{02} = 1,\\
&R_2 = 0,\ K_{01} = 0, \quad R_2 = 0,\ K_{01} = 1.
\end{aligned}
\tag{63}
$$

It is interesting also to remark that for these materials, isotropy is given by the conditions:

$$
T_3 = R_0 = R_1 = R_2 = 0; \tag{64}
$$

there is hence a condition on a linear invariant, T_3, needed to ensure the invariance of the material response under a mirror symmetry about an axis. The relations between the Cartesian and polar components in this case are

$$\begin{aligned}
E_{1111} &= E_{2222} = T_0 + T_1 + T_2,\\
E_{1122} &= -T_0 + T_1 + T_2,\\
E_{1212} &= E_{2121} = T_0 + T_1 - T_2,\\
E_{1221} &= T_0 - T_1 + T_2,
\end{aligned} \tag{65}$$

the remaining components being null. Isotropy is hence determined by three independent moduli, not by two as for classical materials.

For the second case, a tensor without the major symmetries, there are nine independent components and it is

$$\begin{aligned}
E_{1111} &= T_0 + 2T_1 + R_0 \cos 4\Phi_0 + 2R_1 \cos 2\Phi_1 + 2R_2 \cos 2\Phi_2,\\
E_{1112} &= -T_3 + R_0 \sin 4\Phi_0 + 2R_2 \sin 2\Phi_2,\\
E_{1122} &= -T_0 + 2T_1 - R_0 \cos 4\Phi_0 + 2R_1 \cos 2\Phi_1 - 2R_2 \cos 2\Phi_2,\\
E_{1211} &= T_3 + R_0 \sin 4\Phi_0 + 2R_1 \sin 2\Phi_1,\\
E_{1212} &= T_0 - R_0 \cos 4\Phi_0,\\
E_{1222} &= -T_3 - R_0 \sin 4\Phi_0 + 2R_1 \sin 2\Phi_1,\\
E_{2211} &= -T_0 + 2T_1 - R_0 \cos 4\Phi_0 - 2R_1 \cos 2\Phi_1 + 2R_2 \cos 2\Phi_2,\\
E_{1121} &= T_3 - R_0 \sin 4\Phi_0 + 2R_2 \sin 2\Phi_2,\\
E_{2222} &= T_0 + 2T_1 + R_0 \cos 4\Phi_0 - 2R_1 \cos 2\Phi_1 - 2R_2 \cos 2\Phi_2.
\end{aligned} \tag{66}$$

The elastic behaviors in the two cases, in terms of invariants, differ only for a term of the isotropic part, this implying as additional result that the whole discussion of anisotropy does not change with respect to the previous case. In particular, the number and types of orthotropies are quite the same of the previous case. In particular, it is easily seen that isotropy in this case perfectly coincides with that of classical materials.

The two examples briefly considered above clearly show that there is an influence of the tensor symmetries, i.e. of the algebraic structure of the elastic tensor, on the elastic symmetries.

Finally, the following considerations can be done:

- the number of independent invariants, and hence of parameters determining intrinsically the behavior of the material, depends upon the number and type of index symmetries;
- the number and types of algebraically distinct types of orthotropy depend upon the index symmetries of the elastic tensor;
- in some cases of special orthotropy, some Cartesian components are null, or constant or vary with the orientation angle like a second-rank tensor component;

- the number of independent elastic constants in isotropy is two only if the elastic tensor has the minor indicial symmetries, otherwise the constants are three;
- it is easily recognized that classical hyperelastic materials can be recovered from more complex elastic materials as a particular case.

5 Interaction of Geometry and Anisotropy

Anisotropy is a mere fact of material properties and, as such, normally its effects should be the same no matter of the problem at hand, i.e. the anisotropic behaviour should not be altered by other factors. Actually, this is not the case. We show in this section that, in the case of flat plates at least, the geometry and boundary conditions of the plate interact with the anisotropic properties of the plate, so modifying its overall elastic response to, e.g., buckling, or vibrations [36].

The following considerations arise from a research concerning the influence of anisotropy on the flexural response of laminates, [34]. Here, the scope is to show that, in some sense, geometry *filters* the anisotropy of the plate. To better understand, let us consider again Eq. (3); as already remarked, each Cartesian component is just the superposition of a maximum of three contributions: a constant term, the isotropic part, and two oscillating terms, two waves with period, respectively, $\pi/2$ and $\pi/4$, representing together the anisotropic part.

Well, geometry is able, in some cases, to *make one or both of this two waves disappear in the elastic response of a plate*. In some way, *geometry acts on material properties just like, in signal analysis, a filter acts upon the harmonics of a signal: a part or all of the oscillating terms describing the anisotropy of the material disappear from the elastic response of the plate*. That is why we can talk, in such cases, of *filtering anisotropy*, an unusual expression in composite mechanics for a curious phenomenon.

To show this fact, we consider here the flexural behavior of an anisotropic rectangular laminate composed by identical layers and bending-extension uncoupled, with sides length a and b, respectively, along the axes x and y. Along its boundary, the plate is simply supported and its mass per unit area is μ. A constraint is imposed on the anisotropy of the plate: the bending tensor $\mathbb{D}$ is orthotropic and the axes of orthotropy are aligned with the axes of the plate. This assumption, along with that of uncoupling, is needed for having exact solutions for flexural problems, see for, instance, [15]. The question of obtaining uncoupled laminates orthotropic in bending has been addressed in very few works, and the reader can namely refer to [30] for further details on the matter. Here, we bound ourselves to remark that it is possible to find uncoupled laminates with $\mathbb{D}$ orthotropic.

In the above assumptions [34] has shown that (the polar parameters in the following equations are those of the basic layer):

- the compliance J of the plate, that is a measure of its bending stiffness, when loaded by a sinusoidal load orthogonal to its mean surface, is

$$J=\frac{\gamma_{pq}}{p^4h^3(1+\chi^2)^2\sqrt{R_0^2+R_1^2}}\frac{1}{\varphi\left(\xi_0,\xi_1\right)};\tag{67}$$

- the buckling load multiplier λ_{pq} for the mode pq when the plate is loaded by in-plane forces N_x and N_y is given by

$$\lambda_{pq}=\frac{\pi^2p^2h^3}{12a^2}\frac{\left(1+\chi^2\right)^2\sqrt{R_0^2+R_1^2}}{N_x+N_y\chi^2}\varphi\left(\xi_0,\xi_1\right);\tag{68}$$

- the frequency of transversal vibrations ω_{pq} for the mode pq is expressed by

$$\omega_{pq}^2=\frac{\pi^4p^4h^3}{12\mu a^4}\left(1+\chi^2\right)^2\sqrt{R_0^2+R_1^2}\varphi\left(\xi_0,\xi_1\right).\tag{69}$$

In the above equations, h is the laminate's thickness, γ_{pq} is a coefficient depending on the geometry of the plate and on the loading, p and q are the number of half-waves in the directions x and y, respectively, while the dimensionless parameter χ is the ratio of the wavelengths in the two directions, i.e.

$$\chi=\frac{a}{b}\frac{q}{p}.\tag{70}$$

Finally, the function $\varphi(\xi_0,\xi_1)$ is

$$\varphi\left(\xi_0,\xi_1\right)=\tau+\frac{1}{\sqrt{1+\rho^2}}\left[(-1)^k\rho\xi_0\frac{\chi^4-6\chi^2+1}{(1+\chi^2)^2}+4\xi_1\frac{1-\chi^2}{1+\chi^2}\right],\tag{71}$$

and it is a dimensionless function of dimensionless parameters and variables:

$$\tau=\frac{T_0+2T_1}{\sqrt{R_0^2+R_1^2}}\tag{72}$$

is the *isotropy-to-anisotropy ratio*, while ρ is the *anisotropy ratio*

$$\rho=\frac{R_0}{R_1}\tag{73}$$

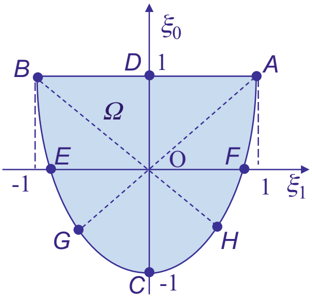


Fig. 1 Domain of the lamination parameters

and

$$\begin{cases} \xi_0 = \dfrac{1}{n^3}\displaystyle\sum_{j=1}^{n} d_j \cos 4\delta_j, \\ \xi_1 = \dfrac{1}{n^3}\displaystyle\sum_{j=1}^{n} d_j \cos 2\delta_j, \end{cases} \qquad d_j = 12j(j-n-1)+4+3n(n+2). \tag{74}$$

are the *bending lamination parameters*, [15], with n the number of layers and δ_j the orientation angles of the plies.

Without rephrasing all the theory of lamination parameters, it is worth noticing that the final bending properties are completely determined by the choice of the material and of a *lamination point*, i.e. of a couple (ξ_0, ξ_1). The set of lamination points defines in the plane ξ_0–ξ_1 a feasible domain Ω having the form of a parabolic sector, bounded by the conditions

$$\begin{gathered} 2\xi_1^2 - 1 \leq \xi_0 \leq 1, \\ -1 \leq \xi_1 \leq 1. \end{gathered} \tag{75}$$

On this domain, some points and lines correspond to particular stacking sequences and in particular, with reference to Fig. 1.

- all the *cross-ply laminates*, i.e. having all the layers at 0 or $\pi/2$, belong to the straight line AB;
- all the *angle-ply laminates*, i.e. having half the layers at the orientation δ and the other half at $-\delta$, belong to the parabolic arc ACB;
- all the *unidirectional laminates*, with all the layers aligned at 0, are represented by the lamination point A;
- all the *unidirectional laminates*, with all the layers aligned at $\pi/2$, are represented by the lamination point B;
- all the *balanced cross-ply laminates*, i.e. having the same number of layers at 0 and at $\pi/2$, are represented by the lamination point D;
- all the angle-ply with $\alpha = \pi/4$ are represented by the lamination point C;

- all the angle-ply with $\alpha = \pi/8$ are represented by the lamination point F;
- all the angle-ply with $\alpha = 3\pi/8$ are represented by the lamination point E;
- all the angle-ply with $\alpha = \pi/6$ are represented by the lamination point H;
- all the angle-ply with $\alpha = \pi/3$ are represented by the lamination point G;
- all the isotropic laminates are represented by the lamination point O.

Nevertheless, there is not a bijective correspondence between lamination points and stacking sequences: a given bending behavior is uniquely determined by one and only one lamination point, but several different stacking sequences can correspond to the same lamination point, and hence be mechanically equivalent. In other words, different laminates can have the same bending behavior. For instance, there is not a unique sequence to obtain isotropy, see, for instance, [39], and on the parabolic boundary of the feasible domain one can find also sequences that do not belong to the angle-ply set.

From the above Eqs. (67)–(71), it is apparent that in the three cases considered here and regarding all the possible situations concerning the flexural response of the plate, this response is always a function of $\varphi(\xi_0, \xi_1)$. Hence, for minimizing J or for maximizing λ_{pq} or ω_{pq}, the problem is always reduced to

$$\max_{\xi_0,\xi_1} \ \varphi\left(\xi_0, \xi_1\right). \tag{76}$$

It is worth noticing that in the dimensionless function $\varphi(\xi_0, \xi_1)$, thanks to the polar formalism, the isotropic part is well separated from the anisotropic one, this last being the only one to interact with geometry. So, in some special cases, $\varphi(\xi_0, \xi_1)$ can be reduced to its only isotropic part:

$$\varphi\left(\xi_0, \xi_1\right) = \tau. \tag{77}$$

In these circumstances, the flexural behavior is no more affected by the stacking sequence, nor by the anisotropy of the material. This annihilation of the anisotropic part of the elastic response happens for particular values of the anisotropy and/or for particular geometries and in such cases, the laminate behaves like an isotropic plate, though it is anisotropic. Let us consider now the possible conditions leading to such a strange situation. First of all, looking at Eq. (71), there are two cases independent from the lamination point $\varphi(\xi_0, \xi_1)$, and given simply by special condition on anisotropy and geometry:

1. $\rho = 0$ and $\chi = 1$; this is the case of R_0-orthotropic materials, for which $R_0 = 0$, and of plates having the same wavelength along x and y:

$$\chi = \frac{a}{b}\frac{q}{p} = 1 \quad \Longleftrightarrow \quad \frac{a}{b} = \frac{p}{q}. \tag{78}$$

This happens for instance for square plates when diagonal modes, i.e. having $p = q$, are considered, but also for other situations. In such cases,

$$\begin{aligned} \varphi\left(\xi_0,\xi_1\right) &= \tau = \frac{T_0+2T_1}{R_1}, \\ J &= \frac{\gamma_{pq}}{4\ p^4h^3\left(T_0+2T_1\right)}, \\ \lambda_{pq} &= \frac{\pi^2p^2h^3}{3\ a^2}\frac{\sqrt{1+\nu^2}}{1+\nu}\frac{T_0+2T_1}{\sqrt{N_x^2+N_y^2}}, \\ \omega_{pq}^2 &= \frac{\pi^4p^4h^3}{3\ \mu\ a^4}\left(T_0+2T_1\right). \end{aligned} \tag{79}$$

So, there is no trace of anisotropy in the expressions of J, λ_{pq} and ω_{pq} in Eq. (79): all the contributions given by the anisotropic part of the material composing the basic layers have disappeared and, despite the fact that the material and the plate are anisotropic, the responses in Eq. (79) depends only upon the isotropic part of the material and they are exactly the same that belong to a plate having the same geometry and composed by an isotropic material whose polar parameters T_0 and T_1 are identical to those of the material actually composing the anisotropic plate. This result is independent of the stack and of the anisotropy of the laminate, i.e. of $\mathbb{D}$. Namely, it is not needed that the laminate be R_0-orthotropic in bending.

2. $\rho=\infty$ and $\chi=\sqrt{2}\pm 1$; this is the case of laminates composed of square-symmetric layers, for which $R_1=0$, and having

$$\frac{a}{b}=\frac{p}{q}\left(\sqrt{2}\pm 1\right). \tag{80}$$

In this case, the observations made for the previous case can be repeated *verbatim*, in particular the laminate does not need to be square-symmetric in bending, and now

$$\begin{aligned} \varphi\left(\xi_0,\xi_1\right) &= \tau = \frac{T_0+2T_1}{R_0}, \\ J &= \frac{\gamma_{pq}}{8\left(3\pm 2\sqrt{2}\right)\ p^4h^3\left(T_0+2T_1\right)}, \\ \lambda_{pq} &= \frac{2\left(3\pm 2\sqrt{2}\right)\pi^2p^2h^3}{3\ a^2}\frac{\sqrt{1+\nu^2}}{1+\nu\left(3\pm 2\sqrt{2}\right)}\frac{T_0+2T_1}{\sqrt{N_x^2+N_y^2}}, \\ \omega_{pq}^2 &= \frac{2\left(3\pm 2\sqrt{2}\right)\pi^4p^4h^3}{3\ \mu\ a^4}\left(T_0+2T_1\right). \end{aligned} \tag{81}$$

There are other sufficient conditions determining an isotropic-like flexural response of the plate: they are all those that render $\varphi(\xi_0, \xi_1) = \tau$, but they depend upon special values of the lamination point, i.e. they are stack-dependent. Some of these sufficient conditions are

3. $\rho = 0$ and $\xi_1 = 0$; this is the case of laminates composed of R_0-orthotropic layers and with a lamination point belonging to the straight line CD; on this line, lie all the combinations of layers with orientations 0, $\pi/2$ and $\pm\pi/4$. Such sequences are called *generalized quasi-isotropic*, and are used extensively in aeronautical composite structures.
4. $\chi = \sqrt{2} \pm 1$ and $\xi_1 = 0$; the lamination points are the same as in the previous case, but now the layers do not have to be R_0-orthotropic, the essential condition concerns now geometry, cfr. case 2.
5. $\rho = \infty$ and $\xi_0 = 0$; it is the case of laminates composed of square symmetric layers and with the lamination point belonging to the line EF. On this line there are the laminates composed by combinations of two angle-ply laminates, one with $\alpha = \pi/8$ and the other one with $\alpha = 3\pi/8$.
6. $\chi = 1$ and $\xi_0 = 0$; the lamination points are the same of the previous case, but now the material properties of the basic layer have no importance, the essential condition concerns now geometry, cfr. case 1.
7. Generally speaking, for a not specially orthotropic material, i.e. for $\rho \neq 0$ and $\rho \neq \infty$, the condition determining $\varphi(\xi_0, \xi_1) = \tau$ is simply

$$\xi_0 = \frac{4}{(-1)^K \rho} \frac{\chi^4 - 1}{\chi^4 - 6\chi^2 + 1} \xi_1. \tag{82}$$

The above equation constitutes, for a given material and geometry, the relation to be satisfied by the lamination parameters in order to obtain a flexural isotropic response. On the domain Ω, this relation corresponds to a straight line, always passing through the origin O (isotropic point).

To end this section, it is interesting to remark that the geometry of the stack and/or of the plate can filter also partially the anisotropy of the material. For instance, Eqs. (70) and (73) show immediately that all the laminates whose lamination point lies on the line EF ($\xi_0 = 0$) and/or having $\chi = \sqrt{2} \pm 1$ cancel the contribution of R_0 to the flexural response of the plate. In this case, the response is just like that of a laminate composed by R_0 -orthotropic plies: the geometry and/or the stack act in such a case as a filter on the anisotropy of the material. A similar effect happens, and this case is more interesting for applications, when the lamination point lies on the line CD ($\xi_1 = 0$) and/or $\chi = 1$. Now, it is the component R_1 to be filtered: the laminate behaves just as one composed of square symmetric layers, i.e. as if it was $R_1 = 0$, without necessarily being square orthotropic.

All the above cases show clearly that in elasticity geometry interacts with the anisotropy of the material or, in other words, that the anisotropy of the same structure can vary according to the cases (for instance, varying the ratio p/q). Of course, the cases presented in this section are particularly simple, due to the geometry of the plate; nevertheless, effects similar to those described above are likely to happen also for other geometries, but analytical solutions cannot be found and a similar analysis is much more difficult to be done.

6 Anisotropic Damage of Isotropic Layers

We consider the anisotropy induced by damage on an initially isotropic layer [37], and in particular the following questions:

- if the elastic tensor of the virgin material is $\mathbb{E}$, which is the final tensor $\widetilde{\mathbb{E}}$?
- what are the bounds for the elastic moduli of $\widetilde{\mathbb{E}}$?
- and those for the characteristics of the damage tensor $\mathbb{D}$?

To this purpose, we define the damage tensor $\mathbb{D}$ as a fourth-rank tensor with minor and major tensor symmetries, such that the elastic tensor $\widetilde{\mathbb{E}}$ of the damaged material linearly depends upon $\mathbb{E}$ and $\mathbb{D}$ [8, 9, 16, 17, 27]:

$$\widetilde{\mathbb{E}} = \mathbb{E} - \widehat{\mathbb{E}}, \quad \text{with} \quad \widehat{\mathbb{E}} = \frac{\mathbb{E}\mathbb{D} + \mathbb{D}\mathbb{E}}{2}, \tag{83}$$

The elastic tensor $\mathbb{E}$ of the virgin material and the damaged elastic tensor $\widetilde{\mathbb{E}}$ must be positive definite, as a consequence of the positiveness of the elastic potential. In a thermodynamical framework, the positive semi-definiteness of the loss of stiffness tensor $\widehat{\mathbb{E}}$ is equivalent to a positive intrinsic dissipation due to linear elasticity-damage coupling.[2] The damage tensor $\mathbb{D}$ is assumed to be positive semi-definite.

The above questions can be effectively investigated using the polar formalism: $\mathbb{E}$ and $\mathbb{D}$ can be represented by the classical polar representation for elasticity-like tensors and the bounds for their positiveness are known, Eq. (7) explicitly.

We then obtain the polar invariants of $\widetilde{\mathbb{E}}$ as functions of those of $\mathbb{E}$ and $\mathbb{D}$; while we assume that the initial material is isotropic, we consider all the possible transformations for the damaged material, leading to a final elastic behavior that can be completely anisotropic, orthotropic, specially orthotropic or also isotropic.

If the damage tensor $\mathbb{D}$ is represented within the polar formalism as

$$\begin{aligned}
D_{1111}(\theta)&=D_0+2D_1+S_0\cos 4\left(\Psi_0-\theta\right)+4S_1\cos 2\left(\Psi_1-\theta\right),\\
D_{1112}(\theta)&=S_0\sin 4\left(\Psi_0-\theta\right)+2S_1\sin 2\left(\Psi_1-\theta\right),\\
D_{1122}(\theta)&=-D_0+2D_1-S_0\cos 4\left(\Psi_0-\theta\right),\\
D_{1212}(\theta)&=D_0-S_0\cos 4\left(\Psi_0-\theta\right),\\
D_{1222}(\theta)&=-S_0\sin 4\left(\Psi_0-\theta\right)+2S_1\sin 2\left(\Psi_1-\theta\right),\\
D_{2222}(\theta)&=D_0+2D_1+S_0\cos 4\left(\Psi_0-\theta\right)-4S_1\cos 2\left(\Psi_1-\theta\right).
\end{aligned} \tag{84}$$

[2] For a proof of this statement, see [37], Sect. 4 and Appendix.

then the damaged stiffness tensor $\widetilde{\mathbb{E}}$ is given by

$$\begin{aligned}
\widetilde{E}_{1111}(\theta) &= T_0(1-2D_0)+2T_1(1-4D_1)-2T_0S_0\cos 4(\Psi_0-\theta)-\\
&\qquad -4(T_0+2T_1)S_1\cos 2(\Psi_1-\theta),\\
\widetilde{E}_{1112}(\theta) &= -2T_0S_0\sin 4(\Psi_0-\theta)-2(T_0+2T_1)S_1\sin 2(\Psi_1-\theta),\\
\widetilde{E}_{1122}(\theta) &= -T_0(1-2D_0)+2T_1(1-4D_1)+2T_0S_0\cos 4(\Psi_0-\theta),\\
\widetilde{E}_{1212}(\theta) &= T_0(1-2D_0)+2T_0S_0\cos 4(\Psi_0-\theta),\\
\widetilde{E}_{1222}(\theta) &= 2T_0S_0\sin 4(\Psi_0-\theta)-2(T_0+2T_1)S_1\sin 2(\Psi_1-\theta),\\
\widetilde{E}_{2222}(\theta) &= T_0(1-2D_0)+2T_1(1-4D_1)-2T_0S_0\cos 4(\Psi_0-\theta)+\\
&\qquad +4(T_0+2T_1)S_1\cos 2(\Psi_1-\theta),
\end{aligned} \tag{85}$$

with T_0 and T_1 the unique two polar components of the undamaged isotropic stiffness tensor $\mathbb{E}$. We find the polar parameters of $\widetilde{\mathbb{E}}$, indicated in the following by a $\sim$, comparing the above equations with the usual polar expressions of an elastic tensor, Eq. (3):

$$\begin{aligned}
&\widetilde{T}_0 = T_0(1-2D_0), \quad \widetilde{T}_1 = T_1(1-4D_1),\\
&\widetilde{R}_0 = 2T_0S_0, \quad \widetilde{R}_1 = (T_0+2T_1)S_1,\\
&\widetilde{\Phi}_0 = \Psi_0+\tfrac{\pi}{4}, \quad \widetilde{\Phi}_1 = \Psi_1+\tfrac{\pi}{2}.
\end{aligned} \tag{86}$$

Some remarks about these results:

- an advantage of the polar formalism, apparent from the developments above, is the uncoupling of the expressions of the parameters of $\widetilde{\mathbb{E}}$ as functions of those of $\mathbb{D}$: each one of the polar parameters of $\widetilde{\mathbb{E}}$ depends exclusively upon the corresponding polar parameter of $\mathbb{D}$;
- ***Equation (86) shows that the damage symmetries, i.e. the corresponding for $\mathbb{D}$ of the ordinary and special orthotropies of $\mathbb{E}$, determine, each one, exactly the same elastic symmetry of the same type for the damaged elastic tensor $\widetilde{\mathbb{E}}$ and inversely;
- ***Equations $(85)_{1,6}$ and $(86)_6$ show that the axis of the strongest component of $\widetilde{\mathbb{E}}$, i.e. $\widetilde{E}_{2222}$, is turned of $\pi/2$ with respect to the direction of the strongest component of $\mathbb{D}, D_{1111}$. This is quite natural, because the material is more severely damaged along the direction of D_{1111}, so that, finally, $\widetilde{E}_{1111}(\theta=0) < \widetilde{E}_{2222}(\theta=0)$. Also the harmonic depending upon 4θ is turned of $\pi/4$, Eq. $(86)_5$, which gives for the angular invariant of $\widetilde{\mathbb{E}}$

$$\widetilde{\Phi}_0-\widetilde{\Phi}_1=\Psi_0-\Psi_1-\frac{\pi}{4}. \tag{87}$$

- the last result shows a rather surprising fact: the damaged elasticity tensor $\widetilde{\mathbb{E}}$ cannot have the same form of ordinary orthotropy of the damage tensor $\mathbb{D}$. In fact, for $\mathbb{D}$ orthotropic with

$$\Psi_0 - \Psi_1 = L\frac{\pi}{4},\ L = \{0, 1\}, \tag{88}$$

$\widetilde{\mathbb{E}}$ is orthotropic with

$$\widetilde{\Phi}_0 - \widetilde{\Phi}_1 = \widetilde{K}\frac{\pi}{4},\ \widetilde{K} = L - 1. \tag{89}$$

So, for $L = 1, \widetilde{K} = 0$ and for $L = 0, \widetilde{K} = 1$ (the sign does not matter).

The conditions of positive semi-definiteness for $\mathbb{D}$ and $\widehat{\mathbb{E}}$ and positive definiteness for $\widetilde{\mathbb{E}}$ provide the conditions to determine the bounds on the values of their moduli, once those on $\mathbb{E}$ known. It can be proved that the positive semi-definiteness of $\widehat{\mathbb{E}}$ always implies the positive semi-definiteness of $\mathbb{D}$, and is even equivalent in some particular cases related to the induced anisotropy by damage.

Using the polar formalism, it is possible to give an explicit expression for the bounds on the polar invariants of $\mathbb{D}$ and $\widehat{\mathbb{E}}$, starting from the simpler case, that of an isotropic tensor $\widetilde{\mathbb{E}}$, we consider all the possible cases of elastic symmetries for $\widetilde{\mathbb{E}}$, until the most general case of a completely anisotropic $\widetilde{\mathbb{E}}$, and show that the admissible domain for the moduli is convex in all the cases; in some of them a graphical representation is also possible. The results in the most general case are summarized in Table 1, where the following ratios have been introduced:

$$\tau_1 = \frac{2T_1}{T_0},\ \tilde{\tau}_0 = \frac{\widetilde{T}_0}{T_0},\ \tilde{\tau}_1 = \frac{2\widetilde{T}_1}{T_0},\ \tilde{\rho}_0 = \frac{\widetilde{R}_0}{T_0},\ \tilde{\rho}_1 = \frac{\widetilde{R}_1}{T_0}. \tag{90}$$

The ratio $\tau_1 > 0$ will hence be the only term representing the mechanical characteristics of the undamaged material.

The general results presented in Table 1 can be easily specialized to the different cases of material symmetry of the final behavior:

Table 1 Minimal set of dimensionless polar bounds in the completely anisotropic case

	Polar bounds for $\mathbb{D}$	Polar bounds for $\widetilde{\mathbb{E}}$
B5	$2(D_0 + S_0) < 1$	$\tilde{\tau}_0 > \tilde{\rho}_0$
B6	$\frac{\tau_1}{4(1+\tau_1)^2}(1 - 4D_1)[(1 - 2D_0)^2 - -4S_0^2] > S_1^2[1 - 2D_0 + 2S_0\cos 4(\Psi_0 - \Psi_1)]$	$\tilde{\tau}_1(\tilde{\tau}_0^2 - \tilde{\rho}_0^2) > 4\tilde{\rho}_1^2\left[\tilde{\tau}_0 - \tilde{\rho}_0\cos 4(\widetilde{\Phi}_0 - \widetilde{\Phi}_1)\right]$
B7	$S_0 \geq 0$	$\tilde{\rho}_0 \geq 0$
B8	$S_1 \geq 0$	$\tilde{\rho}_1 \geq 0$
B9	$D_0 \geq S_0$	$\tilde{\tau}_0 + \tilde{\rho}_0 \leq 1$
B10	$D_1(D_0^2 - S_0^2) \geq \frac{(1+\tau_1)^2}{2\tau_1}S_1^2[D_0$- $-S_0\cos 4(\Psi_0 - \Psi_1)]$	$(\tau_1 - \tilde{\tau}_1)[(1 - \tilde{\tau}_0)^2 - \tilde{\rho}_0^2] \geq 4\tilde{\rho}_1^2[1 - \tilde{\tau}_0 + \tilde{\rho}_0\cos 4(\widetilde{\Phi}_0 - \widetilde{\Phi}_1)]$

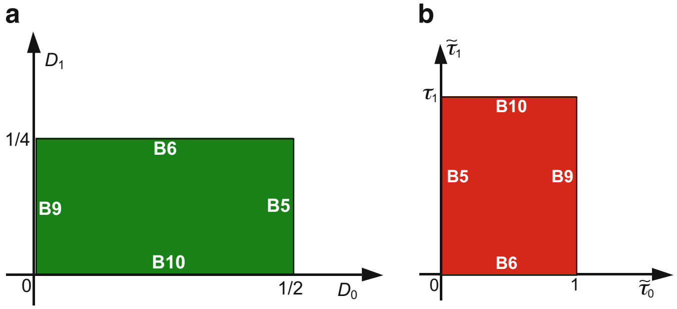


Fig. 2 Admissible domain for the case of isotropic damaged material. (**a**) $\mathbb{D}$. (**b**) $\widetilde{\mathbb{E}}$

- $\widetilde{\mathbb{E}}$ isotropic:

$$\left\{\begin{array}{l} 0 \le D_0 < \frac{1}{2}, \\ 0 \le D_1 < \frac{1}{4}, \end{array}\right. \tag{91}$$

$$\left\{\begin{array}{l} 0 < \tilde{\tau}_0 \le 1, \\ 0 < \tilde{\tau}_1 \le \tau_1, \end{array}\right. \quad \rightarrow \quad \left\{\begin{array}{l} 0 < \widetilde{T}_0 \le T_0, \\ 0 < \widetilde{T}_1 \le T_1. \end{array}\right. \tag{92}$$

The isotropic part of $\mathbb{E}$ is hence diminished by damage; the admissible domain is clearly convex, see Fig. 2.

- $\widetilde{\mathbb{E}}$ square symmetric:

$$\left\{\begin{array}{l} 0 \le D_1 < \frac{1}{4}, \\ 0 \le S_0 < \min\left\{D_0; \frac{1}{2} - D_0\right\}, \end{array}\right. \tag{93}$$

$$\left\{\begin{array}{l} 0 < \tilde{\tau}_1 \le \tau_1, \\ 0 \le \tilde{\rho}_0 < \min\left\{\tilde{\tau}_0; 1 - \tilde{\tau}_0\right\}, \end{array}\right. \quad \rightarrow$$

$$\left\{\begin{array}{l} 0 < \widetilde{T}_0 \le T_0, \\ 0 < \widetilde{T}_1 \le T_1, \\ 0 \le \widetilde{R}_0 < \min\left\{\widetilde{T}_0; T_0 - \widetilde{T}_0\right\}. \end{array}\right. \tag{94}$$

Also for square symmetry, the isotropic part of $\mathbb{E}$ is decreased by damage, but at the same time the anisotropic part, here represented by the only term $\widetilde{R}_0$, grows from zero: damage produces hence a decrease of the averaged stiffness, the isotropic part, but at the same time an increase of the anisotropic part. The admissible domain for the invariants of $\mathbb{D}$ and $\widetilde{\mathbb{E}}$, bounded by linear conditions, is convex, see Fig. 3.

- $\widetilde{\mathbb{E}}$ R_0-orthotropic:

$$\left\{\begin{array}{l} 0 \le D_0 < \frac{1}{2}, \\ 0 \le S_1 < \min\left\{\frac{\sqrt{\tau_1(1-2D_0)(1-4D_1)}}{2(1+\tau_1)}; \frac{\sqrt{2\tau_1 D_0 D_1}}{1+\tau_1}\right\}, \end{array}\right. \tag{95}$$

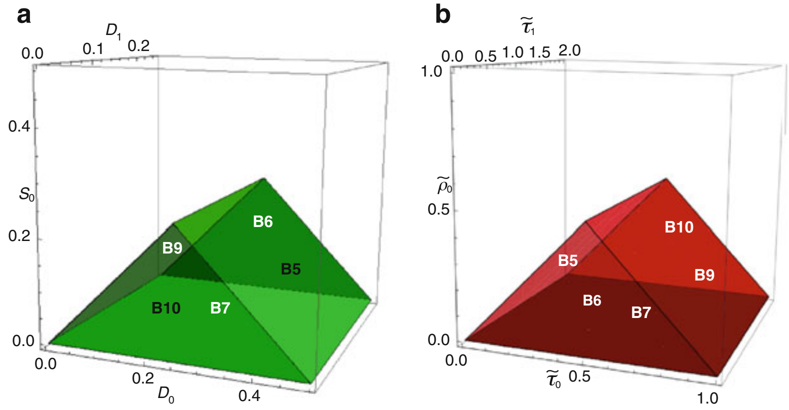


Fig. 3 Admissible domain for the case of square symmetric damaged material; material with $\tau_1 = 2$; the bounds indicated by *white letters* are placed behind and seen in transparency. (**a**) $\mathbb{D}$. (**b**) $\widetilde{\mathbb{E}}$

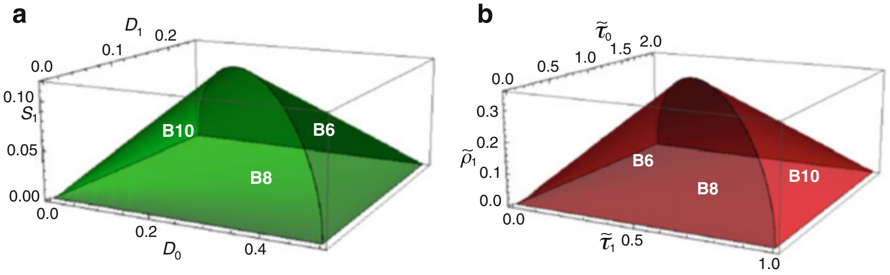


Fig. 4 Admissible domain for the case of R_0-orthotropic damaged material; material with $\tau_1 = 2$; the bounds indicated by *white letters* are placed behind and seen in transparency. (**a**) $\mathbb{D}$. (**b**) $\widetilde{\mathbb{E}}$

$$\begin{cases} 0<\tilde{\tau}_0\le 1,\\ 0\le\tilde{\rho}_1<\min\left\{\frac{\sqrt{\tilde{\tau}_0\tilde{\tau}_1}}{2};\frac{\sqrt{(1-\tilde{\tau}_0)(\tau_1-\tilde{\tau}_1)}}{2}\right\},\end{cases} \rightarrow$$

$$\begin{cases}0<\widetilde{T}_0\le T_0,\\ 0<\widetilde{T}_1\le T_1,\\ 0\le\widetilde{R}_1<\min\left\{\sqrt{\frac{\widetilde{T}_0\widetilde{T}_1}{2}};\sqrt{\frac{(T_0-\widetilde{T}_0)(T_1-\widetilde{T}_1)}{2}}\right\}.\end{cases} \tag{96}$$

In this case too the isotropic part of $\mathbb{E}$ is decreased by damage and the anisotropic part, the term $\widetilde{R}_1$ alone, grows from zero. The above bounds define a convex admissible domain for the invariants of $\mathbb{D}$ and $\widetilde{\mathbb{E}}$, see Fig. 4.

- $\widetilde{\mathbb{E}}$ ordinarily orthotropic:

$$\begin{cases} 0\le S_0<\min\left\{D_0;\frac{1}{2}-D_0\right\},\\ 0\le S_1<\min\left\{\frac{\sqrt{2\tau_1 D_1[D_0+(-1)^L S_0]}}{1+\tau_1};\frac{\sqrt{\tau_1(1-4D_1)[1-2D_0-(-1)^L S_0]}}{2(1+\tau_1)}\right\},\end{cases} \tag{97}$$

$$\left\{\begin{array}{l} 0 \le \tilde{\rho}_0 < \min\left\{\tilde{\tau}_0; 1-\tilde{\tau}_0\right\}, \\ 0 \le \tilde{\rho}_1 < \min\left\{\frac{\sqrt{\tilde{\tau}_1[\tilde{\tau}_0+(-1)^{\widetilde{K}}\tilde{\rho}_0]}}{2}; \frac{\sqrt{[1-\tilde{\tau}_0-(-1)^{\widetilde{K}}\tilde{\rho}_0](\tau_1-\tilde{\tau}_1)}}{2}\right\}, \end{array}\right. \quad \rightarrow$$

$$\left\{\begin{array}{l} 0 < \widetilde{T}_0 \le T_0, \\ 0 < \widetilde{T}_1 \le T_1, \\ 0 \le \widetilde{R}_0 < \min\left\{\widetilde{T}_0; T_0-\widetilde{T}_0\right\}, \\ 0 \le \widetilde{R}_1 < \min\left\{\sqrt{\frac{\widetilde{T}_1[\widetilde{T}_0+(-1)^{\widetilde{K}}\widetilde{R}_0]}{2}}; \sqrt{\frac{[T_0-\widetilde{T}_0-(-1)^{\widetilde{K}}\widetilde{R}_0](T_1-\widetilde{T}_1)}{2}}\right\}. \end{array}\right. \tag{98}$$

Unfortunately, it is not possible to give a graphical representation of the domains defined by the above bounds, because they are functions of four independent quantities. Nevertheless, these domains are convex; in fact, the conditions in Eq. (97) or (98) are either linear or with a Hessian matrix whose eigenvalues are either null or negative, as it can be easily checked. So, such functions are concave and their epigraph convex. The admissible domain, intersection of convex sets, is hence convex. Also in this case, the isotropic part is decreased by damage, while the anisotropic one is increased.

- $\widetilde{\mathbb{S}r_0}$-orthotropic: in this case $\widetilde{\mathbb{E}}$ is orthotropic but depending upon only three non-null invariants thanks to Eq. (24), which inserted into Eqs. (86), (89), and (90) gives

$$S_0 = \frac{(1+\tau_1)^2}{\tau_1}\frac{S_1^2}{1-4D_1}, \quad \tilde{\rho}_0 = 2\frac{\tilde{\rho}_1^2}{\tilde{\tau}_1} = 2\frac{(1+\tau_1)^2}{\tau_1}\frac{S_1^2}{1-4D_1}, \quad L = 1. \tag{99}$$

If an isotropic layer is damaged according to the previous conditions, it is transformed into a material whose behavior is of the same type of that of a sheet of paper, [35]. The bounds in this case are

$$\left\{\begin{array}{l} 0 \le D_0 < \frac{1}{2}, \\ 0 \le D_1 < \frac{1}{4}, \\ 0 \le S_1 < \min\left\{\frac{1}{1+\tau_1}\sqrt{\frac{\tau_1(1-2D_0)(1-4D_1)}{3}}; \frac{1}{1+\tau_1}\sqrt{\frac{2\tau_1 D_0 D_1(1-4D_1)}{1-2D_1}}\right\}. \end{array}\right. \tag{100}$$

$$\left\{\begin{array}{l} 0 < \tilde{\tau}_0 \le 1, \\ 0 < \tilde{\tau}_1 \le \tau_1, \\ 0 \le \tilde{\rho}_1 < \min\left\{\sqrt{\frac{\tilde{\tau}_0\tilde{\tau}_1}{2}}; \sqrt{\frac{\tilde{\tau}_1(1-\tilde{\tau}_0)(\tau_1-\tilde{\tau}_1)}{2(1+\tau_1)}}\right\}, \end{array}\right. \quad \rightarrow$$

$$\left\{\begin{array}{l} 0 < \tilde{T}_0 \le T_0, \\ 0 < \tilde{T}_1 \le T_1, \\ 0 \le \tilde{R}_1 < \min\left\{\sqrt{\tilde{T}_0\tilde{T}_1}; \sqrt{\frac{2\tilde{T}_1(T_0-\tilde{T}_0)(T_1-\tilde{T}_1)}{T_0+2T_1}}\right\}. \end{array}\right. \tag{101}$$

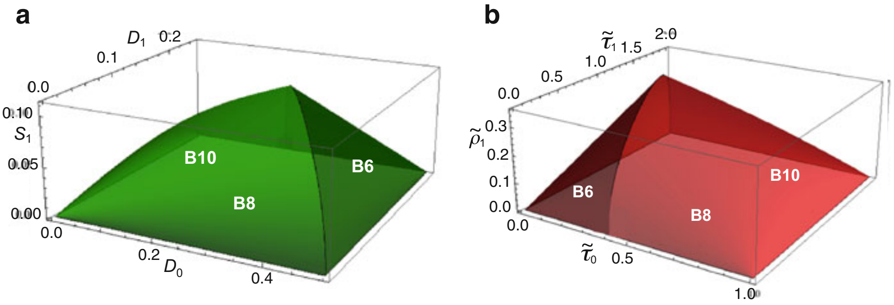


Fig. 5 Admissible domain for the case of r_0-orthotropic damaged material; material with $\tau_1 = 2$; the bounds indicated by *white letters* are placed behind and seen in transparency. (**a**) $\mathbb{D}$. (**b**) $\widetilde{\mathbb{E}}$

We remark that Fig. 5 clearly confirms that the admissible domain is convex, as it has to be, being this one a particular case of ordinary orthotropy. Once more, the isotropic part decreases and the anisotropic one increases as a consequence of damage.

References

1. Benvenuto, E.: An Introduction to the History of Structural Mechanics, 2 vols. Springer, Berlin (1991)
2. Born, M.: Dynamik der Krystallgitter. Teubner, Leipzig (1915)
3. Boscovich, R.G.: Theoria Philosophiae Naturalis Redacta ad Unicam Legem Virium in Natura Existentium. Officina Libraria Kaliwodiana, Vienna (1758)
4. Campbell, J.G.: The in-plane elastic constants of paper. Aust. J. Appl. Sci. **12**, 356–357 (1961)
5. Capecchi, D., Ruta, G., Trovalusci, P.: Voigt and Poincaré's mechanistic–energetic approaches to linear elasticity and suggestions for multiscale modelling. Arch. Appl. Mech. **81**, 1573–1584 (2011)
6. Cauchy, A.L.: De la pression ou tension dans un système de points matériels. Exerc. Math. **2**, 213–236 (1828)
7. Cauchy, A.L.: Sur l'équilibre et le mouvement d'un système de points matériels sollicités par des forces d'attraction ou de repulsion mutuelle. Exerc. Math. **3**, 188–212 (1828)
8. Chaboche, J.L.: Le concept de contrainte effective appliqué à l'elasticité et à la viscoplasticité en présence d'un endommagement anisotrope. In: Proceedings of Colloque Euromech 115 (Villard-de-Lans, 1979): Comportement Mécanique des Matériaux Anisotropes, Paris, pp. 737–760 (1982)
9. Chow, C.L.: On evolution laws of anisotropic damage. Eng. Fract. Mech. **34**, 679–701 (1987)
10. de Saint-Venant, A.B.: Sur la question de savoir s'il existe des masses continues et sur la nature probable des dernières particules des corps. Société Philomatique de Paris 3–15 (1844)
11. Doyle, T.C., Ericksen, J.L.: Nonlinear elasticity. Adv. Appl. Mech. **4**, 53–115 (1956)
12. Dugas, R.: Histoire de la Mécanique. Editions du Griffon, Neuchâtel (1950)
13. Green, G.: On the laws of reflexion and refraction of light at the common surface of two non-crystallized media. Camb. Philos. Soc. Trans. **7**, 1–24 (1839)

14. Hono, M., Onogi, S.: Dynamic measurements of physical properties of pulp and paper by audiofrequency sound. J. Appl. Phys. **22**, 971–977 (1951)
15. Jones, R.M.: Mechanics of Composite Materials, 2nd edn. Taylor & Francis, London (1999)
16. Leckie, F.A., Onat, E.T.: Tensorial nature of damage measuring internal variables. In: Hult, J., Lemaitre, J. (eds). Proceedings of IUTAM Colloquium "Physical Non-linearities in Structural Mechanics", Senlis, pp. 140–155 (1980)
17. Lemaitre, J., Chaboche, J. L., Benallal, A., Desmorat, R.: Mécanique des Matériaux Solides. Dunod, Paris (2009)
18. Love, A.E.H.: A Treatise on the Mathematical Theory of Elasticity. Dover, New York (1944)
19. Muskhelishvili, N.I.: Some Basic Problems of the Mathematical Theory of Elasticity. Noordhoff, Groningen (1953)
20. Navier, L.: Mémoire sur les lois de l'équilibre et du mouvement des solides élastiques. Mémoires de l'Académie Royale des Sciences de l'Institut National **7**, 375–393 (1827)
21. Newton, I.: Philosophiae Naturalis Principia Mathematica. J. Streater, London (1687)
22. Ostoja-Starzewski, M.: Microstructural Randomness and Scaling in Mechanics of Materials. Chapmann and Hall/CRC, New York (2007)
23. Poincaré, H.: Leçons sur la Théorie de l'élasticité. Carré, Paris (1892)
24. Poisson, S.D.: Traité de Mécanique. Courcier, Paris (1811)
25. Poisson, S.D.: Mémoire sur les surfaces élastiques. Mémoires de l'Académie Royale des Sciences de l'Institut National **8**, 167–225 (1814)
26. Poisson, S.D.: Mémoire sur l'équiibre et le mouvement des corps élastiques. Mémoires de l'Académie Royale des Sciences de l'Institut National **8**, 357–570 (1829)
27. Sidoroff, F.: Description of anisotropic damage. Application to elasticity. In: Hult, J., Lemaitre, J. (eds.) Proceedings of IUTAM Colloquium "Physical Non-linearities in Structural Mechanics", Senlis, pp. 237–244 (1980)
28. Stackgold, I.: The cauchy relations in a molecular theory of elasticity. Q. Appl. Math. **8**, 169–186 (1950)
29. Todhunter, I., Pearson, K.: History of the Theory of Elasticity, vol. 1. Cambridge University Press, Cambridge (1886)
30. Valot, E., Vannucci, P.: Some exact solutions for fully orthotropic laminates. Compos. Struct. **69**, 157–166 (2005)
31. Vannucci, P.: A special planar orthotropic material. J. Elast. **67**, 81–96 (2002)
32. Vannucci, P.: HDR thesis, University of Burgundy (2002)
33. Vannucci, P.: Plane anisotropy by the polar method. Meccanica **40**, 437–454 (2005)
34. Vannucci, P.: Influence of invariant material parameters on the flexural optimal design of thin anisotropic laminates. Int. J. Mech. Sci. **51**, 192–203 (2009)
35. Vannucci, P.: On special orthotropy of paper. J. Elast. **99**, 75–83 (2010)
36. Vannucci, P.: Strange laminates. Math. Methods Appl. Sci. **35**, 1532–1546 (2012)
37. Vannucci, P., Desmorat, B.: Analytical bounds for damage induced planar anisotropy. Int. J. Solids Struct. **60–61**, 96–106 (2015)
38. Vannucci, P., Desmorat, B.: Plane anisotropic rari-constant materials. Math. Methods Appl. Sci. **39**, 3271–3281 (2016) doi:10.1002/mma.3770
39. Vannucci, P., Verchery, G.: A new method for generating fully isotropic laminates. Compos. Struct. **58**, 75–82 (2002)
40. Vannucci, P., Verchery, G.: Anisotropy of plane complex elastic bodies. Int. J. Solids Struct. **47**, 1154–1166 (2010)
41. Verchery, G.: Les invariants des tenseurs d'ordre 4 du type de l'élasticité. In: Proceedings of Colloque Euromech 115 (Villard-de-Lans, 1979): Comportement Mécanique des Matériaux Anisotropes, pp. 93–104. Editions du CNRS, Paris (1982)
42. Voigt, W.: Lehrbuch der Kristallphysik. B. G. Teubner, Leipzig (1910)