Numer Algor (2018) 78:983–1000
DOI 10.1007/s11075-017-0409-7

ORIGINAL PAPER

Inexactly constrained discrete adjoint approach for steepest descent-based optimization algorithms

David A. Brown[1] · **Siva Nadarajah**[1]

Received: 21 May 2017 / Accepted: 28 August 2017 / Published online: 7 September 2017

Abstract The problem of constrained optimization via the gradient-based discrete adjoint steepest descent method is studied under the assumption that the constraint equations are solved inexactly. Error propagation from the constraint equations to the gradient is studied analytically, as is the convergence rate of the inexactly constrained algorithm as it relates to the exact algorithm. A method is developed for adapting the residual tolerance to which the constraint equations are solved. The adaptive tolerance method is applied to two simple test cases to demonstrate the potential gains in computational efficiency.

Keywords Discrete adjoint · Optimization · Gradient · Steepest descent · Inexact · Constrained optimization · Computational fluid dynamics

1 Introduction

The discrete adjoint method, pioneered in the field of computational aerodynamics by Pironneau [9] and Reuther and Jameson [11], has seen widespread use in the last two decades for numerically finding optimal designs for engineering problems—for example, the shape of an aircraft wing with minimum drag under given operating conditions. This method is popular due to its reliability as well as numerical efficiency. However, there remain two philosophies on how the discrete adjoint method can be most efficiently applied. In the first approach, the adjoint and constraint equations are solved to a relaxed tolerance, reducing the computational effort of each

✉ David A. Brown
davidbrown172@hotmail.com

[1] McGill University, Montreal, QC Canada

optimization iteration [5, 6]. This is typically combined with a steepest descent gradient method. The second approach is to solve both the adjoint and constraint equations accurately [3], allowing for the use of a more rapidly converging iterative procedure such as the BFGS algorithm. Either approach can be effective and there is no clear consensus on which is more efficient [2].

Despite the popularity of the method, to our knowledge, no formal studies have been performed to investigate the relationship between the solver tolerances and algorithm convergence. In our analysis, we model the error as being small and controllable. The objective is to determine how to select the tolerances such that the gradient retains sufficient accuracy but to avoid wasting computational effort in over-solving. The current study is limited to the steepest descent algorithm.

2 Preliminaries

2.1 Fréchet differentiability

Definition 1 (cf. Ortega and Rheinboldt [8], Def. 3.1.5, p. 61) The mapping $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^m$ is Fréchet differentiable at $x \in \mathbb{R}^n$ if there exists a linear operator $\mathcal{A} \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$\lim_{h\to 0} \frac{1}{\|h\|} \|\mathcal{R}(x+h) - \mathcal{R}(x) - \mathcal{A}h\| = 0.$$

The linear operator $\mathcal{A}$ is the Jacobian matrix of $\mathcal{R}$, which we denote $\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{R}(x)$.

2.2 Big-$\mathcal{O}$ notation

Definition 2 Let x_k be a sequence which converges to x^*. Then, for positive-valued continuous scalar mappings $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, we denote

$$f(x_k) = \mathcal{O}(g(x_k)) \quad \text{if} \quad \limsup_{k\to\infty} \frac{f(x_k)}{g(x_k)} < +\infty.$$

Similarly,

$$\begin{aligned} f(x_k) = \mathcal{O}(g(x_k), h(x_k)) \qquad &\text{if} \quad \limsup\nolimits_{k\to\infty} \tfrac{f(x_k)}{g(x_k)} < +\infty \\ &\text{and} \quad \limsup\nolimits_{k\to\infty} \tfrac{f(x_k)}{h(x_k)} < +\infty \end{aligned}$$

2.3 Q-convergence

(cf. Nocedal and Wright [7], Ch. 2.2, pp. 28-29) A sequence $\{x_j\}_{j\geq 0}$ is said to converge to its limit x^* with Q-convergence rate p if there exist $L \in \mathbb{R}$, $L > 0$ and $k \in \mathbb{Z}$, $k \geq 0$ such that

$$\|x_{j+1} - x^*\| < L \|x_j - x^*\|^p \tag{1}$$

for all $j \geq k$.

3 The discrete adjoint approach

Consider the optimization problem:

$$\min_{F} \mathcal{I}\left(w_s, F\right) \tag{2}$$

subject to the constraints

$$\mathcal{R}\left(w_s, F\right) = \mathbf{0}, \tag{3}$$

where $w_s \in \mathbb{R}^n$, $F \in \mathbb{R}^m$, $\mathcal{I} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $(w, F) \mapsto \mathcal{I}\left(w_s, F\right)$, $\mathcal{R} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $(w, F) \mapsto \mathcal{R}\left(w_s, F\right)$. The Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $(w, F) \mapsto \mathcal{L}\left(w, F\right)$ of the system is formed by introducing the Lagrangian multipliers $\psi \in \mathbb{R}^n$:

$$\mathcal{L}\left(w_s, F\right) = \mathcal{I}\left(w_s, F\right) - \psi^T \mathcal{R}\left(w_s, F\right). \tag{4}$$

One approach to deriving the discrete adjoint method is to consider the differential element

$$\delta\mathcal{L}\left(w_s, F\right) = \delta\mathcal{I}\left(w_s, F\right) - \psi^T \delta\mathcal{R}\left(w_s, F\right). \tag{5}$$

Expanding $\delta\mathcal{I}\left(w_s, F\right)$ and $\delta\mathcal{R}\left(w_s, F\right)$ and simplifying:

$$\begin{aligned} \delta\mathcal{L}\left(w_s, F\right) &= \frac{\partial}{\partial w}\mathcal{I}\left(w_s, F\right)\delta w_s + \frac{\partial}{\partial F}\mathcal{I}\left(w_s, F\right)\delta F \\ &\quad -\psi^T\left(\frac{\partial}{\partial w}\mathcal{R}\left(w_s, F\right)\delta w_s + \frac{\partial}{\partial F}\mathcal{R}\left(w_s, F\right)\delta F\right) \\ &= \left(\frac{\partial}{\partial w}\mathcal{I}\left(w_s, F\right) - \psi^T\left[\frac{\partial}{\partial w}\mathcal{R}\left(w_s, F\right)\right]\right)\delta w_s \\ &\quad +\left(\frac{\partial}{\partial F}\mathcal{I}\left(w_s, F\right) - \psi^T\left[\frac{\partial}{\partial F}\mathcal{R}\left(w_s, F\right)\right]\right)\delta F. \end{aligned}$$

Setting the Lagrangian multipliers such that

$$\frac{\partial}{\partial w}\mathcal{I}\left(w_s, F\right) - \psi^T\left[\frac{\partial}{\partial w}\mathcal{R}\left(w_s, F\right)\right] = \mathbf{0} \tag{6}$$

eliminates the first term, resulting in

$$\delta\mathcal{L}\left(w_s, F\right) = \mathcal{G}\delta F, \tag{7}$$

where

$$\mathcal{G}\left(w_s, F\right) = \frac{\partial}{\partial F}\mathcal{I}\left(w_s, F\right) - \psi^T\left[\frac{\partial}{\partial F}\mathcal{R}\left(w_s, F\right)\right], \tag{8}$$

$\mathcal{G} \in \mathbb{R}^m$ is the gradient and is interpreted as a row vector. A basic pseudo-code of the algorithm is shown in Algorithm 1. Additional details and discussion are provided, for example, by Jameson [5].

Algorithm 1 The exact discrete adjoint algorithm

```
while Not converged do
    $w \leftarrow$ Solve $\mathcal{R}(w, F) = \mathbf{0}$ for $w$;
    $\mathcal{I} \leftarrow \mathcal{I}(w, F)$;
    $\frac{\partial}{\partial w}\mathcal{I} \leftarrow \frac{\partial}{\partial w}\mathcal{I}(w, F)$;
    $\mathcal{A} \leftarrow \frac{\partial}{\partial w}\mathcal{R}(w, F)$;
    $\psi \leftarrow \left[\mathcal{A}^T\right]^{-1} \frac{\partial}{\partial w}\mathcal{I}$;
    $\frac{\partial}{\partial F}\mathcal{I} \leftarrow \frac{\partial}{\partial F}\mathcal{I}(w, F)$;
    $\psi^T \frac{\partial}{\partial F}\mathcal{R} \leftarrow \psi^T \frac{\partial}{\partial F}\mathcal{R}(w, F)$;
    $\mathcal{G} \leftarrow \frac{\partial}{\partial F}\mathcal{I} - \psi^T \frac{\partial}{\partial F}\mathcal{R}$;
    $F \leftarrow$ Gradient-based optimization update;
end
```

4 The inexactly constrained discrete adjoint approach

4.1 Relationship between the gradient and the constraint error

From (6) and (8), we see that

$$\mathcal{G}(w_s, F) = \frac{\partial}{\partial F}\mathcal{I}(w_s, F) - \frac{\partial}{\partial w}\mathcal{I}(w_s, F)\left[\frac{\partial}{\partial w}\mathcal{R}(w_s, F)\right]^{-1}\frac{\partial}{\partial F}\mathcal{R}(w_s, F), \quad (9)$$

assuming that $\left[\frac{\partial}{\partial w}\mathcal{R}(w_s, F)\right]^{-1}$ exists. Assuming that $\frac{\partial}{\partial w}\mathcal{R}(w_s, F)$ and its inverse are smooth with respect to w, the gradient can be found from a nearby Taylor expansion:

$$\mathcal{G}(w_s, F) = \mathcal{G}(w_s + \epsilon_w, F) - \frac{\partial}{\partial w}\mathcal{G}(w_s + \epsilon_w, F)\,\epsilon_w + \mathcal{O}\left(\|\epsilon_w\|^2\right). \quad (10)$$

The quantity $\mathcal{G}(w_s + \epsilon_w, F)$, henceforth denoted $\mathcal{G}_a$, represents the “actual” gradient, evaluated based on the inexact solution to the constraint equations. The quantity $\epsilon_w \in \mathbb{R}^n$ is the error in the constraint equation variables. We also introduce $\epsilon_{\mathcal{G}} \in \mathbb{R}^m$, the resulting error in the gradient. These three quantities take the following formal definitions:

$$\epsilon_w \equiv w - w_s, \quad (11)$$

$$\mathcal{G}_a \equiv \mathcal{G}(w, F) = \mathcal{G}(w_s + \epsilon_w, F), \quad (12)$$

$$\epsilon_{\mathcal{G}} \equiv \mathcal{G}_a - \mathcal{G}, \quad (13)$$

where we have abbreviated $\mathcal{G} \equiv \mathcal{G}(w_s, F)$, and will use such abbreviations for $\mathcal{I}$ and $\mathcal{R}$ as well. Hence, (10) can be written as

$$\mathcal{G} = \mathcal{G}_a - \frac{\partial \mathcal{G}_a}{\partial w}\epsilon_w + \mathcal{O}\left(\|\epsilon_w\|^2\right), \quad (14)$$

or, more compactly, as

$$\epsilon_{\mathcal{G}} = \frac{\partial \mathcal{G}_a}{\partial w} \epsilon_w + \mathcal{O}\left(\|\epsilon_w\|^2\right). \tag{15}$$

4.2 Q-convergence rate

In this section, we establish that it is possible to achieve the same order of Q-convergence with an inexactly constrained gradient algorithm as it is with the exact algorithm for suitable choice of $\left\|\epsilon_{w,j}\right\|$, where j is the iteration index. Gradient-based optimization algorithms are typically of the form

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_j, \tag{16}$$

where $\gamma_j > 0$, $\gamma \in \mathbb{R}$ is the step size, and $\mathcal{D}_j$ is a square matrix, commonly assumed to be positive definite. In the case of the steepest descent algorithm, $\mathcal{D}_j$ is the identity matrix. However, a higher convergence rate can often be achieved by setting $\mathcal{D}_j$ to an approximation to the inverse Hessian matrix [1].

Both γ_j and $\mathcal{D}_j$ can vary with iteration index j and can thus be interpreted as a sequence. The analysis in this paper is limited to the case where $\mathcal{D}_j$ and γ_j are independent of ϵ_w as doing otherwise would require consideration of specific algorithms which is left as future work.

Theorem 1 *Assume that $\mathcal{I}$ and $\mathcal{R}$ are smooth in the Fréchet sense with respect to both w and F and assume that the sequence generated by*

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_j, \tag{17}$$

where $\mathcal{G}_j \equiv \mathcal{G}\left(w_{s,j}, F_j\right)$, converges to a local minimizer F^*. Furthermore, assume that there exists some fixed $L > 0$ and $p \geq 1$ such that

$$\left\|F_{j+1} - F^*\right\| < L \left\|F_j - F^*\right\|^p \tag{18}$$

for all $j \geq k$ for all F_k in some ball $\mathcal{B}_r\left(F^*\right)$ of radius $r > 0$ centred at F^*. Then there exists a sequence δ_j and fixed $L' > 0$ such that the algorithm

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_{a,j}, \tag{19}$$

where $\mathcal{G}_{a,j} \equiv \mathcal{G}_a\left(w_j, F_j\right)$, converges according to

$$\left\|F_{j+1} - F^*\right\| < L' \left\|F_j - F^*\right\|^p \tag{20}$$

for all $0 < \left\|\epsilon_{w,j}\right\| < \delta_j$. Furthermore, for any fixed $\epsilon > 0$, there exists a sequence δ_j such that $L' < L + \epsilon$.

Proof
Without loss of generality, consider some $F_j \in \mathcal{B}_r\left(F^*\right)$ to which an iterative step of either the exact or inexact gradient algorithm could be applied. The inexact update is

given by

$$\begin{aligned} F_{j+1} &= F_j - \gamma_j \mathcal{D}_j \mathcal{G}_{a,j} \\ F_{j+1} - F^* &= F_j - F^* - \gamma_j \mathcal{D}_j \mathcal{G}_j - \gamma_j \mathcal{D}_j \frac{\partial}{\partial w} \mathcal{G}_{a,j} \epsilon_{w,j} + \gamma_j \mathcal{D}_j \mathcal{O}\left(\left\| \epsilon_{w,j} \right\|^2 \right), \\ \left\| F_{j+1} - F^* \right\| &\le \left\| F_j - F^* - \gamma_j \mathcal{D}_j \mathcal{G}_j \right\| + \gamma_j \left\| \mathcal{D}_j \frac{\partial}{\partial w} \mathcal{G}_{a,j} \epsilon_{w,j} \right\| \\ &\quad + \gamma_j \left\| \mathcal{D}_j \mathcal{O}\left(\left\| \epsilon_{w,j} \right\|^2 \right) \right\| \end{aligned} \tag{21}$$

where we have subtracted F^* from both sides and used (14). Since $\mathcal{G}_a$ is Fréchet differentiable, then for every $\epsilon_{1,j} > 0$ there exists $\delta_{1,j} > 0$ such that

$$\left\| \frac{\partial}{\partial w} \mathcal{G}_{a,j} \epsilon_{w,j} \right\| \le \left\| \mathcal{G}_j - \mathcal{G}_{a,j} \right\| + \epsilon_{1,j} \tag{22}$$

for all $\left\| \epsilon_{w,j} \right\| < \delta_{1,j}$. Since Fréchet differentiability also implies Lipschitz continuity [8], it also follows that for every $\epsilon_{2,j} > 0$ there exists $\delta_{2,j} > 0$ such that

$$\left\| \mathcal{G}_j - \mathcal{G}_{a,j} \right\| \le \epsilon_{2,j} \tag{23}$$

for all $\left\| \epsilon_{w,j} \right\| < \delta_{2,j}$. Furthermore, let the sequence $\epsilon_{3,j}$ be an upper bound on the $\mathcal{O}\left(\left\| \epsilon_{w,j} \right\|^2 \right)$ terms. Note that it is possible to construct a sequence $\delta_{3,j}$ such that the $\mathcal{O}\left(\left\| \epsilon_{w,j} \right\|^2 \right)$ terms are bounded above by any sequence $\epsilon_{3,j} > 0$ for all $\left\| \epsilon_{w,j} \right\| < \delta_{3,j}$. These three inequalities lead to the following chain of inequalities:

$$\begin{aligned} \left\| \mathcal{D}_j \frac{\partial}{\partial w} \mathcal{G}_{a,j} \epsilon_{w,j} \right\| + \left\| \mathcal{D}_j \mathcal{O}\left(\left\| \epsilon_{w,j} \right\|^2 \right) \right\| &\le \left\| \mathcal{D}_j \right\| \left\| \frac{\partial}{\partial w} \mathcal{G}_{a,j} \epsilon_{w,j} \right\| + \left\| \mathcal{D}_j \right\| \mathcal{O}\left(\left\| \epsilon_{w,j} \right\|^2 \right) \\ &\le \left\| \mathcal{D}_j \right\| \left\| \mathcal{G}_j - \mathcal{G}_{a,j} \right\| + \left\| \mathcal{D}_j \right\| \epsilon_{1,j} + \left\| \mathcal{D}_j \right\| \epsilon_{3,j} \\ &\le \left\| \mathcal{D}_j \right\| \left(\epsilon_{2,j} + \epsilon_{1,j} + \epsilon_{3,j} \right). \end{aligned} \tag{24}$$

Using (24) and (18) with (21) gives

$$\left\| F_{j+1} - F^* \right\| \le L \left\| F_j - F^* \right\|^p + \gamma_j \left\| \mathcal{D}_j \right\| \left(\epsilon_{1,j} + \epsilon_{2,j} + \epsilon_{3,j} \right), \tag{25}$$

where, in addition to (22), we have invoked the convergence rate of the exact algorithm as given by (18).

Hence, for any fixed ϵ, choosing $\delta_{1,j} = \delta_{2,j} = \delta_{3,j} = \delta_j$ such that

$$\epsilon_{1,j} + \epsilon_{2,j} + \epsilon_{3,j} < \frac{\epsilon}{\left\| \mathcal{D}_j \right\| \gamma_j} \left\| F_j - F^* \right\|^p \tag{26}$$

for $\left\| \epsilon_{w,j} \right\| < \delta_j$ will ensure that the convergence rate (20) is achieved with $L' < L + \epsilon$. □

Thus, we see that it is possible to maintain the convergence rate of the theoretical algorithm with the inexactly constrained algorithm and that the convergence-related constant L is even recovered in the limit of δ_j, the upper bound on $\left\| \epsilon_{w,j} \right\|$, tending to zero.

It is apparent from the proof of Theorem 1 that it will be necessary to adapt δ_j based on at least one of γ_j, $\left\| F_j - F^* \right\|$, or $\left\| \mathcal{G}_j - \mathcal{G}_{a,j} \right\|$. We now show that the key lies in maintaining the relative gradient error below a certain threshold.

Theorem 2 *Assume that the sequence generated by*

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_j, \tag{27}$$

where $\mathcal{G}_j \equiv \mathcal{G}\left(w_{s,j}, F_j\right)$, converges to a local minimizer F^* in the sense that there exists some fixed $L > 0$ and $p \geq 1$ such that

$$\left\| F_{j+1} - F^* \right\| < L \left\| F_j - F^* \right\|^p \tag{28}$$

for all $j \geq k$ for all F_k in some ball $\mathcal{B}_r\left(F^*\right)$ of radius $r > 0$ centred at F^*. Then there exists some fixed $C > 0$ such that the algorithm

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_{a,j}, \tag{29}$$

where $\mathcal{G}_{a,j} \equiv \mathcal{G}_a\left(w_j, F_j\right)$, converges according to

$$\left\| F_{j+1} - F^* \right\| < L' \left\| F_j - F^* \right\|^p \tag{30}$$

so long as

$$\left\| \mathcal{D}_j \epsilon_{\mathcal{G},j} \right\| < C \left\| \mathcal{D}_j \mathcal{G}_j \right\|^p. \tag{31}$$

Furthermore, for any fixed $\epsilon > 0$, there exists $C > 0$ such that $L' < L + \epsilon$. Moreover, such a C satisfies

$$C \leq \frac{\epsilon}{\left(L r^{p-1} + 1\right)^p}. \tag{32}$$

Proof As before, consider some $F_j \in \mathcal{B}_r\left(F^*\right)$ to which an iterative step of either the exact or inexact gradient algorithm could be applied. The inexact update is given by

$$\begin{aligned} F_{j+1} &= F_j - \gamma_j \mathcal{D}_j \mathcal{G}_{a,j} \\ F_{j+1} - F^* &= F_j - F^* - \gamma_j \mathcal{D}_j \mathcal{G}_j - \gamma_j \mathcal{D}_j \epsilon_{\mathcal{G}}, \\ \left\| F_{j+1} - F^* \right\| &\leq \left\| F_j - F^* - \gamma_j \mathcal{D}_j \mathcal{G}_j \right\| + \left\| \gamma_j \mathcal{D}_j \epsilon_{\mathcal{G},j} \right\|. \end{aligned} \tag{33}$$

Invoking (28) and (31):

$$\left\| F_{j+1} - F^* \right\| \leq L \left\| F_j - F^* \right\|^p + C \left\| \gamma_j \mathcal{D}_j \mathcal{G}_j \right\|^p. \tag{34}$$

To proceed, we use (27) and (28) to develop the following inequality:

$$\begin{aligned} \left\| \gamma_j \mathcal{D}_j \mathcal{G}_j \right\| &= \left\| F_{j+1} - F_j \right\| \\ &= \left\| F_{j+1} - F^* - F_j + F^* \right\| \\ &\leq \left\| F_{j+1} - F^* \right\| + \left\| F_j - F^* \right\| \\ &\leq L \left\| F_j - F^* \right\|^p + \left\| F_j - F^* \right\| \\ &= \left\| F_j - F^* \right\| \left(L \left\| F_j - F^* \right\|^{p-1} + 1 \right) \\ &\leq \left\| F_j - F^* \right\| \left(L r^{p-1} + 1 \right). \end{aligned} \tag{35}$$

Using (35) with (34):

$$\begin{aligned}\left\| F_{j+1} - F^{*} \right\| &\leq L \left\| F_j - F^{*} \right\|^{p} + C \left(L r^{p-1} + 1 \right)^{p} \left\| F_j - F^{*} \right\|^{p} \\ &= \left[L + C \left(L r^{p-1} + 1 \right)^{p} \right] \left\| F_j - F^{*} \right\|^{p} . \end{aligned} \tag{36}$$

For any $\epsilon > 0$, choosing C according to the inequality (32) completes the proof. □

It is now a simple matter to extend the result to ϵ_w.

Theorem 3 *Assume that the sequence generated by*

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_j , \tag{37}$$

where $\mathcal{G}_j \equiv \mathcal{G}\left(w_{s,j}, F_j\right)$, converges to a local minimizer F^{*} in the sense that there exists some fixed $L > 0$ and $p \geq 1$ such that

$$\left\| F_{j+1} - F^{*} \right\| < L \left\| F_j - F^{*} \right\|^{p} \tag{38}$$

for all $j \geq k$ for all F_k in some ball $\mathcal{B}_r\left(F^{*}\right)$ of radius $r > 0$ centred at F^{*}. Assume furthermore that $\frac{\partial}{\partial w}\mathcal{G}_a$ is bounded on $\mathcal{B}_r\left(F^{*}\right)$ such that

$$\left\| \frac{\partial}{\partial w} \mathcal{G}_a \right\| < C'_{\epsilon_w} . \tag{39}$$

Then there exists some fixed $C_{\epsilon_w} > 0$ such that the algorithm

$$F_{j+1} = F_j - \gamma_j \mathcal{D}_j \mathcal{G}_{a,j} , \tag{40}$$

where $\mathcal{G}_{a,j} \equiv \mathcal{G}_a\left(w_j, F_j\right)$, converges according to

$$\left\| F_{j+1} - F^{*} \right\| < L' \left\| F_j - F^{*} \right\|^{p} \tag{41}$$

so long as

$$\left\| \epsilon_{w,j} \right\| < C_{\epsilon_w} \left\| \mathcal{G}_j \right\|^{p} . \tag{42}$$

Proof From (15),

$$\left\| \epsilon_{\mathcal{G}} \right\| \leq \left\| \frac{\partial \mathcal{G}_a}{\partial w} \epsilon_w \right\| + \mathcal{O}\left(\left\| \epsilon_w \right\|^{2} \right) . \tag{43}$$

Consider C''_{ϵ_w} such that the $\mathcal{O}\left(\left\|\epsilon_w\right\|^2\right)$ terms above are bounded by $C''_{\epsilon_w} \left\|\epsilon_w\right\|$ on $\mathcal{B}_r\left(F^{*}\right)$. Then

$$\left\| \epsilon_{\mathcal{G}} \right\| \leq \left\| \frac{\partial \mathcal{G}_a}{\partial w} \right\| \left\| \epsilon_w \right\| + C''_{\epsilon_w} \left\| \epsilon_w \right\| . \tag{44}$$

Considering inequality (39), we get

$$\left\| \epsilon_{\mathcal{G}} \right\| \leq C'_{\epsilon_w} \left\| \epsilon_w \right\| + C''_{\epsilon_w} \left\| \epsilon_w \right\| . \tag{45}$$

Thus, choosing $\left\|\epsilon_w\right\|$ such that

$$\left\| \epsilon_w \right\| < C_{\epsilon_w} \left\| \mathcal{G}_a \right\|^{p} , \tag{46}$$

where

$$C_{\epsilon_w} = \frac{C}{C'_{\epsilon_w} + C''_{\epsilon_w}}, \tag{47}$$

ensures that $\left\|\epsilon_{\mathcal{G}}\right\| < C \left\|\mathcal{G}_a\right\|^p$, and Q-order p convergence follows from Theorem 2 with the same value of C. □

4.3 Adaptive constraint tolerance algorithm for the steepest descent method

Based on Theorem 3, we can obtain the design order of convergence by adapting $\|\epsilon_w\|$ based on $\|\epsilon_w\| < \|\epsilon_w\|_{\text{tar}}$ for sufficiently small constant C_{ϵ_w}, with

$$\|\epsilon_w\|_{\text{tar}} = C_{\epsilon_w} \left\|\mathcal{G}_a\right\|^p, \tag{48}$$

with C_{ϵ_w} treated as a user parameter. However, selection of C_{ϵ_w} is not intuitive and is most likely case-dependent, especially considering that ϵ_w and $\mathcal{G}$ can be in different units. This makes it impossible for the user to know how to select C_{ϵ_w} for a given test case.

To overcome this deficiency, the constant C_{ϵ_w} is replaced by

$$C \to \frac{\eta}{S_k}, \tag{49}$$

where $\eta \in \mathbb{R}$ is a user-defined constant parameter and

$$S_k = \max_{0 \le k \le j} \frac{\left\|\mathcal{G}_a\left(w_k, F_k\right) - \mathcal{G}_a\left(w_{k-1}, F_{k-1}\right)\right\|}{\left\|\Delta w_{k-1}\right\|} \tag{50}$$

is a sort of “unit conversion” from $\mathcal{G}$ to w. Though C is no longer a constant, it is clearly upper bounded by η/S_0, and hence, Theorem 3 still ensures convergence for small enough η.

To rationalize the inclusion of the S_k term, we consider (47). The parameter C''_{ϵ_w} is a bound on the second-order terms in the Taylor expansion of the gradient and cannot easily be estimated. C'_{ϵ_w} however must satisfy

$$C'_{\epsilon_w} > \max_{F \in \mathcal{B}_r(F^*)} \left\| \frac{\partial \mathcal{G}_a}{\partial w} \right\| = \max_{F \in \mathcal{B}_r(F^*)} \max_{v} \frac{\left\| \frac{\partial \mathcal{G}_a}{\partial w} v \right\|}{\|v\|}. \tag{51}$$

Normally, the matrix $\frac{\partial}{\partial w}\mathcal{G}_a$ cannot easily be formed nor its norm estimated. Under such circumstances, we use a reference value

$$C'_{\epsilon_w} \sim \max_{0 \le k \le j} \frac{\left\| \frac{\partial \mathcal{G}_{a,k}}{\partial w} \Delta w_{k-1} \right\|}{\left\|\Delta w_{k-1}\right\|}, \tag{52}$$

where $\Delta w_{k-1} \equiv w_k - w_{k-1}$. Using this vector, the Fréchet approximation to this matrix-vector product

$$\frac{\partial \mathcal{G}_{a,k}}{\partial w} \Delta w_{k-1} \approx \mathcal{G}_a\left(w_k, F_k\right) - \mathcal{G}_a\left(w_{k-1}, F_k\right) \tag{53}$$

involves quantities which are known or calculable.

The final step is to replace $\mathcal{G}_a(w_{k-1}, F_k)$ with $\mathcal{G}_a(w_{k-1}, F_{k-1})$ in the above expression, since the former is expensive to compute whereas the latter is already available. This is reasonable for many cases of interest since in many cases, the direct dependence of the cost functional on F is very weak. For example, considering the problem of finding the drag-minimizing wing shape of an aircraft, the drag coefficient is expected to be far more sensitive to the changes in the flow field variables w resulting from changes to the shape than it is directly sensitive to the shape change. Hence, the following expression might be used for the target constraint error:

$$\|\epsilon_w\|_{\text{tar},j+1} = \eta \frac{\left\|\mathcal{G}_a\left(w_j, F_j\right)\right\|^p \left\|\Delta w_{j-1}\right\|}{\left\|\mathcal{G}_a\left(w_j, F_j\right) - \mathcal{G}_a\left(w_{j-1}, F_{j-1}\right)\right\|}. \tag{54}$$

Considering (47) and (31), η might be on the order of the relative gradient error.

The inexactly constrained discrete adjoint algorithm with adaptive tolerance as described in this section is shown as Algorithm 2.

Algorithm 2 The inexactly constrained discrete adjoint algorithm with constraint tolerance adaptation based on (19). The boxes indicate additions to the original algorithm

```
$\boxed{S \leftarrow 0;}$
while Not converged do
    $\boxed{w_{\text{prev}} \leftarrow w;}$
    $w \leftarrow$ Solve $\mathcal{R}(w, F) = \mathbf{0}$
    $\boxed{\text{inexactly such that } \|w - w_s\| < \|\epsilon_w\|_{\text{tar}}, \text{ approximately;}}$
    $\mathcal{I} \leftarrow \mathcal{I}(w, F);$
    $\frac{\partial}{\partial w}\mathcal{I} \leftarrow \frac{\partial}{\partial w}\mathcal{I}(w, F);$
    $\mathcal{A} \leftarrow \frac{\partial}{\partial w}\mathcal{R}(w, F);$
    $\psi \leftarrow \left[\mathcal{A}^T\right]^{-1} \frac{\partial}{\partial w}\mathcal{I};$
    $\frac{\partial}{\partial F}\mathcal{I} \leftarrow \frac{\partial}{\partial F}\mathcal{I}(w, F);$
    $\psi^T \frac{\partial}{\partial F}\mathcal{R} \leftarrow \psi^T \frac{\partial}{\partial F}\mathcal{R}(w, F);$
    $\mathcal{G} \leftarrow \frac{\partial}{\partial F}\mathcal{I} - \psi^T \frac{\partial}{\partial F}\mathcal{R};$
    if $i > 1$ then
        $\boxed{S \leftarrow \max\left(S, \frac{\|\mathcal{G}-\mathcal{G}_{\text{prev}}\|}{\|w-w_{\text{prev}}\|}\right);}$
        $\boxed{\|\epsilon_w\|_{\text{tar}} \leftarrow \frac{\eta\|\mathcal{G}\|^p}{S};}$
    end
    $F \leftarrow$ Gradient-based optimization update;
    $\boxed{\mathcal{G}_{\text{prev}} \leftarrow \mathcal{G};}$
end
```

4.4 Constraint error estimation based on the residual

The error ϵ_w can be estimated from the residual. Consider, at the kth iteration of an iterative root-finding solver for the constraint equations, the Taylor series relating $\mathcal{R}(w_k, F)$ to $\mathcal{R}(w^*, F)$, which is given by

$$\mathcal{R}(w_k, F) = \mathcal{R}\left(w^*, F\right) - \frac{\partial}{\partial w}\mathcal{R}(w_k, F)\,\epsilon_{w,k} + \mathcal{O}\left(\left\|\epsilon_{w,k}\right\|^2\right). \tag{55}$$

Since $\mathcal{R}(w^*, F) = \mathbf{0}$, and ignoring higher order terms, we get an approximation to the error

$$\epsilon_{w,k} \approx -\left[\frac{\partial}{\partial w}\mathcal{R}(w_k, F)\right]^{-1}\mathcal{R}(w_k, F), \tag{56}$$

which is simply a Newton update. If an inexact linear solver is used such that the Newton update is given by

$$\left\|\frac{\partial}{\partial w}\mathcal{R}\left[w_k - w_{k-1}\right] - \mathcal{R}\right\| = \omega\left\|\mathcal{R}\right\|, \tag{57}$$

where $\omega \in \mathbb{R}$, $0 < \omega < 1$ is a (user-specified) tolerance, then $\left\|\epsilon_{w,k-1}\right\|$ can be estimated as

$$\left\|\epsilon_{w,k-1}\right\| \approx \frac{\left\|\Delta w_{k-1}\right\|}{1-\omega} \tag{58}$$

and $\left\|\epsilon_{w,k}\right\|$ can be extrapolated as

$$\left\|\epsilon_{w,k}\right\| \approx \frac{\left\|\mathcal{R}_k\right\|}{\left\|\mathcal{R}_{k-1}\right\|}\left\|\epsilon_{w,k-1}\right\| \approx \frac{\left\|\mathcal{R}_k\right\|\left\|\Delta w_{k-1}\right\|}{\left\|\mathcal{R}_{k-1}\right\|(1-\omega)}. \tag{59}$$

5 Results

5.1 Nonlinear heat transfer

The first test case is essentially a steady one-dimensional heat transfer problem, though we do not bother to make the problem physically realistic. The spatial domain is $x \in [0, 1]$ divided into equal intervals with length Δx. The discrete form of the governing equations at internal node i is

$$\frac{\alpha}{\Delta x^2}\left(w_{i-1} - 2w_i + w_{i+1}\right) + h\left(w_i - w_\infty\right) + k\left(w_i^4 - w_\infty^4\right) = 0, \tag{60}$$

where α, h, k, and w_∞ are constants. The variables w in this case may be interpreted as temperature and the terms in the above equation may be interpreted as conduction, external convection, and radiation, respectively. Dirichlet boundary conditions are applied at the endpoints.

The optimization problem is posed as an "inverse-design" problem. The physical system described above is solved for a given $k = k_0$ and the solution w^* is recorded at several grid points. Denote $\mathcal{S}$ as the set of indices at which the solution is recorded. The objective of the optimization problem is then to recover this value of k by attempting to match the solution profile. Explicitly, the objective function is

$$\mathcal{I} = \sum_{i \in \mathcal{S}} \left(w_i - w_i^*\right)^2 \tag{61}$$

and the constraints are the discrete equations representing the physical system. The constraint equations are solved directly using Newton's method. The linear system was solved using a direct method but included a relaxation $\omega = 0.6$ on the update. This relaxation factor is applied in order to relax the convergence of the nonlinear solver so that the final error does not over-shoot the target error too much for the sake of the analysis. The constraint equation error at each iteration is estimated based on (59).

The parameters for the test case are $w_{\text{left}} = 5$, $w_{\text{right}} = 6$, $\alpha = 0.1$, $h = 1$, $k = 0.01$, $w_\infty = 2$, with $\Delta x = 0.01$ and nine equi-spaced points at which the temperature is recorded for the inverse design problem. The inverse design problem is solved using the steepest descent method with step size $\gamma = 0.0001$. Algorithm 2 is used and convergence rate $p = 1$ is assumed. The optimization problem is considered converged when $\|\mathcal{G}\|$ drops 5 orders of magnitude. Setting k to 0, the inverse design problem is solved in 48 iterations of the optimization algorithm. The solution is displayed in Fig. 1.

We begin by investigating the relationship between the quantity $\|\mathcal{G} - \mathcal{G}_a\| / \|\mathcal{G}\|$ and the user parameter η. The correlation is obtained by calculating the gradient twice at each optimization iteration: once with an accurate solution to the constraint equation and again using the adapted error tolerance. The update is performed using the accurate gradient so that the study is consistent for each value of η investigated. The data are shown in Fig. 2, from which we see that the relative gradient error remains fairly constant throughout convergence and that it is consistently less than η. Hence, if η is interpreted as the target relative gradient error, then (54) produces a conservative value of $\|\epsilon_w\|_{\text{tar}}$ in this case.

Figure 3 shows the correlation between the residual to which the constraint equations are solved and the gradient. Despite solving the constraint equations to an increasingly tighter tolerance, the relative error in the gradient remains consistent, emphasizing the importance of the adaptive tolerance if deep convergence is desired. This is further emphasized in Fig. 4, where we see that using an adaptive tolerance of

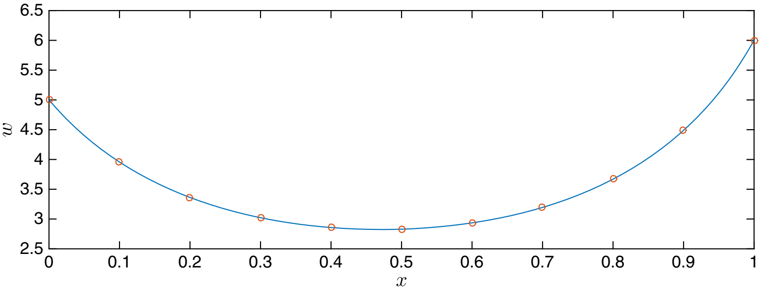

Fig. 1 Solution to the inverse design problem; target values are circled

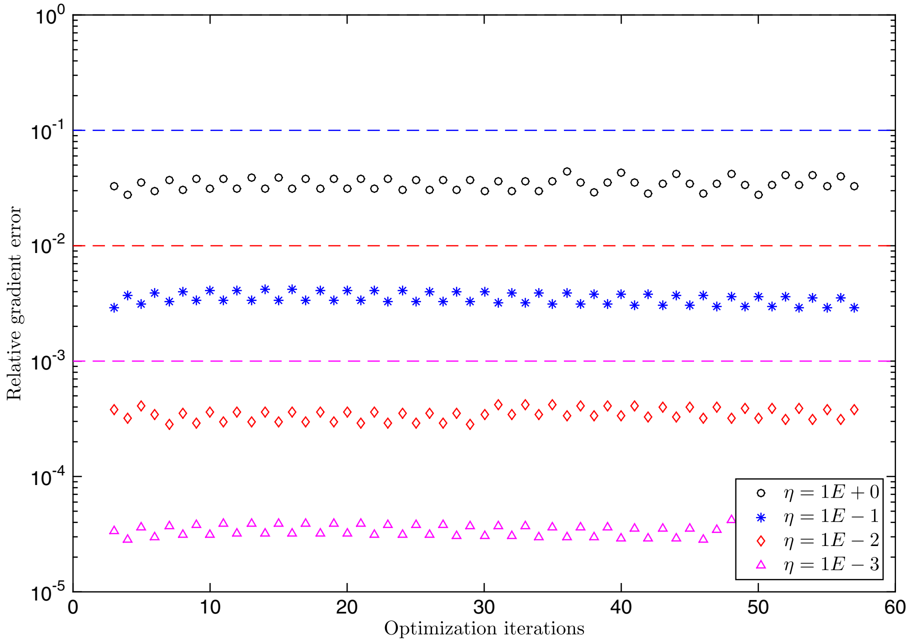

Fig. 2 Relative error in the gradient for the heat transfer case—the value of η for each corresponding colour is indicated with a dashed line

$\eta = 10^{-2}$ leads to essentially the same convergence rate as a fixed constraint equation residual tolerance of $\tau = 10^{-12}$ and allows for deep convergence, whereas less conservative values of τ result in stagnation of the cost functional $\mathcal{I}$. From Fig. 5, we see the apparent efficiency benefit of using the adaptive tolerance.

5.2 Inviscid compressible flow through a nozzle

The second test case is quasi-one-dimensional inviscid compressible air flow through a converging/diverging nozzle. The nozzle shape $S(x)$ is given by

$$S(x) = \begin{cases} 1 + k_1 \left(1 - \frac{x}{5}\right)^2 & 0 \leq x \leq 5 \\ 1 + k_2 \left(1 - \frac{x}{5}\right)^2 & 5 < x \leq 10, \end{cases} \tag{62}$$

with $k_1 = 1.5$ and $k_2 = 0.5$. The air is considered to be a perfect gas with ideal gas constant $R = 287\mathrm{N} \cdot \mathrm{m} \cdot \mathrm{kg}^{-1} \cdot \mathrm{K}^{-1}$, heat capacity ratio $\gamma = 1.4$, total temperature $T_0 = 300$ K, and total inlet pressure $p_{01} = 100$ kPa. The critical area is $S^* = 0.8$. The critical area is used to calculate the Mach number at the inlet in this case and can also be used to calculate the Mach number for all x analytically. Both problems are described by Pulliam and Zingg [10], who in turn reference Hirsch [4]. More details of the problem can be found in either textbook, including the analytical solution.

A finite-difference discretization is again used in this case with 101 cells and Dirichlet boundary conditions (density and velocity are fixed at the inlet, energy is

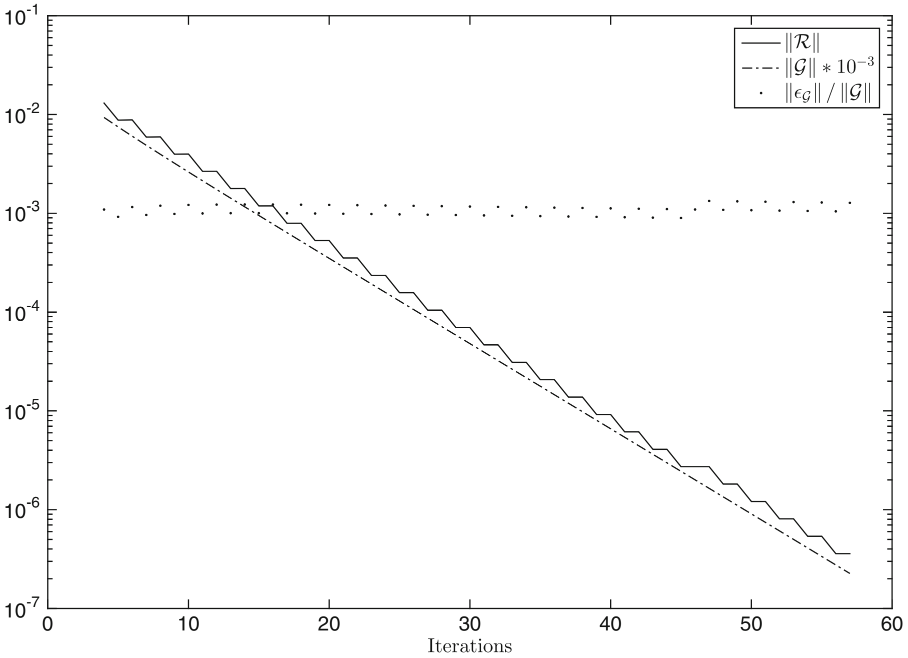

Fig. 3 Correlation of the constraint residual and gradient for the heat transfer case

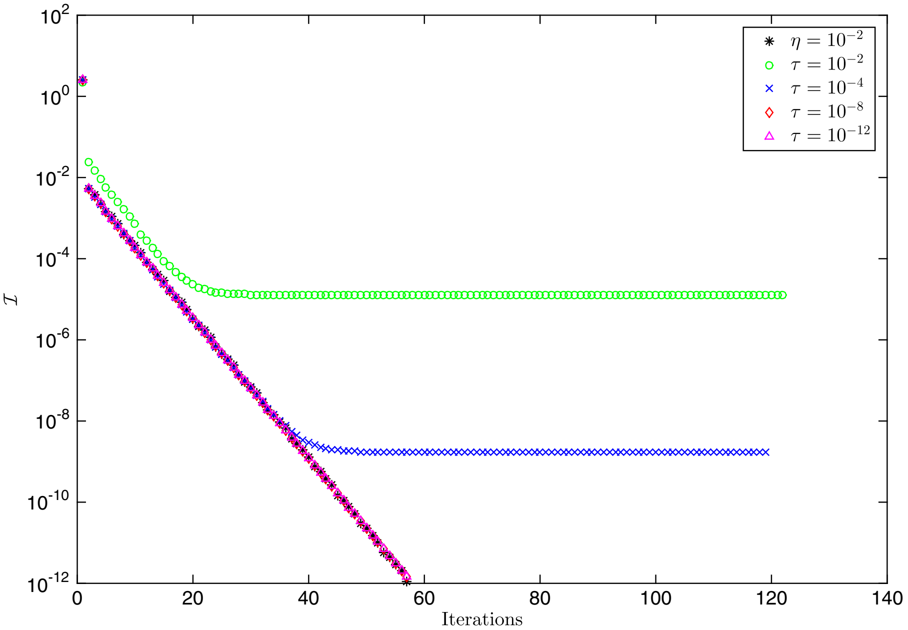

Fig. 4 Optimization algorithm convergence for the heat transfer case for several fixed τ cases and an adaptive τ case with fixed η (the $\tau = 10^{-8}$ and $\tau = 10^{-12}$ cases are not visible as they overlap with the $\eta = 10^{-2}$ case)

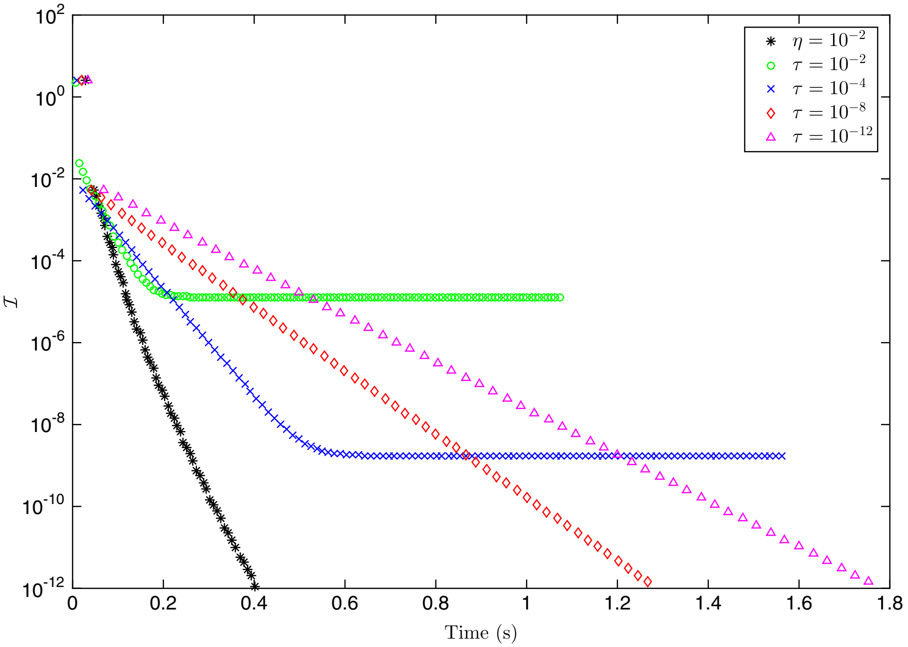

Fig. 5 Optimization algorithm performance for the heat transfer case for several fixed τ cases and an adaptive τ case with fixed η

fixed at the outlet). The optimization problem is again an inverse design problem where the goal this time is to recover the correct values of k_1 and k_2 and the objective function is based on the velocity recorded at nine equi-spaced points. If S is the set of indices corresponding to these points and u_i^*, $i \in S$ are the velocities at these points then the objective function is given by

$$\mathcal{I} = \sum_{i \in \mathcal{S}} \left(u_i - u_i^*\right)^2 . \tag{63}$$

For the results presented, we have used $k_1 = 1.4$ and $k_2 = 0.45$ as the starting guesses for k_1 and k_2. Newton's method is applied directly to the linear system, though we relax the linear solution by the factor $\omega = 0.1$. The step size used in the steepest descent algorithm is $\gamma = 1$. This time we have sought either a 15 order of magnitude drop in $\|\mathcal{G}\|$ from its initial value or $\|\mathcal{I}\| < 10^{-15}$ as termination criteria. The optimizer converges in 60 iterations under these conditions.

The same studies are performed as with the heat transfer case. The correlation between the relative gradient error $\|\mathcal{G} - \mathcal{G}_{\text{true}}\| \,/\, \|\mathcal{G}_{\text{true}}\|$ and η is shown in Fig. 6 for several values of η. As with the heat transfer case, we see that the relative gradient error remains relatively consistent throughout convergence and that it is less than the user-supplied value of η, indicating that (54) has again produced a conservative value of $\|\epsilon_w\|_{\text{tar}}$. Figure 7 shows that the constraint error and gradient are again clearly correlated for this case. Preservation of the convergence rate when using the adaptive

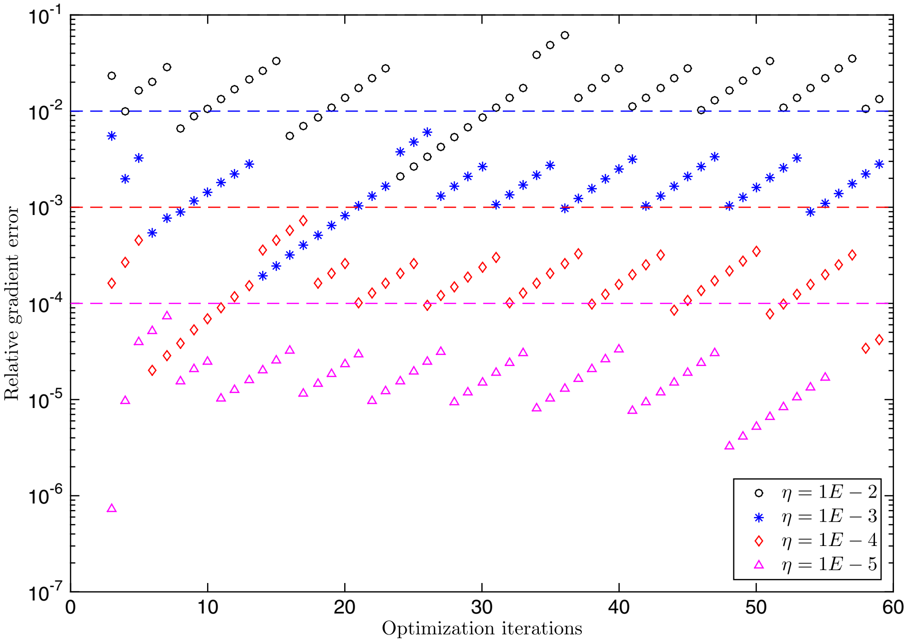

Fig. 6 Relative error in the gradient for the convection case—the value of η for each corresponding colour is indicated with a dashed line

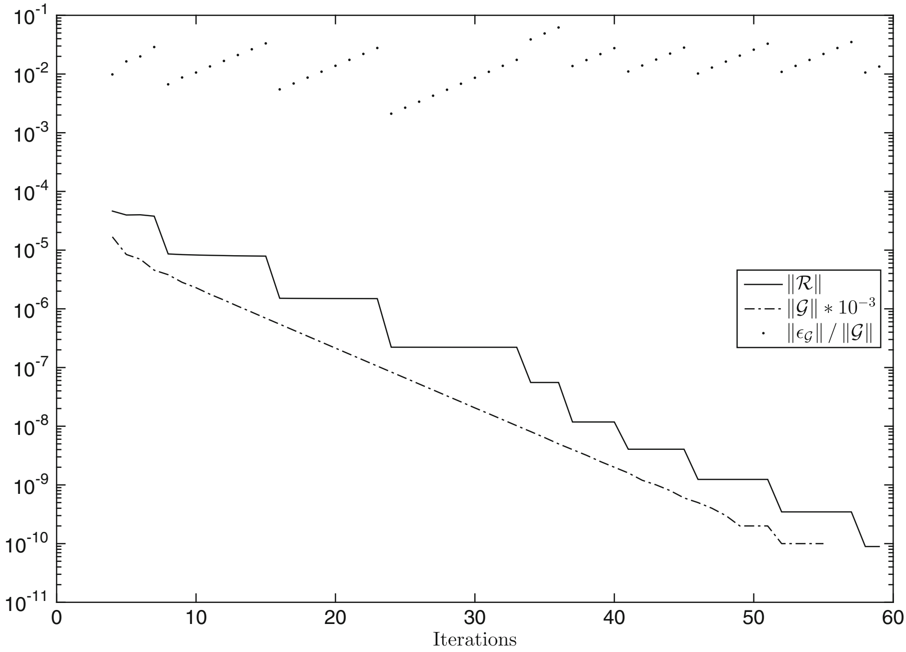

Fig. 7 Correlation of the constraint residual and gradient for the convection case

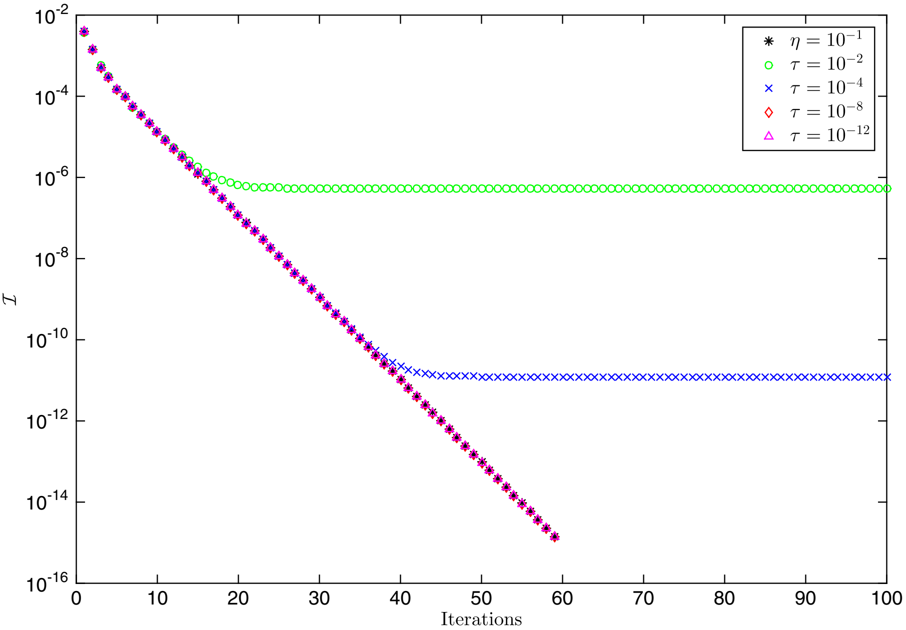

Fig. 8 Optimization algorithm convergence for the convection case for several fixed τ cases and an adaptive τ case with fixed η (the $\tau = 10^{-8}$ and $\tau = 10^{-12}$ cases are not visible as they overlap with the $\eta = 10^{-1}$ case)

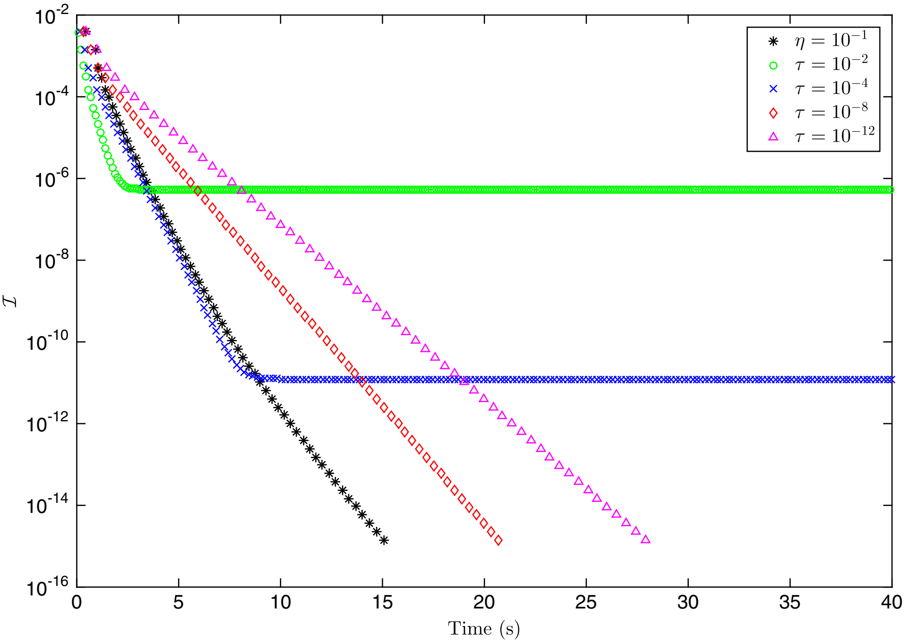

Fig. 9 Optimization algorithm performance for the convection case for several fixed τ cases and an adaptive τ case with fixed η

constraint tolerance is demonstrated in Fig. 8 and the efficiency gained by using the adaptive tolerance is demonstrated in Fig. 9.

6 Conclusions

Convergence properties of the inexactly constrained gradient-based discrete adjoint steepest descent optimization algorithm were studied. A relationship was derived between the error in the solution to the constraint equations and the resulting error in the gradient and it was established analytically that it is possible to maintain the Q-convergence rate of an exact gradient-based algorithm by adapting the constraint equation error based on the norm of the gradient and the convergence rate.

A suitable formula for adapting the constraint equation error was presented and validated for inverse design problems on a discretized one-dimensional heat transfer system as well as a one-dimensional compressible air flow system. It was found that the inexactly constrained algorithm could retain the convergence properties compared to tightly solving the constraint equations. The value of the adaptation method was demonstrated through timing comparisons as we were able to achieve much greater algorithm efficiency when adapting the constraint equation tolerance than when using a fixed tolerance.

Future investigation will be directed toward the effect of error resulting from inexactly solving the adjoint system, analysis of more sophisticated gradient-based algorithms, and application to more difficult computational problems.

Acknowledgments The authors gratefully acknowledge funding from the National Sciences and Engineering Research Council of Canada (NSERC) as well as Bombardier Aerospace.

References

1. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific, Belmont (1996)
2. Giles, M.B., Pierce, N.A.: An introduction to the adjoint approach to design. Flow Turbulence and Combustion **65**, 393–415 (2000)
3. Hicken, J.E., Zingg, D.W.: Aerodynamic optimization algorithm with integrated geometry parameterization and mesh movement. AIAA J. **48**(2), 401–413 (2010)
4. Hirsch, C.: Numerical Computation of Internal and External Flows. Wiley, New York (1988)
5. Optimum aerodynamic design using CFD and control. AIAA-95-1729 (1995)
6. Nadarajah, S.K., Jameson, A.: A comparison of the continuous and the discrete adjoint approach to automatic aerodynamic optimization. AIAA-2000-0667 (2000)
7. Nocedal, J., Wright, S.J. Numerical Optimization, 2nd edn. Springer, Berlin (2006)
8. Ortega, J.M., Rheinboldt, W.C.: Iterative solution of nonlinear equations in several variables. SIAM (1970)
9. Pironneau, O.: On optimum design in fluid mechanics. J. Fluid Mech. **64**, 97–110 (1974)
10. Pulliam, T.H., Zingg, D.W.: Fundamental algorithms in computational fluid dynamics. Springer, Berlin (2014)
11. Reuther, J., Jameson, A.: Control theory based airfoil design for potential flow and a finite volume discretization. AIAA-94-0499 (1994)