Springer Proceedings in Mathematics & Statistics

Chrysafis Vogiatzis
Jose L. Walteros
Panos M. Pardalos *Editors*

Dynamics of Information Systems

Computational and Mathematical Challenges

Springer

Springer Proceedings in Mathematics & Statistics

Volume 105

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Chrysafis Vogiatzis • Jose L. Walteros
Panos M. Pardalos
Editors

Dynamics of Information Systems

Computational and Mathematical Challenges

Springer

Editors
Chrysafis Vogiatzis
Center for Applied Optimization
Department of Industrial
and Systems Engineering
University of Florida
Gainesville, FL, USA

Jose L. Walteros
Center for Applied Optimization
Department of Industrial
and Systems Engineering
University of Florida
Gainesville, FL, USA

Panos M. Pardalos
Center for Applied Optimization
Department of Industrial
and Systems Engineering
University of Florida
Gainesville, FL, USA

Laboratory of Algorithms and Technologies
for Network Analysis (LATNA)
National Research University
Higher School of Economics
Moscow, Russia

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-10045-6 ISBN 978-3-319-10046-3 (eBook)
DOI 10.1007/978-3-319-10046-3
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014951355

Mathematics Subject Classification (2010): 90

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Information systems, now more than ever, are a vital part of modern societies. They are used in many of our everyday actions, including our online social network interactions, business and bank transactions, and sensor communications, among many others. The rapid increase in their capabilities has enabled us with more powerful systems, readily available to sense, control, disperse, and analyze information.

In 2013, we were honored to host the Fifth International Conference on the Dynamics of Information Systems. The conference focused on sensor networks and related problems, such as signal and message reconstruction, community and cohesive structures in complex networks and state-of-the-art approaches to detect them, network connectivity, cyber and computer security, and stochastic network analysis.

The Fifth International Conference on the Dynamics of Information Systems was held in Gainesville, Florida, USA, during February 25–27, 2013.

There were four plenary lectures:

- **Roman Belavkin**, Middlesex University, UK
  Utility, Risk and Information
- **My T. Thai**, University of Florida, USA
  Interdependent Networks Analysis
- **Viktor Zamaraev**, Higher School of Economics, Russia
  On coding of graphs from hereditary classes
- **Jose Principe**, University of Florida, USA
  Estimating entropy with Reproducing Kernel Hilbert Spaces

All manuscripts submitted to this book were independently reviewed by at least two anonymous referees. Overall, this book consists of ten contributed chapters, each dealing with a different aspect of modern information systems with an emphasis on interconnected network systems and related problems.

The conference would not have been as successful without the participation and contribution of all the attendees and thus we would like to formally thank them. We would also like to extend a warm thank you to the members of the local organizing committee and the Center for Applied Optimization.

We would also like to extend our appreciation to the plenary speakers and to all the authors who worked hard on submitting their research work to this book. Last, we thank Springer for making the publication of this book possible.

Gainesville, FL, USA
June 2014

Chrysafis Vogiatzis
Jose L. Walteros
Panos M. Pardalos

Contents

Asymmetry of Risk and Value of Information

Roman V. Belavkin

Abstract The von Neumann and Morgenstern theory postulates that rational choice under uncertainty is equivalent to maximization of expected utility (EU). This view is mathematically appealing and natural because of the affine structure of the space of probability measures. Behavioural economists and psychologists, on the other hand, have demonstrated that humans consistently violate the EU postulate by switching from risk-averse to risk-taking behaviour. This paradox has led to the development of descriptive theories of decisions, such as the celebrated prospect theory, which uses an S-shaped value function with concave and convex branches explaining the observed asymmetry. Although successful in modelling human behaviour, these theories appear to contradict the natural set of axioms behind the EU postulate. Here we show that the observed asymmetry in behaviour can be explained if, apart from utilities of the outcomes, rational agents also value information communicated by random events. We review the main ideas of the classical value of information theory and its generalizations. Then we prove that the value of information is an S-shaped function and that its asymmetry does not depend on how the concept of information is defined, but follows only from linearity of the expected utility. Thus, unlike many descriptive and 'non-expected' utility theories that abandon the linearity (i.e. the 'independence' axiom), we formulate a rigorous argument that the von Neumann and Morgenstern rational agents should be both risk-averse and risk-taking if they are not indifferent to information.

Keywords Decision-making • Expected utility • Prospect theory • Uncertainty • Information

R.V. Belavkin (✉)
Middlesex University, London NW4 4BT, UK
e-mail: R.Belavkin@mdx.ac.uk

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_1

1 Introduction

A theory of decision-making under uncertainty is extremely important, because it suggests models of rational choice used in many practical applications, such as optimization and control systems, financial decision-support systems and economic policies. Therefore, the fact that one of the most fundamental principles of such a theory remains disputed for more than half a century is not only intriguing, but points at a lack of understanding with potentially dangerous consequences. The principle is the von Neumann and Morgenstern expected utility postulate [18], which follows very naturally from some fundamental ideas of probability theory, and it has become an essential part of game theory, operations research, mathematical economics and statistics (e.g. [20,31]). Several researchers, however, were sceptical about the validity of the postulate and devised clever counter-examples undermining the expected utility idea (e.g. [1,6]). Psychologists and behavioural economists have studied such examples in experiments and demonstrated consistently over several decades that the expected utility fails to explain human behaviour in some situations of making choice under uncertainty (e.g. see [8, 30]). The attempts to dismiss these observations simply by humans' ignorance about game and probability theories were quickly challenged, when professional traders were shown to conform to these 'irrational' patterns of decision-making [13]. A suggestion that the human mind is somehow inadequate for making decisions under uncertainty should be taken with caution, considering that it has evolved over millions of years to do exactly that.

One of the most successful behavioural theories explaining the phenomenon is prospect theory [9], which suggests that humans value prospects of gains differently from prospects of losses, and therefore their attitude to risk is different in these situations. To model this asymmetry of risk an S-shaped value function with concave and convex branches was proposed (e.g. see Fig. 1). Unfortunately, it is precisely this asymmetry that appears to be in conflict with the expected utility theory and specifically with the axioms that imply its linear (or affine) properties (the so-called independence axiom [15]). Many attempts to develop theories without such axioms have been made, such as the regret theory [14] and other 'non-expected' utility theories (see [16, 17, 22]). The main aim of this work is to show that another approach is possible, and it involves one important concept emerging from physics and now entering new areas of science, and it is the concept of *entropy*.

Entropy is an information potential, and decision-making under uncertainty can be improved, if some additional information is provided. This improvement implies that information has utility, and the amalgamation of these two concepts is known as the value of information theory, which was developed in the mid-1960s by Stratonovich and Grishanin as a branch of information theory and theoretical cybernetics [7,23–28]. This theory considered variational problems of maximization or minimization of expected utility subject to constraints on information. One of many interesting results is an S-shaped value function representing the value of information, which resembles the S-shaped value function in prospect theory. Analysis shows that this geometric property is the consequence of linearity of

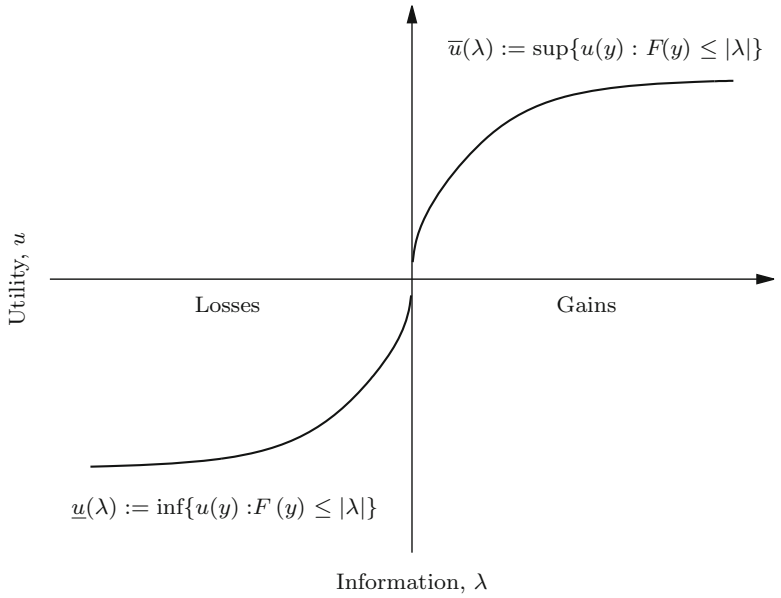

Fig. 1 An S-shaped value function with concave and convex branches used in prospect theory [9] to model risk-aversion for gains and risk-taking for losses. These properties also characterize two branches of the value of information: $\overline{u}(\lambda)$ is concave and plotted here against 'positive' information associated with gains; $\underline{u}(\lambda)$ is convex and plotted against 'negative' information associated with losses

the expected utility functional, and it is independent of any specific definition of information [2]. Thus, rational agents that are not indifferent to information should value information about gains differently from information about losses, and this may explain the observed asymmetry in humans' attitude towards risk. The advantage of the proposed approach is that it does not contradict, but generalizes the expected utility postulate.

In the next section, we review the main mathematical principles behind the expected utility postulate. The presentation of axioms follows the theory of ordered vector spaces, and it allows the author to give a very short and simple proof of the postulate in Theorem 2. The aim of this section is to show that the ideas behind the expected utility are very natural and fundamental. Section 3 overviews several classical examples that are often used in psychological experiments to test humans' preferences and attitude towards risk. Some examples are presented in a slightly simplified form to illustrate the idea. The basic concepts of information theory and the classical value of information theory are presented in the first half of Sect. 4. Then an abstraction will be made using convex analysis to show that the S-shape characterizes the value of an abstract information functional. We conclude with a brief discussion of the paradoxes.

2 Linear Theory of Utility

We review the definition of a preference relation, its utility representation and the condition of its existence. Then we show that in the category of linear spaces, such as the vector space of measures, the preference relation should be linear and represented by a linear functional, such as the expected utility.

2.1 Abstract Choice Sets and Their Representations

A set Ω is called an abstract *choice* set, if any pair of its elements can be compared by a transitive binary relation $\precsim$, called the *preference* relation:

Definition 1 (Preference relation). A binary relation $\precsim \subseteq \Omega \times \Omega$ that is

1. Total[1]: $a \precsim b$ or $a \succsim b$ for all $a, b \in \Omega$.
2. Transitive: $a \precsim b$ and $b \precsim c$ implies $a \precsim c$.

One can see that $\precsim$ is a total *pre-order* (reflexivity of $\precsim$ follows from the fact that it is total). We shall denote by $\succsim$ the inverse relation $(\precsim)^{-1}$. We shall distinguish between the strict and non-strict preference relations, which are defined respectively as follows:

$$a \prec b := (a \precsim b) \wedge \neg (a \succsim b)$$
$$a \sim b := (a \precsim b) \wedge (a \succsim b).$$

Non-strict preference $\sim$ is also called an *indifference*, and it is an equivalence relation. The quotient set $\Omega/\sim$ defined by this equivalence relation is the set of equivalence classes $[a] := \{b \in \Omega : a \sim b\}$, which are totally ordered.

It is quite natural in applications to map the choice set to some standard ordered set, such as $\mathbb{N}$ or $\mathbb{R}$. Such numerical mapping is called a *utility representation*:

Definition 2 (Utility representation of $\precsim$). A real function $u : (\Omega, \precsim) \to (\mathbb{R}, \leq)$ such that:

$$a \precsim b \quad \Longleftrightarrow \quad u(a) \leq u(b).$$

Observe that the mapping above is monotonic in both directions, which means that utility defines an order-embedding of $(\Omega/\sim, \leq)$ into $(\mathbb{R}, \leq)$. Clearly, a utility

[1] This property is sometimes called *completeness*, but this term often has other meanings in order theory (e.g. complete partial order) or topology (e.g. complete metric space).

representation exists for any countable choice set Ω. For uncountable Ω, the existence of a utility representation is not guaranteed, and it is given by the following condition:

Theorem 1 (Debreu [5]). *A utility representation of uncountable* $(\Omega, \lesssim)$ *exists if and only if there is a countable subset* $Q \subset \Omega$ *that is* order dense*: for all* $a < b$ *in* $\Omega \setminus Q$ *there is* $q \in Q$ *such that* $a < q < b$.

Note that in optimization theory and its applications one often begins the analysis with a given real objective function $u : \Omega \to \mathbb{R}$ (e.g. a utility function u or a cost function $-u$). The preference relation $\lesssim$ is then induced on Ω by the values $u(\omega) \in \mathbb{R}$ as a pullback of order $\leq$ on $\mathbb{R}$. This *nuclear* binary relation $\lesssim$ is clearly total and transitive. Therefore, although some works consider non-total or nontransitive preferences, as well as relations without a utility function, this paper focuses only on choice sets with utility representations.

2.2 Choice Under Uncertainty

By definition, a utility representation $u : \Omega \to \mathbb{R}$ is an embedding of the pre-ordered set $(\Omega, \lesssim)$ into $(\mathbb{R}, \leq)$, so that the quotient set $(\Omega/\sim, \leq)$ is order-isomorphic to the subset $u(\Omega) \subseteq \mathbb{R}$. Recall, that the set of real numbers $(\mathbb{R}, \leq)$ is more than just an ordered set—it is a totally ordered field, in which the order is compatible with the algebraic operations of addition and multiplication, is Archimedean (see below), and it is the only such field. Suppose that the choice set Ω is also equipped with some algebraic operations. Then it appears quite natural if utility $u : \Omega \to \mathbb{R}$ is compatible also with these algebraic operations, acting as a homomorphism. In the language of category theory, utility should be a morphism between objects Ω and $u(\Omega) \subseteq \mathbb{R}$ of the same category. For example, if Ω is a subset of a real vector space Y, then in the category of linear spaces or algebras, like $(\mathbb{R}, \leq)$, pre-order $(Y, \lesssim)$ (extended from $\Omega \subseteq Y$) should be compatible with the vector space operations

$$x \lesssim y \iff \lambda x \lesssim \lambda y\,, \qquad \forall\, \lambda > 0 \tag{1}$$

$$x \lesssim y \iff x + z \lesssim y + z\,, \qquad \forall\, z \in Y \tag{2}$$

and Archimedean

$$n x \lesssim y\,, \quad \forall\, n \in \mathbb{N} \quad \Rightarrow \quad x \lesssim 0. \tag{3}$$

These three axioms are often assumed in the category of pre-ordered vector spaces. Note that classical texts on expected utility (e.g. [18, 20]) present these axioms in a different form, because of a restriction to an affine subspace of a vector space due to

normalization and positivity conditions for probability measures. Thus, axioms (1) and (2) are combined into the so-called independence axiom:

$$x \lesssim y \iff \lambda x + (1-\lambda)z \lesssim \lambda y + (1-\lambda)z\,, \qquad \forall\, z \in Y\,,\ \lambda \in (0,1].$$

The Archimedean axiom (3) is replaced by the *continuity* axiom:

$$x \lesssim y \lesssim z \quad \Rightarrow \quad y \sim \lambda x + (1-\lambda)z\,, \qquad \exists\, \lambda \in [0,1].$$

The author finds it more convenient to work in the category of linear spaces and making the restriction to an affine subspace when necessary. Thus, we shall assume axioms (1)–(3). Substituting $z = -x - y$ into (2) gives also

$$x \lesssim y \quad \iff \quad -x \gtrsim -y \tag{4}$$

and together with axiom (1) this property also implies that

$$x \sim y \quad \iff \quad \lambda x \sim \lambda y\,, \qquad \forall\, \lambda \in \mathbb{R} \tag{5}$$

The linear or affine algebraic structures occur naturally in measure theory and probabilistic models of uncertainty. Indeed, consider a probability space $(\Omega, \mathcal{F}, P)$, where Ω is the set of elementary events, $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra of events and $P : \mathcal{F} \to [0,1]$ is a probability measure. In the context of game theory or economics, the probability measure P, defined over the choice set $(\Omega, \lesssim)$ with utility $u : \Omega \to \mathbb{R}$, is often referred to as a *lottery*, emphasizing the fact that utility is now a random variable (assuming it is $\mathcal{F}$-measurable). The *expected utility* associated with event $E \subseteq \Omega$ is given by the integral:

$$\mathbb{E}_P\{u\}(E) = \int_E u(\omega)\, dP(\omega).$$

In particular, the utility associated with elementary event $a \in \Omega$ can be defined as $u(a) = \int_\Omega u(\omega)\, \delta_a(\omega)$, where δ_a is the elementary probability measure (i.e. the Dirac δ-measure concentrated entirely on $a \in \Omega$).

Probability measures, or 'lotteries', are elements of a vector space, for example, the space $Y = \mathcal{M}_c(\Omega)$ of signed Radon measures on Ω [4]. We remind that signed Radon measures are bounded linear functionals $y(f) = \int f\, dy$ on the space $X = \mathcal{C}_c(\Omega, \mathbb{R})$ of continuous functions $f : \Omega \to \mathbb{R}$ with compact support (i.e. $Y = X'$ is the space of distributions dual of the space X of test functions). Measures that are non-negative $y(E) = \int_E dy \geq 0$ for all $E \subseteq \Omega$ form a convex cone in Y. The normalization condition $y(\Omega) = 1$ defines an affine set in Y, and its intersection with the positive cone defines its base:

$$\mathcal{P}(\Omega) := \{y \in Y : y(E) \geq 0\,,\ y(\Omega) = 1\}.$$

The base $\mathcal{P}(\Omega)$ is the set of all Radon probability measures on Ω. It is a weakly compact convex set, and by the Krein-Milman theorem each point $p \in \mathcal{P}(\Omega)$ can be represented as a convex combination of its extreme points δ_ω—the elementary measures on Ω. In fact, $\mathcal{P}(\Omega)$ is a simplex, so that representations are unique and the set ext $\mathcal{P}(\Omega)$ of extreme points is identified with the set Ω of elementary events. Figure 2 shows an example of two-simplex, which is the set $\mathcal{P}(\Omega)$ of lotteries over three outcomes $\Omega = \{\omega_1, \omega_2, \omega_3\}$.

A question that arises in this construction is: How should the preference relation $\lesssim$ on Ω be extended to the set $\mathcal{P}(\Omega)$ of all 'lotteries' over Ω? Because $\mathcal{P}(\Omega)$ is a subset of a vector space, it is quite natural to require that $\lesssim$ satisfies axioms (1), (2) and (3), and this leads immediately to the following result.

Theorem 2 (Expected utility). *A totally pre-ordered vector space $(Y, \lesssim)$ satisfies axioms (1)–(3) if and only if $(Y, \lesssim)$ has a utility representation by a closed[2] linear functional $u : Y \to \mathbb{R}$.*

Proof.

($\Leftarrow$) The necessity of axioms (1) and (2) follows immediately from linearity of functional $u : Y \to \mathbb{R}$, representing $(Y, \lesssim)$. The Archimedean axiom (3) is necessary if u is closed: $u(x) = \upsilon$ for every convergent sequence $x_n \to x$ such that $u(x_n) \to \upsilon$ (i.e. $u(\lim x_n) = \lim u(x_n)$). Indeed, assume $nz \lesssim y$ for all $n \in \mathbb{N}$ and some $z > 0$. Then $x_n = y/n \gtrsim z > 0$ for all $n \in \mathbb{N}$, and therefore $\lim u(x_n) \geq u(z) > 0$, because u is a representation of $(Y, \lesssim)$. But $\lim x_n = y \lim(1/n) = y \cdot 0 = 0$, meaning that u is not closed.

($\Rightarrow$) First, we show that axioms (1) and (2) imply that the equivalence classes $[x] := \{y : x \sim y\}$ are affine. Indeed, assume they are not affine. Then there exist two points x, y in $[x]$ such that the line passing through them contains a point that does not belong to $[x]$. That is $(1-\lambda)x + \lambda y \notin [x]$ for some $\lambda \in \mathbb{R}$ and x, $y \in [x]$. This means, for example, that

$$x \sim y < (1-\lambda)x + \lambda y.$$

Using property (5), let us replace λy by the equivalent λx, so that we have

$$x \sim y < (1-\lambda)x + \lambda x = x.$$

But $y < x$ contradicts our assumption $x \sim y$ (and $x < x$ is a contradiction as well). Therefore, for any x and y in $[x]$, the whole line $\{z : z = (1-\lambda)x + \lambda y\,,\ \lambda \in \mathbb{R}\}$ is also in $[x]$. Thus, if $(Y, \lesssim)$ has a utility representation, then it can be taken to be an affine or a linear functional $u(y) = \int u\, dy$, because it must have affine level sets $[x] = \{y : u(y) = u(x)\}$.[3]

[2] We use the notion of a closed functional, because the topology in Y is not defined.

[3] An affine functional h and a linear functional $u(y) = h(y) - h(0)$ have isomorphic level sets.

Second, we prove that axiom (3) implies that there is a countable order-dense subset $Q \subset Y$, so that $(Y, \lesssim)$ has a utility representation by Theorem 1. Indeed, take $Q := \{mz/n : z > 0,\ m/n \in \mathbb{Q}\}$. Case $x < 0 < y$ is trivial; therefore consider the case $0 < x < y$ (or equivalently $x < y < 0$). Because $z > 0$, axiom (3) implies that $z/n < y - x$ for some $n \in \mathbb{N}$ or

$$x < x + z/n < y.$$

If $x \sim mz/n \in Q$, then $x < q < y$ for $q = (m+1)z/n \in Q$. Otherwise, if $x \nsim mz/n \in Q$ for all $m/n \in \mathbb{Q}$, then $mz/n < x < (m+1)z/n$ for some $m, n \in \mathbb{N}$. But this means $(m+1)z/n < y$, because $(m+1)z/n = z/n + mz/n < x + z/n < y$. Thus, we have found $q = (m+1)z/n \in Q$ with the property $x < q < y$. □

The restriction of the linear functional $u(y) = \int u\, dy$ to the set $\mathcal{P}(\Omega)$ of probability measures is the expected utility: $u(y)|_{\mathcal{P}} = \mathbb{E}_P\{u\} = \int u\, dP$. Thus, Theorem 2 generalizes the EU postulate [18]: the preference relation $(\mathcal{P}(\Omega), \lesssim)$ satisfies axioms (1)–(3) if and only if there exists $u : \Omega \to \mathbb{R}$ such that

$$Q \lesssim P \quad \Longleftrightarrow \quad \mathbb{E}_Q\{u\} \le \mathbb{E}_P\{u\} \qquad \forall\, Q, P \in \mathcal{P}(\Omega). \tag{6}$$

Note that the proof of the above result can be quite complicated (e.g. it spans five pages in [11]), while the proof of Theorem 2 appears to be simpler.

3 Violations of Linearity and Asymmetry of Risk

The linear theory described above is quite beautiful because it follows naturally from some basic mathematical principles. However, its final conclusion, the EU postulate (6), appears to be over-simplistic: according to it, a decision-maker should pay attention only to the first moments of utility distributions; all other information, such as their variance or higher-order statistics, should be disregarded. The fact that this idea is rather naive becomes obvious, when one attempts to apply it in practical situations involving money. Many counter-examples and paradoxes have been discussed in the literature (e.g. see [1, 6, 30]). Here we review some of them with the aim to show that the expected utility does not fully characterize an important aspect of decision-making under uncertainty, and that is the concept of risk.

3.1 Risk-Aversion

Consider the following example:

Example 1. Let $\Omega = \{\omega_1, \ldots, \omega_4\}$ be four elementary outcomes that carry utilities $u(\omega) \in \{-\$1000, -\$1, \$1, \$1000\}$. Consider two lotteries over these outcomes:

$$P(\omega) \in \{0, 0.5, 0.5, 0\}, \quad Q(\omega) \in \{0.5, 0, 0, 0.5\}.$$

Both lotteries have zero expected utility $\mathbb{E}_P\{u\} = \mathbb{E}_Q\{u\} = \$0$. Thus, according to the EU postulate (6), a rational agent should be indifferent $P \sim Q$. However, lottery Q appears to be more 'risky', as there is an equal chance of losing or winning \$1000 in Q as opposed to losing or winning just \$1. Thus, a risk-averse agent should prefer $P > Q$.

This example illustrates that risk is related somehow to the higher-order moments of utility distribution, such as variance $\sigma^2(u)$ (i.e. expected squared deviation from the mean). In fact, financial risk is often defined as the probability of an outcome that is preferred much less than the expected outcome (i.e. the probability of negative deviation $u(\omega) - \mathbb{E}_P\{u\} < 0$). Other higher-order statistics can also be useful, and in the next section we discuss entropy and information in relation to risk. The following example supports this idea.

Example 2 (The Ellsberg paradox [6]). The lotteries P and Q are represented by two urns with 100 balls each. There are 50 red and 50 white balls in urn P; the ratio of red and while balls in urn Q is unknown. The player is offered to draw a ball from any of the two urns. If the ball is red, then the player wins \$100. Which of the urns should the player prefer?

The choice can be represented by two lotteries:

P: The probabilities of winning \$100 and winning nothing are equal: $P(\$100) = P(\$0) = 0.5$.
Q: The probability of winning \$100 is unknown: $Q(\$100) = t \in [0, 1]$.

One can check that $\mathbb{E}_P\{u\} = \mathbb{E}_Q\{u\} = \int_0^1 (\$100 \cdot t + \$0 \cdot (1-t))\, dt = \$50$. Thus, the player should be indifferent $P \sim Q$ according to the EU postulate (6). There is an overwhelming evidence, however, that most humans prefer $P > Q$, which suggests that they prefer more information about the parameters of the distribution in this game.

Whether an agent is risk-averse or not may depend on its wealth. However, it is generally assumed that most rational agents are risk-averse, when unusually high amounts of money are involved, and this is represented by a concave 'utility of money' function [11]. This is justified by the idea that the utility of gaining \$1 relative to some amount $C > 0$ is decreasing as C grows. The origin of this idea is in the St. Petersburg paradox due to Nicolas Bernoulli (1713).

Example 3 (The St. Petersburg lottery). The lottery is played by tossing a fair coin repeatedly until the first head appears. Thus, the set Ω of elementary events is the set of all sequences of $n \in \mathbb{N}$ coin tosses. If the head appeared on the nth toss, then

the player wins $\$2^n$. Clearly, it is impossible to loose in this lottery. However, to play the lottery the player must pay an entree fee $C > 0$. The question is: What amount $C > 0$ should a rational agent pay?

According to the EU postulate (6) the fee C should not exceed the expected utility $\mathbb{E}_P\{u\}$ of the lottery. It is easy to see, however, that for a fair coin $P(\omega_n) = 2^{-n}$, and therefore the expected utility diverges

$$\mathbb{E}_P\{u\} = \sum_{n=1}^{\infty} \frac{2^n}{2^n}.$$

Thus, any amount $C > 0$ appears to be a rational fee to pay. The paradox is that not many people would pay more than $C = \$2$. The solution proposed by Daniel Bernoulli in [3] was to convert the utility $2^n \mapsto \log_2 2^n = n$. Although this does not resolve the general problem of unbounded expectations (e.g. one can introduce another lottery Q such that $\mathbb{E}_Q\{\log_2(u)\}$ diverges), this was the first example of a concave function used to represent risk-averse utility.

Note that although the 'utility of money' can be concave as a function of $x(\omega) \in \mathbb{R}$ amount, the expected utility is still a linear functional on the set $\mathcal{P}(\Omega)$ of lotteries. The level sets of the expected utility are affine sets corresponding to equivalence classes of lotteries with respect to $\lesssim$ that are parallel to each other (see Fig. 2). The risk-averse concave modification simply gives less weight to higher values $x(\omega)$. This modification also reduces the variance of the lottery.

3.2 Risk-Taking

It is not difficult to introduce a lottery in which risk-taking appears to be rational.

Example 4 (The 'Northern Rock' lottery). A player is allowed to borrow any amount $C > 0$ from a bank. When repayment is due, the amount to repay is decided in the St. Petersburg lottery: a fair coin is tossed repeatedly until the first head appears. If the head appeared on the nth toss, then the player has to repay $\$2^n$ to the bank (i.e. the utility is $u(\omega_n) = -\$2^n$). The question is: What amount $C > 0$ should a rational agent borrow?

Again, according to the EU postulate (6), one should not borrow an amount C that is less than the expected repayment $\mathbb{E}_P\{-u\}$. However, assuming that the probability $P(\omega_n) = 2^{-n}$ for a fair coin, it is easy to see that the expected repayment diverges, and therefore a rational agent should not borrow at all. Although the author did not conduct a systematic study of this problem, anecdotal evidence suggests that many people do borrow substantial amounts. The solution to this paradox can be made similar to [3] by modifying the utility $-2^n \mapsto -\log_2 2^n = -n$. Observe that the utility for repayments is not concave, but convex (negative logarithm), and therefore it appears to represent not a risk-averse, but a risk-taking utility.

One of the most striking counter-examples to the expected utility postulate was introduced by Allais [1]. Similar problems were studied by psychologists [30], which demonstrated the importance of how the outcomes are 'framed' or perceived by an agent. There are many versions of this problem, and the version below was used by the author in multiple talks on the subject.

Example 5 (The Allais paradox [1]). Consider which of the two lotteries you prefer to play:

P: Win \$300 with probability $P(\$300) = 1/3$ or nothing with $P(\$0) = 2/3$.
Q: Win \$100 with certainty $Q(\$100) = 1$.

One can check that $\mathbb{E}_P\{u\} = \mathbb{E}_Q\{u\} = \$100$, which implies indifference $P \sim Q$ according to the EU postulate (6). There is an overwhelming evidence, however, that most humans prefer $P < Q$, which suggests that they are risk averse in this game. Consider now another set of two lotteries:

P: Lose \$300 with probability $P(-\$300) = 1/3$ or nothing with $P(\$0) = 2/3$.
Q: Lose \$100 with certainty $Q(-\$100) = 1$.

Again, it is easy to check that $\mathbb{E}_P\{u\} = \mathbb{E}_Q\{u\} = -\$100$, corresponding to indifference $P \sim Q$ according to the EU postulate (6). However, most humans prefer $P > Q$, which suggests a risk-taking behaviour.

A risk-averse preference is usually observed when the outcomes are associated with gains (positive change of utility), while a risk-taking preference is observed when the outcomes are associated with losses. This phenomenon of switching from risk-averse to risk-taking behaviour is sometimes referred to as the 'reflection effect'. Note that gains can be converted into losses by multiplying their utility by -1 and vice versa. In fact, this reflection was used in the construction of Example 4 from the St. Petersburg lottery. The use of concave functions for a risk-averse utility and convex functions for a risk-taking utility can also be explained using this reflection: recall that function $u(x)$ is concave if and only if $-u(x)$ is convex.

3.3 Why Is This a Paradox?

The reflection effect is quite systematic [14, 30], and the Allais paradox was demonstrated in numerous experiments [8] including professional traders [13]. This asymmetric perception of risk has been modelled in prospect theory [9] by an S-shaped value function, such as a function shown in Fig. 1, which has a concave branch for outcomes associated with gains and convex branch for outcomes associated with losses. Although this descriptive theory has gained significant recognition among psychologists and behavioural economists, it is not clear how the concave-convex properties of the prospect value function can be derived mathematically;

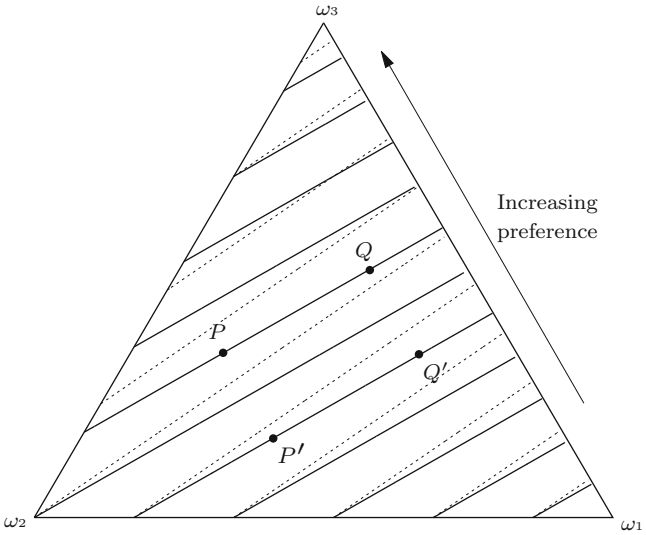

Fig. 2 Level sets of expected utility on the two-simplex of probability measures over set $\Omega = \{\omega_1, \omega_2, \omega_3\}$ with preference $\omega_1 < \omega_2 < \omega_3$. *Dotted lines* represent level sets after a risk-averse modification of the utility function

they are just postulated. Moreover, the reflection effect it models appears to violate the beautiful and natural set of axioms behind the expected utility postulate [18] [specifically, axioms (1) and (2)].

As mentioned earlier, the expected utility $\mathbb{E}_P\{u\} = \int u\, dP$ is a linear functional on the set $\mathcal{P}(\Omega)$ of probability measures (lotteries) regardless of the 'shape' of the utility function $u : \Omega \to \mathbb{R}$ on the extreme points ext $\mathcal{P}(\Omega) \equiv \Omega$. The equivalence classes of the preference relation $\lesssim$ induced on the set of lotteries $\mathcal{P}(\Omega)$ by the expected utility are the level sets $[\upsilon] := \{P : \mathbb{E}_P\{u\} = \upsilon\}$, and they are affine sets. These level sets are shown in Fig. 2 by parallel lines, where the triangle (a two-simplex) represents the set $\mathcal{P}(\Omega)$ of lotteries over three elements. Assuming the preference relation $\omega_1 < \omega_2 < \omega_3$ and taking the utility of ω_2 as the reference level, lotteries above the reference level set (e.g. shown by points P and Q) can be considered as gains, while lotteries below the reference (points P' and Q') as losses. To model a risk-averse pattern, one has to modify the utility by giving lower values to the most preferable outcomes (i.e. to decrease the utility of ω_3). This modification of utility changes the level sets of expected utility, as shown in Fig. 2 by dotted parallel lines. One can notice that lotteries with higher variances or entropies (these are lotteries closer to the middle point of the simplex) are preferred less than they were before the 'risk-averse' modification (they are below the dotted lines). However, because the level sets are parallel to each other, this change applies equally to gains and losses (i.e. lotteries P and Q above and P' and Q' below the reference level). Thus, if a rational agent uses the expected utility model to

rank lotteries, then they only can be risk-averse or risk-taking, but not both. This observation was illustrated on a two-simplex in [15], and it clearly showed why the reflection effect cannot be explained by the expected utility theory alone. Thus, it appears that human decision-makers violate the linear axioms (1) and (2), and several 'non-expected' utility theories have been proposed, such as the regret theory [14] (see [16, 17, 22] for a review of many others).

4 Risk and Value of Information

As discussed previously, risk is related to a deviation from expected utility, and many examples suggest its relation to variance or higher-order statistics of the utility distribution. Another functional characterizing the distribution is *entropy*, which is closely related to variance and higher-order cumulants of a random variable. Entropy defines the maximum amount of information that a random variable can communicate. Although information is measured in *bits* or *nats* that have no monetary value, when put in the context of decision-making or estimation, information defines the upper and lower bounds of the expected utility. This amalgamation of expected utility and information is known as the *value of information* theory pioneered by [23]. Remarkably, the value of information function has two distinct branches—one is concave, representing the upper frontier of expected utility, while another is convex, representing the lower frontier of expected utility. Interestingly, it was shown recently that these geometric properties do not depend on the definition of information itself, but follow only from the linearity of expected utility [2]. In this section, we discuss the classical notion of value of information, its generalization and how it can be related to asymmetry of risk.

4.1 Information and Entropy

Information measures the ability of two or more systems to communicate and therefore depend on each other. System A influences system B (or B depends on A) if the conditional probability $P(B \mid A)$ is different from the prior probability $P(B)$; or equivalently, if the joint probability $P(A \cap B)$ is different from the product probability $Q(A) \otimes P(B)$ of the marginals. Shannon defined *mutual information* [21] as the expectation of the logarithmic difference of these probabilities:

$$I_S(A, B) := \int_{A \times B} \left[\ln \frac{dP(b \mid a)}{dP(b)} \right] dP(a, b).$$

Mutual information is always non-negative with $I_S(A, B) = 0$ if and only if A and B are independent (i.e. $P(B \mid A) = P(B)$). The supremum of $I_S(A, B)$ is attained

for $P(B \mid A)$ corresponding to an injective mapping $f : A \to B$, and it can be infinite. Note that mutual information in this case equals the *entropy* of the marginal distributions.

Indeed, recall that entropy of distribution $P(B)$ is defined as the expectation of its negative logarithm:

$$H(B) := -\int_B [\ln dP(b)]\, dP(b).$$

One can rewrite the definition of mutual information as the difference of marginal and conditional entropies:

$$I_S(A, B) = H(B) - H(B \mid A) = H(A) - H(A \mid B).$$

When $P(B \mid A)$ corresponds to a function $f : A \to B$, the conditional entropy is zero $H(B \mid A) = 0$, and the mutual information equals entropy $H(B)$. For example, by considering $A \equiv B$, one can define entropy as *self-information* $I_S(B, B) = H(B) - H(B \mid B) = H(B)$ (i.e. $P(B \mid B)$ is the identity mapping $\mathrm{id} : B \to B$). More generally, conditional entropies are zero for any bijection $f : A \to B$, so that $I_S(A, B) = H(A) = H(B)$ is the supremum of $I_S(A, B)$. Thus, we can give the following variational definition of entropy:

$$H(B) = I_S(B, B) = \sup_{P(A \cap B)} \left\{ I_S(A, B) : \int_A dP(B \mid a)\, dQ(a) = P(B) \right\}$$

where the supremum is taken over all joint probability measures $P(A \cap B)$ such that $P(B)$ is its marginal. This definition shows that entropy $H(B)$ is an information *potential*, because it represents the maximum information that system B with distribution $P(B)$ can communicate about another system. In this context, it is called Boltzmann information, and its supremum $\sup H(B) = \ln |B|$ is called Hartley information.

The relation of entropy to information may help in the analysis of choice under uncertainty. Indeed, lotteries with higher entropy have greater information potential. Thus, although lotteries P and Q in Example 5 have the same expected utilities, their entropies or information potentials are very different. In fact, because lottery Q in Example 5 offers a fixed amount of money with certainty, its entropy is zero. Information may be useful to a decision-maker and therefore may also carry a utility.

4.2 Classical Value of Information

The idea that information may improve the performance of statistical estimation and control systems was developed into a rigorous theory in the mid-1960s by Stratonovich and Grishanin [7, 23–28]. Consider a composite system $A \times B$ with

joint distribution $P(A \cap B) = P(B \mid A) \otimes Q(A)$ and a utility function $u : A \times B \to \mathbb{R}$. For example, A may represent a system to be estimated or controlled, B may represent an estimator or a controller and $u(a, b)$ measures the quality of estimation or control (e.g. a negative error). In game theory, $A \times B$ may represent the set of pure strategies of two players, and $u(a, b)$ a reward function to player B. If there is no information communicated between A and B, then the expected utility $\mathbb{E}_P\{u(a, b)\}$ can be maximized in a standard way by choosing elements $b \in B$ based on the distribution $Q(A)$. On the other hand, if there is complete information (i.e. $a \in A$ is known or observed), then $u(a, b)$ can be maximized by choosing $b \in B$ for each $a \in A$. The *value of Shannon's information amount* λ (or λ-*information*) was defined as the maximum expected utility that can be achieved subject to the constraint that mutual information $I_S(A, B)$ does not exceed λ:

$$\overline{u}_S(\lambda) := \sup_{P(B|A)} \{\mathbb{E}_P\{u(a, b)\} : I_S(A, B) \le \lambda\}.$$

Note that the expected utility and mutual information above are computed using the joint distributions $P(A \cap B) = P(B \mid A) \otimes Q(A)$, while the maximization is over the conditional probabilities $P(B \mid A)$ with the marginal distribution $Q(A)$ considered to be fixed. The subscript in $\overline{u}_S(\lambda)$ denotes that it is the value of information of Shannon type. Stratonovich also defined the value of information $\overline{u}_B(\lambda)$ of Boltzmann type, in which maximization is done with the additional constraint that $P(B \mid A)$ must be a function $f : A \to B$ such that the entropy $H(B) = H(f(A)) \le \lambda$, and value of information $\overline{u}_H(\lambda)$ of Hartley type with the constraint on cardinality $\ln|f(A)| \le \lambda$ [26]. Stratonovich also showed the inequality $\overline{u}_S(\lambda) \ge \overline{u}_B(\lambda) \ge \overline{u}_H(\lambda)$, which follows from the fact that $I_S(A, B) \le H(f(A)) \le \ln|f(A)|$, and proved a theorem about asymptotic equivalence of all types of λ-information (Theorems 11.1 and 11.2 in [26]).

The function $\overline{u}_S(\lambda)$ defines the upper frontier of the expected utility. One may also be interested in the lower frontier (i.e. the worst case scenario) defined similarly using minimization:

$$\underline{u}_S(\lambda) := \inf_{P(B|A)} \{\mathbb{E}_P\{u(a, b)\} : I_S(A, B) \le \lambda\}.$$

Functions $\overline{u}_S(\lambda)$ and $\underline{u}_S(\lambda)$ were referred to in [26] as *normal* and *abnormal* branches of λ-information, representing, respectively, the maximal gain $\overline{u}_S(\lambda) - \overline{u}_S(0) \ge 0$ and the maximal loss $\underline{u}_S(\lambda) - \underline{u}_S(0) \le 0$. Observe that $\underline{u}_S(\lambda) = -\overline{(-u)}_S(\lambda)$[4] (because $\inf u = -\sup(-u)$), which uses the reflection $u(x) \mapsto -u(x)$ to switch between gains and losses, as discussed in Sect. 3.2 (Example 5). It was shown in [26] that the normal branch $\overline{u}_S(\lambda)$ is concave and non-decreasing, while abnormal branch $\underline{u}_S(\lambda)$ is convex and non-increasing. These properties can be used to give the following information-theoretic interpretation of humans' perception of risk.

[4]Note that $\underline{u}_S(\lambda) \ne -\overline{u}_S(\lambda)$ in general, and one of the branches may be empty.

Indeed, lotteries with non-zero entropy have a non-zero information potential, which means that after playing the lottery, information λ may increase or decrease by the amount $\Delta\lambda$. The value of this potential information, however, can be represented either by the normal branch $\overline{u}_S(\lambda)$, if lotteries are associated with gains, or by the abnormal branch $\underline{u}_S(\lambda)$, if lotteries are associated with losses. Using the absolute value $|\lambda|$ in the constraint $I_S \leq |\lambda|$, one can plot the normal branch $\overline{u}_S(\lambda)$ against 'positive' information $\lambda \geq 0$, associated with gains, while the abnormal branch $\underline{u}_S(\lambda)$ against 'negative' information $\lambda \leq 0$, associated with losses. The graph of the resulting function is shown in Fig. 1, and it is similar to the S-shaped value function in prospect theory [9], because $\overline{u}_S(\lambda)$ is concave and $\underline{u}_S(\lambda)$ is convex. The normal branch implies risk-aversion in choices associated with gains, because the potential increase $\overline{u}_S(\lambda + \Delta\lambda) - \overline{u}_S(\lambda)$ associated with $\Delta\lambda$ is less than the potential decrease $\overline{u}_S(\lambda) - \overline{u}_S(\lambda - \Delta\lambda)$. On the other hand, convexity of the abnormal branch $\underline{u}_S(\lambda)$ implies risk-taking in choices associated with losses, because the potential increase $\underline{u}_S(\lambda) - \underline{u}_S(\lambda + \Delta\lambda)$ is greater than potential decrease $\underline{u}_S(\lambda - \Delta\lambda) - \underline{u}_S(\lambda)$ (here, we assume $\lambda \leq 0$ as in Fig. 1).

Unfortunately, this explanation may appear simply as a curious coincidence, because proofs that $\overline{u}_S(\lambda)$ is concave and $\underline{u}_S(\lambda)$ is convex are usually based on very specific assumptions about information, such as convexity and differentiability of Shannon's information $I_S(A, B)$ as a functional of probability measures. It can be shown, however, that the discussed properties of λ-information hold in a more general setting, when information is understood more abstractly [2], and they follow only from the linearity of the expected utility, that is from axioms (1) and (2).

4.3 Value of Abstract Information

In this section, we discuss generalizations of the concept of information and show that the corresponding value functions have concave and convex branches. Recall that the definition of Shannon's information, as well as entropy, involves a very specific functional—the Kullback-Leibler divergence $D_{KL}(P, Q)$ [12]. If P and Q are two probability measures defined on the same σ-ring $\mathcal{R}(\Omega)$ of subsets of Ω and P is absolutely continuous with respect to Q, then KL-divergence of Q from P is the expectation $\mathbb{E}_P\{\ln(P/Q)\}$:

$$D_{KL}(P, Q) := \int_{\Omega} \left[\ln \frac{dP(\omega)}{dQ(\omega)} \right] dP(\omega).$$

It plays the role of a distance between distributions, because $D_{KL}(P, Q) \geq 0$ for all P, Q and $D_{KL}(P, Q) = 0$ if and only if $P = Q$, but it is not a metric (in general, symmetry and the triangle inequality do not hold). The unique property

of the KL-divergence is that it satisfies the axiom of additivity of information from independent sources [10]:

$$D_{KL}(P_1 \otimes P_2, Q_1 \otimes Q_2) = D_{KL}(P_1, Q_1) + D_{KL}(P_2, Q_2).$$

One can see that Shannon's mutual information $I_S(A, B)$ is the KL-divergence of the prior distribution $P(B)$ from posterior $P(B \mid A)$ (or equivalently of the product $Q(A) \otimes P(B)$ of marginals from the joint distribution $P(A \cap B)$). Entropy can be interpreted as negative KL-divergence $-D_{KL}(P, \mu)$ of some reference measure μ (e.g. the Lebesgue measure on Ω) from P. One way to generalize the notion of information is to consider other information distances.

By a *distance* one understands a non-negative function $D : Y \times Y \to \mathbb{R} \cup \{\infty\}$ such that $y = z$ implies $D(y, z) = 0$. When D is restricted to the set $\mathcal{P}(\Omega) \subset Y$ of probability measures, we refer to it as an *information distance*. If a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$ is minimized at y_0, then the distance $D(y, y_0)$ can be defined by the non-negative difference $F(y) - F(y_0)$. More generally, a distance associated with F can be defined as follows:

Definition 3 (F-information distance). A restriction to $\mathcal{P}(\Omega) \subset Y$ of function $D_F : Y \times Y \to \mathbb{R} \cup \{\infty\}$ associated with a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$ as follows:

$$D_F(y, z) := \inf\{F(y) - F(z) - x(y - z) : x \in \partial F(z)\}, \tag{7}$$

where $\partial F(z) := \{x \in X : x(y - z) \leq F(y) - F(z),\ \forall y \in Y\}$ is subdifferential of F at z. We define $D_F(y, z) = \infty$ if $\partial F(z) = \varnothing$ or $F(y) = \infty$.

It follows immediately from the definition of $\partial F(z)$ that $D_F(y, z) \geq 0$. We note also that the notion of subdifferential can be applied to a non-convex function F. However, nonempty $\partial F(z)$ implies $F(z) < \infty$ and $F(z) = F^{**}(z)$, $\partial F(z) = \partial F^{**}(z)$ ([19], Theorem 12). Generally, $F^{**} \leq F$, so that $F(y) - F(z) \geq F^{**}(y) - F^{**}(z)$ if $\partial F(z) \neq \varnothing$. If F is Gâteaux differentiable at z, then $\partial F(z)$ has a single element $x = \nabla F(z)$, called the *gradient* of F at z. One can see that distance (7) is a generalization of the Bregman divergence for the case of a non-convex and non-differentiable F. The KL-divergence is a particular example of Bregman divergence associated with strictly convex and smooth functional $F(y) = \int (\ln y - 1)\, dy$. Thus, an information constraint can be understood geometrically as constraint $D_F(P, Q) \leq \lambda$ on some F-information distance, and the value of information has the following geometric interpretation.

Let X and Y be two linear spaces in duality, and let $x(y) = \int x\, dy$ be a linear functional on Y. Recall that the *support function* of set $C \subseteq Y$ is sublinear mapping $sC : X \to \mathbb{R} \cup \{\infty\}$ defined as

$$sC(x) := \sup\{x(y) : y \in C\}.$$

Because expected utility $\mathbb{E}_P\{u\}$ is the restriction to $\mathcal{P}(\Omega) \subset Y$ of linear functional $u(y) = \int u\, dy$, the value of Shannon's mutual information $\overline{u}_S(\lambda)$ coincides with

the support function $sC(\lambda)(u)$ of set $C(\lambda) \subseteq \mathcal{P}(\Omega)$, defined by the information constraint $I_S(A,B) \leq \lambda$ and evaluated at $u \in X$ corresponding to the utility function $u : \Omega \to \mathbb{R}$. Another way to define subsets $C(\lambda)$ is based on the notion of information *resource* [2].

Let $\{C(\lambda)\}_{\lambda\in\mathbb{R}}$ be a family of non-empty closed sets such that $C(\lambda_1) \subseteq C(\lambda_2)$ for any $\lambda_1 \leq \lambda_2$. Then the support $sC(\lambda)(x)$ is a non-decreasing function of λ. The union of all sets $C(\lambda) \times [\lambda, \infty)$ is the epigraph of some closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$. In fact, this functional can be defined as $F(y) = \inf\{\lambda : y \in C(\lambda)\}$. Then each closed set $C(\lambda)$ is a sublevel set $\{y : F(y) \leq \lambda\}$.

Definition 4 (Information resource). A restriction to $\mathcal{P}(\Omega) \subset Y$ of a closed functional $F : Y \to \mathbb{R} \cup \{\infty\}$.

A generalized notion of the value of information is given by the support function of subsets $C(\lambda) \subseteq \mathcal{P}(\Omega)$, defined by constraints either on F-information distance from some reference point or on an information resource:

$$\overline{u}(\lambda) := \sup\{u(y) : F(y) \leq \lambda\},$$
$$\underline{u}(\lambda) := \inf\{u(y) : F(y) \leq \lambda\}.$$

Properties of the above value functions were studied in [2]. In particular, the following result was proven (Proposition 3, [2]).

Theorem 3. *Function $\overline{u}(\lambda)$ is strictly increasing and concave. Function $\underline{u}(\lambda)$ is strictly decreasing and convex.*

Here we outline the proof assuming the reader has some knowledge of convex analysis (e.g. see [19, 29] for references). See Propositions 1–3 in [2] for more details.

Proof. Variational problem $\overline{u}(\lambda) = \sup\{u(y) : F(y) \leq \lambda\}$ is solved using the method of Lagrange multipliers. The Lagrange function is

$$K(y, \beta^{-1}) = u(y) + \beta^{-1}[\lambda - F(y)],$$

where β^{-1} is the Lagrange multiplier associated with constraint $F(y) \leq \lambda$. The necessary conditions of extremum of $K(y, \beta^{-1})$ are

$$\bar{y}(\beta) \in \partial F^*(\beta u), \qquad F(\bar{y}(\beta)) = \lambda,$$

where $\partial F^*(x) := \{y \in Y : y(z-x) \leq F^*(z) - F^*(x),\ \forall z \in X\}$ is subdifferential of the dual functional $F^*(x) = \sup\{x(y) - F(y)\}$. If the convex closure $\operatorname{co}\operatorname{cl}\{y : F(y) \leq \lambda\}$ of the sublevel set of F coincides with the sublevel set $\{y : F^{**}(y) \leq \lambda\}$ of its bi-dual F^{**}, then the above conditions are also sufficient.

The function $\beta^{-1}(\lambda)$ is the derivative $d\overline{u}(\lambda)/d\lambda$, because $\overline{u}(\lambda) = u(\bar{y}) + \beta^{-1}[\lambda - F(\bar{y})]$. Also, $\beta^{-1} = d\overline{u}(\lambda)/d\lambda \geq 0$, because $\overline{u}(\lambda)$ is non-decreasing. In fact, $\beta^{-1} = 0$ if and only if $\lambda = \sup F(y)$, so that $\overline{u}(\lambda)$ is strictly increasing.

The fact that $\overline{u}(\lambda)$ is concave is proven by showing that its derivative $\beta^{-1}(\lambda)$ is non-increasing. Consider two solutions $\bar{y}(\beta_1)$, $\bar{y}(\beta_2)$ for $\lambda_1 \leq \lambda_2$. Because $\bar{y}(\beta_i) \in \partial F^*(\beta_i u)$ and ∂F^* is a monotone operator, we have

$$(\beta_2 - \beta_1)u(\bar{y}(\beta_2) - \bar{y}(\beta_1)) \geq 0.$$

The difference $u(\bar{y}(\beta_2) - \bar{y}(\beta_1)) \geq 0$, because $\overline{u}(\lambda) = u(\bar{y}(\beta))$ is nondecreasing. Therefore, $\beta_2 - \beta_1 \geq 0$, which proves that $\beta^{-1}(\lambda)$ is nonincreasing.

The strictly decreasing and convex properties of $\underline{u}(\lambda)$ follow from the fact that $\underline{u}(\lambda) = -\overline{(-u)}(\lambda)$, and $\overline{(-u)}(\lambda)$ is strictly increasing and concave, as was shown above. □

5 Discussion

In this paper, we have reviewed mathematical arguments for the expected utility theory and some behavioural arguments against it. Our hope is that by the end of Sect. 3 the reader was sufficiently intrigued by the paradox following from the conflict between the logic and structure of the axiomatic theory of utility on one hand, and our own behaviour that appears to contradict it on the other. Perhaps the key to this puzzle is in the fact that the contradiction occurs only when humans are presented with the problems, and humans are information-seeking agents. Our mind has evolved to learn and adapt to new information, and this suggests we need to take potential information (entropy) into account.

Analysis shows that the value of information is an S-shaped function, which mirrors some of the ideas of prospect theory [9], and therefore the value of information theory may explain humans' attitude to risk. Unlike the descriptive nature of the value function for prospects, however, properties of the value of information are based on rigorous results. Furthermore, because the value of information is defined as conditional extremum of expected utility, this normative theory does not contradict the axioms of expected utility. Rather, it generalizes the von Neumann and Morgenstern theory by adding a non-linear component that reflects the agent's preferences about potential information.

Acknowledgements This work was supported by UK EPSRC grant EP/H031936/1.

References

1. Allais, M.: Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'École americaine. Econometrica **21**, 503–546 (1953)
2. Belavkin, R.V.: Optimal measures and Markov transition kernels. J. Global Optim. **55**, 387–416 (2013)

3. Bernoulli, D.: Commentarii acad. Econometrica **22**, 23–36 (1954)
4. Bourbaki, N.: Eléments de mathématiques. Intégration. Hermann, Paris (1963)
5. Debreu, G.: Representation of a preference relation by a numerical function. In: Thrall, R.M., Coombs, C.H., Davis, R.L. (eds.) Decision Process. Wiley, New York (1954)
6. Ellsberg, D.: Risk, ambiguity, and the Savage axioms. Q. J. Econom. **75**(4), 643–669 (1961)
7. Grishanin, B.A., Stratonovich, R.L.: Value of information and sufficient statistics during an observation of a stochastic process (in Russian). Izvestiya USSR Acad. Sci. Tech. Cybern. **6**, 4–14 (1966)
8. Huck, S., Müller, W.: Allais for all: revisiting the paradox in a large representative sample. J. Risk Uncertainty **44**(3), 261–293 (2012)
9. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**(2), 263–292 (1979)
10. Khinchin, A.I.: Mathematical Foundations of Information Theory. Dover, New York (1957)
11. Kreps, D.M.: Notes on the Theory of Choice. Westview Press, Colorado (1988)
12. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
13. List, J.A., Haigh, M.S.: A simple test of expected utility theory using professional traders. PNAS **102**(3), 945–948 (2005)
14. Loomes, G., Sugden, R.: Regret theory: an alternative theory of rational choice under uncertainty. Econom. J. **92**(368), 805–824 (1982)
15. Machina, M.J.: "Expected utility" Analysis without the independence axiom. Econometrica **50**(2), 277–323 (1982)
16. Machina, M.J.: States of the world and the state of decision theory, chap. 2. In: Meyer, D. (ed.) The Economics of Risk. W. E. Upjohn Institute for Employment Research, Kalamazoo (2003)
17. Machina, M.J.: Nonexpected utility theory. In: Teugels, J.L., Sundt, B. (eds.) Encyclopedia Of Actuarial Science, vol. 2, pp. 1173–1179. Wiley, Chichester (2004)
18. von Neumann, J., Morgenstern, O.: Theory of games and economic behavior, first edn. Princeton University Press, Princeton (1944)
19. Rockafellar, R.T.: Conjugate duality and optimization. In: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 16. Society for Industrial and Applied Mathematics, Philadelphia (1974)
20. Savage, L.: The Foundations of Statistics. Wiley, New York (1954)
21. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
22. Starmer, C.: Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. J. Econ. Liter. pp. 332–382 (2000)
23. Stratonovich, R.L.: On value of information (in Russian). Izvestiya USSR Acad. Sci. Tech. Cybern. **5**, 3–12 (1965)
24. Stratonovich, R.L.: Value of information during an observation of a stochastic process in systems with finite state automata (in Russian). Izvestiya USSR Acad. Sci. Tech. Cybern. **5**, 3–13 (1966)
25. Stratonovich, R.L.: Extreme problems of information theory and dynamic programming (in Russian). Izvestiya USSR Acad. Sci. Tech. Cybern. **5**, 63–77 (1967)
26. Stratonovich, R.L.: Information Theory (in Russian). Sovetskoe Radio, Moscow (1975)
27. Stratonovich, R.L., Grishanin, B.A.: Value of information when an estimated random variable is hidden (in Russian). Izvestiya USSR Acad. Sci. Tech. Cybern. **3**, 3–15 (1966)
28. Stratonovich, R.L., Grishanin, B.A.: Game-theoretic problems with information constraints (in Russian). Izvestiya USSR Acad. Sci. Tech. Cybern. **1**, 3–12 (1968)
29. Tikhomirov, V.M.: Analysis II,chap. In: Convex Analysis. Encyclopedia of Mathematical Sciences, vol. 14, pp. 1–92. Springer, New York (1990)
30. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. Science **211**, 453–458 (1981)
31. Wald, A.: Statistical Decision Functions. Wiley, New York (1950)

A Risk-Averse Differential Game Approach to Multi-agent Tracking and Synchronization with Stochastic Objects and Command Generators

Khanh Pham and Meir Pachter

Abstract This chapter presents the formulation of a class of distributed stochastic multi-agent systems where local interconnections among cautious and defensive decision makers and/or trackers are supported by connectivity graphs. Associated with autonomous decision makers and/or trackers are finite-horizon performance measures for conflict-free coordination and cohesive object and/or command tracking. The current analysis is limited to the class of distributed linear stochastic systems and measurement subsystems. It is shown that optimal rules for ordering uncertain prospects are feasible for all self-directed decision makers and/or trackers with output-feedback Nash decision making and risk-averse utility functions.

Keywords Distributed Control • Performance-Measure Statistics • Downside Performance Risk Measure • Connectivity Graphs • Person-by-Person Decision and Control

1 Introduction

One of the best ways to understand the growing interest in multi-agent tracking and distributed control systems is to review the history of these old and new engineering problems. Examples include multi-agent architectures for tracking and estimation [7, 10, 15] and control design with pre-specified information structures and control under communication constraints [8]. Yet relatively little work has focused on understanding quantitatively the downside risk measures in multi-agent systems for various tasks in terms of performance robustness and risk aversion.

In noncooperative stochastic games and distributed controls, there are more than two capable decision makers who optimize different goals and utilities. Each

K. Pham (✉)
Air Force Research Laboratory, Kirtland A.F.B., NM 87117, Rome, NewYork
e-mail: AFRL.RVSV@kirtland.af.mil

M. Pachter
Air Force Institute of Technology, Wright-Patterson A.F.B., OH 45433, USA

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_2

decision maker wishes to influence to his/her advantage a shared interaction process by exerting his/her control decisions. To the best knowledge of the authors, most studies, e.g., [2] and [4], have mainly concentrated on the selection of open- and/or closed-loop Nash strategy equilibria in accordance of expected utilities under the structural constraints of linear system dynamics, quadratic cost functionals, and additive independent white Gaussian noises corrupting the system dynamics and measurements. Very little work, if any, has been published on the subject of higher-order assessment of performance uncertainty and risks beyond expected performance.

For this reason attention in the research investigation that follows is directed primarily toward a linear-quadratic class of noncooperative stochastic games and/or distributed controls, which in turn has linear system dynamics, quadratic rewards and/or costs, and independent white zero-mean Gaussian noises additively corrupting the system dynamics and output measurements. Notice that, under these conditions, the quadratic rewards or costs are random variables with the generalized chi-squared probability distributions. If a measure of uncertainty such as the variance of the possible rewards or costs was used in addition to the expected reward or costs, the decision makers should be able to correctly order preferences for alternatives. This claim seems plausible, but it is not always correct. Various investigations have indicated that any evaluation scheme based on just the expected reward or cost and reward/cost variance would necessarily imply indifference between some courses of action; therefore, no criterion based solely on the two attributes of means and variances can correctly represent their preferences. See [14] and [9] for further details.

The present research contributions include significant extensions of the existing results [11] toward some completely unexplored areas as such: i) the design of distributed filtering via private observations for self-directed decision makers and/or autonomous controllers with distributed noisy information structures about the uncertain interaction process; ii) an efficiently computational procedure for all the mathematical statistics associated with the generalized chi-squared rewards/costs when respective mean-risk aware utilities are formed; and iii) the synthesis of distributed risk-sensitive decision policies with output feedback for distributed noncooperative solutions of Nash type that now guarantee performance robustness with certainty much stronger than ensemble averaging measures of performance.

The remainder of the chapter is organized as follows: In Sect. 2, the setting which involved necessary background and terminologies associated with a class of distributed multi-agent tracking and synchronization is provided. The purpose of Sect. 3 is to continue the discussion of the research development in using preferences of risk, dynamic game decision optimization, and distributed decision making with local output-feedback measurements tailored toward the worst-case scenarios. The feasibility of person-by-person risk-averse strategies supported by distributed Kalman-like estimators is subsequently put forward in Sect. 4. Some final remarks are given in Sect. 5.

2 The Setting

In this section, some preliminaries are in order. For instance, a fixed probability space with filtration is denoted by $(\Omega, \mathbb{F}, \{\mathbb{F}_{t_0,t} : t \in [t_0, t_f], \mathbb{P}\})$ where all filtrations are right continuous and complete. In addition, $L^2_{\mathbb{F}_{t_f}}([t_0, t_f]; \mathbb{R}^n)$ denotes the space of $\mathbb{F}_{t_f}$-adapted random processes $\{z(t) : t \in [t_0, t_f]\}$ such that $E\{\int_{[t_0,t_f]} ||z(t)||^2_{\mathbb{R}^n} dt\} < \infty$ and $\mathbb{F}_{t_f} \triangleq \{\mathbb{F}_{t_0,t} : t \in [t_0, t_f]\}$.

2.1 Distributed Multi-agent Tracking and Synchronization

As for a model specification, there is a stochastic object or command generator that evolved in the fixed probability space $(\Omega, \mathbb{F}, \{\mathbb{F}_{t_0,t} : t \in [t_0, t_f], \mathbb{P}\})$ and is subject to the following stochastic dynamical decision system

$$dx_o(t) = Ax_o(t)dt + G_o dw_o(t), \quad x_o(t_0) \tag{1}$$

where the initial state $x_o(t_0) = x_o^0$, the state space is in $\mathbb{R}^{n_0}$, and the exogenous state noise $\{w_o(t) : t \in [t_0, t_f]\}$ is an $\mathbb{R}^{n_o}$-valued stationary Wiener process adapted to $\mathbb{F}_{t_f}$, independent of $x_o(t_0)$ and having the correlation of independent increments

$$E\left\{[w_o(\tau_1) - w_o(\tau_2)][w_o(\tau_1) - w_o(\tau_2)]^T\right\} = W_o|\tau_1 - \tau_2|, \quad \forall\ \tau_1, \tau_2 \in [t_0, t_f].$$

Moreover, there are N identical decision makers and/or trackers which are also described by

$$dx_{ti}(t) = (Ax_{ti}(t) + Bu_{ti}(t))dt + Gdw_{ti}(t), \quad x_{ti}(t_0). \tag{2}$$

Of note, each stochastic dynamical decision system i and $i \in \overline{N} \triangleq \{1, \ldots, N\}$ has an initial state $x_{ti}(t_0) = x_{ti}^0$, state space $\mathbb{R}^{n_{ti}}$, an action space $\mathbb{A}^i \subset \mathbb{R}^{m_i}$, and an exogenous state noise space $\{w_{ti}(t) : t \in [t_0, t_f]\}$ defined by an $\mathbb{R}^{p_{ti}}$-valued stationary Wiener process adapted to $\mathbb{F}_{t_f}$, independent of $x_{ti}(t_0)$ and having the correlation of independent increments

$$E\left\{[w_{ti}(\tau_1) - w_{ti}(\tau_2)][w_{ti}(\tau_1) - w_{ti}(\tau_2)]^T\right\} = W_{ti}|\tau_1 - \tau_2|, \quad \forall\ \tau_1, \tau_2 \in [t_0, t_f].$$

The decentralized partial information structure available to decision maker i or u_{ti} is generated by noisy relative observation

$$dy_{ti}(t) = C(x_{ti}(t) - x_o(t))dt + Hdv_{ti}(t), \quad i \in \overline{N} \tag{3}$$

where the exogenous measurement noise is an $\mathbb{R}^{q_{ti}}$-valued stationary Wiener process adapted to $\mathbb{F}_{t_f}$ and independent of $\{w_{ti}(t) : t \in [t_0, t_f]\}$ with the correlation of independent increments

$$E\left\{[v_{ti}(\tau_1) - v_{ti}(\tau_2)][v_{ti}(\tau_1) - v_{ti}(\tau_2)]^T\right\} = V_{ti}|\tau_1 - \tau_2|\,, \quad \forall\ \tau_1, \tau_2 \in [t_0, t_f].$$

A concern has grown in the relative states, e.g., $x_i \triangleq x_{ti} - x_o$ that the evolutions are then described by

$$dx_i(t) = (Ax_i(t) + Bu_{ti}(t))dt + Gdw_{ti}(t) - G_o dw_o(t)\,, \quad x_i(t_0) = x_i^0 \tag{4}$$

with the local noisy observations

$$dy_{ti}(t) = Cx_i(t)dt + Hdv_{ti}(t)\,, \quad i \in \overline{N}\,. \tag{5}$$

With the advent of connectivity graphs widely used in cooperative control and formation among unmanned systems [5] and [16], the role of formation graphs supporting networks of local decision makers or trackers herein has also been significant. To this end, a vertex of the graph corresponds to a decision maker or tracker and the edges of the graph convey the dependence of the interconnections. For instance, a directed graph $\mathcal{G}^i = (\mathcal{V}^i, \mathcal{E}^i)$ associated with decision maker or tracker i consists of a set of vertices $\mathcal{V}^i \triangleq \{v_{i_1}, \ldots, v_{i_{N_i}}\}$, indexed by local decision makers or trackers in the N_i-neighborhood and a set of edges $\mathcal{E}^i \triangleq \{(v_{i_1}, v_{i_2}) \in \mathcal{V}^i \times \mathcal{V}^i\}$, containing ordered pairs of distinct vertices.

As part of the effort to approach distributed multi-agent tracking and synchronization, a local neighborhood of N_i immediate decision makers (or trackers) associated with decision maker (or tracker) i and supported by an appropriate directed graph includes two key elements. First, the augmented vectors are sought to be

$$z_i \triangleq \begin{bmatrix} x_{i_1} \\ \vdots \\ x_{i_{N_i}} \end{bmatrix}, \quad w_i \triangleq \begin{bmatrix} w_{ti_1} \\ \vdots \\ w_{ti_{N_i}} \\ w_o^T \end{bmatrix}, \quad v_i \triangleq \begin{bmatrix} v_{ti_1} \\ \vdots \\ v_{ti_{N_i}} \end{bmatrix}, \quad i \in \overline{N}.$$

Second, the Kronecker product of matrices are defined as follows:

$$A_{N_i} \triangleq I_{N_i \times N_i} \otimes A\,, \qquad G_{N_i} \triangleq \left[\, I_{N_i \times N_i} \otimes G \ -1_{N_i} \otimes G_o \,\right]$$

$$B_i \triangleq \left[0 \ldots 0\ 1\ 0 \ldots 0\right]^T \otimes B\,, \qquad C_{N_i} \triangleq I_{N_i \times N_i} \otimes C$$

$$H_{N_i} \triangleq I_{N_i \times N_i} \otimes H$$

where $I_{N_i \times N_i}$ is the $N_i \times N_i$ identity matrix and $1_{N_i} \triangleq \left[1 \ldots 1\right]^T$ is the column vector of one of size N_i and will result in the distributed stochastic system dynamics with controls and observations from decision maker or tracker i and all of its immediate neighbors $\overline{N}_i$ and $\overline{N}_i \triangleq \{i_1, \ldots, i_{N_i}\}$

$$dz_i(t) = (A_{N_i} z_i(t) + B_i u_{ti}(t) + \sum_{j \in \overline{N}_i} e_{ij} B_{ij}(t) u_{tj}(t))dt + G_{N_i} dw_i(t) \tag{6}$$

$$dy_i(t) = C_{N_i} z_i(t)dt + H_{N_i} dv_i(t) \tag{7}$$

where e_{ij} is the edge weights and $z_i(t_0)$ is the initial system state with the value of $z_i^0 \triangleq \left[x_{ti_1}^T(t_0) \ldots x_{ti_{N_i}}^T(t_0)\right]^T$.

Continuing the practice of private observations, each decision maker or tracker i can presumably observe a noise corrupted version of all best responses $\sum_{j \in \overline{N}_i} e_{ij} B_{ij}(t) u_{tj}(t)$ from the immediate neighbors

$$du_{-ti}(t) \triangleq u_{-ti}(t)dt = \sum_{j \in \overline{N}_i} e_{ij} B_{ij}(t) u_{tj}(t)dt + d\eta_i(t)\,. \tag{8}$$

Notice that decision makers or trackers i operate within their own noisy environments modeled by the uncorrelated p_i-, n_i-, and q_i-dimensional stationary Wiener processes adapted for $[t_0, t_f]$

$$E\left\{[w_i(\tau_1) - w_i(\tau_2)][w_i(\tau_1) - w_i(\tau_2)]^T\right\} = W_i|\tau_1 - \tau_2|$$
$$E\left\{[\eta_i(\tau_1) - \eta_i(\tau_2)][\eta_i(\tau_1) - \eta_i(\tau_2)]^T\right\} = M_i|\tau_1 - \tau_2|$$
$$E\left\{[v_i(\tau_1) - v_i(\tau_2)][v_i(\tau_1) - v_i(\tau_2)]^T\right\} = V_i|\tau_1 - \tau_2|$$

whose a priori second-order statistics $W_i \triangleq diag(W_{ti_1}, \ldots, W_{ti_{N_i}}, W_o) > 0$, $M_i > 0$, and $V_i \triangleq diag(V_{ti_1}, \ldots, V_{ti_{N_i}}) > 0$ for $i \in \overline{N}$ are also assumed known.

For decentralized filtering, each decision maker or tracker i has σ-algebras

$$\mathbb{F}_{t_0,t}^i \triangleq \sigma\{z_i(t_0), w_i(s), v_i(s) : t_0 \le s \le t\} \tag{9}$$

$$\mathcal{G}_{t_0,t}^{y^{i,u}} \triangleq \sigma\{y_i(s) : t_0 \le s \le t\}, \quad t \in [t_0, t_f], \quad i \in \overline{N} \tag{10}$$

and the minimum σ-algebras generated by (9)–(10) are therefore given by

$$\mathbb{F}_{t_0,t} \triangleq \vee_{i=1}^N \mathbb{F}_{t_0,t}^i \tag{11}$$

$$\mathcal{G}_{t_0,t}^{y^u} \triangleq \vee_{i=1}^N \mathcal{G}_{t_0,t}^{y^{i,u}}\,. \tag{12}$$

2.2 Distributed Decision/Control and Filtering

As a critical element of the effort to move toward the distributed decision strategies, $\mathcal{G}_{t_f}^{y^{i,u}} \triangleq \{\mathcal{G}_{t_0,t}^{y^{i,u}} : t \in [t_0, t_f]\} \subset \{\mathbb{F}_{t_0,t} : t \in [t_0, t_f]\}$ is denoted for the information available to decision maker and/or tracker i and $i \in \overline{N}$. The admissible set of distributed feedback strategies for decision maker and/or tracker i is defined by

$$\mathbb{U}^{y^{i,u}}[t_0, t_f] \triangleq \{u_{ti} \in L^2_{\mathcal{G}_{t_f}^{y^{i,u}}}([t_0, t_f], \mathbb{R}^{m_i}) : u_{ti}^t \in \mathbb{A}^i \subset \mathbb{R}^{m_i}, \text{almost everywhere}$$
$$t \in [t_0, t_f], \mathbb{P} - \text{almost surely}\}, \quad i \in \overline{N} \tag{13}$$

where $\mathbb{U}^{y^{i,u}}[t_0, t_f]$ is a closed convex subset of $L^2_{\mathbb{F}_{t_f}}([t_0, t_f]; \mathbb{R}^{m_i})$ for $i \in \overline{N}$.

At this point, each of the N distributed filters whose outputs are the state estimates $\hat{z}_i(t) \triangleq E\{z_i(t)|\mathcal{G}_{t_0,t}^{y^u}\}$ of (6) has the form

$$d\hat{z}_i(t) = (A_{N_i}(t)\hat{z}_i(t) + B_i(t)u_{ti}(t) + u_{-ti}(t))dt$$
$$+L_i(t)(dy_i(t) - C_{N_i}\hat{z}_i(t)dt), \quad \hat{z}_i(t_0) = z_i^0 \tag{14}$$

where the local filter gain $L_i(t)$ is given by

$$L_i(t) = \Sigma_i(t)C_{N_i}^T(H_{N_i}V_iH_{N_i})^{-1} \tag{15}$$

and the estimate error covariance $\Sigma_i(t) \triangleq E\{[z_i(t) - \hat{z}_i(t)][z_i(t) - \hat{z}_i(t)]^T|\mathcal{G}_{t_0,t}^{y^u}\}$

$$\frac{d}{dt}\Sigma_i(t) = A_{N_i}\Sigma_i(t) + \Sigma_i(t)A_{N_i}^T + G_{N_i}W_iG_{N_i}^T + M_i$$
$$-\Sigma_i(t)C_{N_i}^T(H_{N_i}V_iH_{N_i})^{-1}C_{N_i}\Sigma_i(t), \quad \Sigma_i(t_0) = 0. \tag{16}$$

In the background is the substitution of (6), (8), and (14) in a setting shaped by the estimate errors, e.g., $\tilde{z}_i(t) \triangleq z_i(t) - \hat{z}_i(t)$. Thus, it can be shown that

$$d\tilde{z}_i(t) = (A_{N_i} - L_i(t)C_{N_i})\tilde{z}_i(t)dt$$
$$+G_{N_i}dw_i(t) - L_i(t)H_{N_i}dv_i(t) - d\eta_i(t), \quad \tilde{z}_i(t_0) = 0. \tag{17}$$

2.3 Person-by-Person Performance Measure

Recall that decision maker or tracker i is assumed to act purely on the basis of his own information, e.g.,

$$\mathcal{G}_{t_f}^{y^{i,u}} \triangleq \{\mathcal{G}_{t_0,t}^{y^{i,u}} : t \in [t_0, t_f]\} \subset \{\mathbb{F}_{t_0,t} : t \in [t_0, t_f]\}.$$

And the set of admissible decentralized feedback policies $\mathbb{U}^{y^{i,u}}[t_0, t_f]$ is a closed convex subset of $L^2_{\mathbb{F}_{t_f}}([t_0, t_f]; \mathbb{R}^{m_i})$ for $i \in \overline{N}$. The objective of distributed multi-agent tracking and synchronization is then to regulate the dynamical states of all the decision makers or trackers to those of stochastic command generators or objects while being subject to transient trade-offs between the state regulatory and effectiveness of decision policies and/or control inputs.

Associated with each admissible 2-tuple $(u_{ti}(\cdot), u_{-ti}(\cdot))$ is the person-by-person performance measure with the generalized chi-squared type for each decision maker and/or tracker i defined as

$$J_i(u_{ti}, u_{-ti}) = g_i(t_f, z_i(t_f)) + \int_{t_0}^{t_f} C_i(\tau, z_i(\tau), u_{ti}(\tau), u_{-ti}(\tau)) d\tau\,, \quad i \in \overline{N} \tag{18}$$

where the cohesive tracking and regulation criteria are given by

$$g_i(t_f, z_i(t_f)) = \sum_{(v_{i_r}, v_{i_s}) \in \mathcal{E}^i} w_{i_r i_s} ||x_{ti_r}(t_f) - x_{ti_s}(t_f)||^2 + ||x_{ti}(t_f)||^2_{S_{if}} = ||z_i(t_f)||^2_{Q_{if}}$$

and

$$C_i(\tau, z_i(\tau), u_{ti}(\tau), u_{-ti}(\tau)) = \sum_{(v_{i_r}, v_{i_s}) \in \mathcal{E}^i} v_{i_r i_s} ||x_{ti_r}(\tau) - x_{ti_s}(\tau)||^2$$
$$+ ||x_{ti}(\tau)||^2_{S_i} + ||u_{ti}(\tau)||^2_{R_i} = ||z_i(\tau)||^2_{Q_i} + ||u_{ti}(\tau)||^2_{R_i}$$

provided that the design parameters S_{if}, S_i, and R_i are positive semidefinite with R_i invertible and

$$Q_{if} = \hat{D}_i \hat{W}_{if} \hat{D}_i^T + diag(0_{n_{ti} \times n_{ti}}, \ldots, 0_{n_{ti} \times n_{ti}}, S_{if}, 0_{n_{ti} \times n_{ti}}, \ldots, 0_{n_{ti} \times n_{ti}})$$
$$Q_i = \hat{D}_i \hat{W}_i \hat{D}_i^T + diag(0_{n_{ti} \times n_{ti}}, \ldots, 0_{n_{ti} \times n_{ti}}, S_i, 0_{n_{ti} \times n_{ti}}, \ldots, 0_{n_{ti} \times n_{ti}})$$
$$\hat{D}_i = D_i \otimes I_{n_{ti} \times n_{ti}}; \quad \hat{W}_i = W_i \otimes I_{n_{ti} \times n_{ti}}; \quad \hat{W}_{if} = W_{if} \otimes I_{n_{ti} \times n_{ti}}$$
$$W_{if} = diag(w_{i_r i_s}) \text{ of dimension } |\mathcal{E}^i|; \quad W_i = diag(v_{i_r i_s}) \text{ of dimension } |\mathcal{E}^i|$$
$$D_i = \text{incidence matrix of the directed graph } \mathcal{G}^i(\mathcal{V}^i, \mathcal{E}^i) \text{ withsize } N_i \times |\mathcal{E}^i|\,.$$

2.4 Person-by-Person Decision and/or Control Policies

The realization of admissible feedback policies is discussed next. In the case of incomplete information, an admissible feedback policy u_{ti} for a local best response

to relevant immediate decision makers or trackers u_{-ti} must be of the form, for some $\eth^i(\cdot,\cdot)$,

$$u_{ti}(t) = \eth^i(t, y_i(\tau)), \quad \tau \in [t_0, t], \quad i \in \overline{N}. \tag{19}$$

In general, the conditional density $p^i(z_i(t)|\mathcal{G}^{y^u}_{t_0,t})$, which is the density of $z_i(t)$ conditioned on $\mathcal{G}^{y^u}_{t_0,t}$ (i.e., induced by the observation $\{y_i(\tau) : \tau \in [t_0, t]\}$), represents the sufficient statistics for describing the conditional stochastic effects of future feedback policy u_{ti}. Under the linear-Gaussian assumption the conditional density $p^i(z_i(t)|\mathcal{G}^{y^u}_{t_0,t})$ is parameterized by the locally available state estimate $\hat{z}_i(t)$ and estimate error covariance $\Sigma_i(t)$. In addition, $\Sigma_i(t)$ is independent of feedback policy $u_{ti}(t)$ and observations $\{y_i(\tau) : \tau \in [t_0, t]\}$. Henceforth, to look for an optimal control and/or decision policy $u_{ti}(t)$ of the form (19), it is only required that

$$u_{ti}(t) = \gamma^i(t, \hat{z}_i(t)), \quad t \in [t_0, t_f], \quad i \in \overline{N}.$$

Given the linear-quadratic properties of the distributed multi-agent tracking and synchronization problem governed by (6), (7), and (18), the search for an optimal feedback solution is productively restricted to a linear time-varying feedback policy generated from the locally accessible state $\hat{z}_i(t)$ by

$$u_{ti}(t) = K^i(t)\hat{z}_i(t), \quad t \in [t_0, t_f], \quad i \in \overline{N} \tag{20}$$

with $K^i \in C([t_0, t_f]; \mathbb{R}^{m_i \times n_i})$ an admissible feedback form whose further defining properties will be stated shortly.

For the admissible pair (t_0, z_i^0), the a priori knowledge about neighboring disturbances $u_{-ti}(\cdot)$ and the admissible feedback policy (20), the aggregation of the dynamics (14) and (17) associated with decision maker or tracker i, is described by the controlled stochastic differential equation

$$dz^i(t) = (F^i(t)z^i(t) + E^i(t)u_{-ti}(t))dt + G^i(t)dw^i(t), \quad z^i(t_0) = z_0^i \tag{21}$$

and the performance measure (18) is rewritten as follows:

$$J_i(u_{ti}, u_{-ti}) = (z^i)^T(t_f)N_f^i z^i(t_f) + \int_{t_0}^{t_f} (z^i)^T(\tau)N^i(\tau)z^i(\tau)d\tau \tag{22}$$

where the aggregate dynamical states and system coefficients are given by

$$z^i(t) \triangleq \begin{bmatrix} \hat{z}_i(t) \\ \tilde{z}_i(t) \end{bmatrix}, \quad F^i(t) \triangleq \begin{bmatrix} A_{N_i} + B_i K^i(t) & L_i(t)C_{N_i} \\ 0 & A_{N_i} - L_i(t)C_{N_i} \end{bmatrix}$$

$$G^i(t) \triangleq \begin{bmatrix} 0 & 0 & L_i(t)H_{N_i} \\ G_{N_i} & -I_{n_i \times n_i} & -L_i(t)H_{N_i} \end{bmatrix}, \; E^i(t) \triangleq \begin{bmatrix} I_{n_i \times n_i} \\ 0 \end{bmatrix}, \; W^i \triangleq \begin{bmatrix} W_i & 0 & 0 \\ 0 & M_i & 0 \\ 0 & 0 & V_i \end{bmatrix}$$

$$N^i(t) \triangleq \begin{bmatrix} Q_i + (K^i)^T(t)R_i(t)K^i(t) & Q_i \\ Q_i & Q_i \end{bmatrix}, \quad N_f^i \triangleq \begin{bmatrix} Q_{if} & Q_{if} \\ Q_{if} & Q_{if} \end{bmatrix},$$

whereas the aggregate Wiener process noise $w^i \triangleq \begin{bmatrix} w_i^T & \eta_i^T & v_i^T \end{bmatrix}^T$ has the correlation of independent increments $E\left\{[w^i(\tau_1) - w^i(\tau_2)][w^i(\tau_1) - w^i(\tau_2)]^T\right\} = W^i|\tau_1 - \tau_2|$ for all $\tau_1, \tau_2 \in [t_0, t_f]$.

2.5 Person-by-Person Downside Risk Measures

In the sequel, moving from the background of the generalized chi-squared random performance (22) and its complex behavior, one productive step involved in the discussion of the use of downside risk measures in person-by-person decision and/or control analysis is modeling and management of all the mathematical statistics (also known as semi-invariants) associated with (22). The major target in the downside risk measure debate is the measure of all the higher-order statistics associated with (22) as used in mean-risk optimization. To this end, the results that follow highlight the rather crucial role played by the endeavor of extracting higher-order statistics pertaining to random distributions of (22).

Theorem 1 (Person-by-Person Cumulant-Generating Function). *Let the states $z^i(\cdot)$ of the distributed stochastic dynamics (21) subject to the performance measure (22) be associated with risk-averse decision maker or tracker i. Further, let initial states $z^i(\tau) \equiv z_\tau^i$ and $\tau \in [t_0, t_f]$ and moment-generating functions with risk-sensitive parameter θ^i be defined by*

$$\varphi^i\left(\tau, z_\tau^i, \theta^i\right) \triangleq \varrho^i\left(\tau, \theta^i\right) \exp\left\{(z_\tau^i)^T \Upsilon^i(\tau, \theta^i) z_\tau^i + 2(z_\tau^i)^T \ell^i(\tau, \theta^i)\right\} \tag{23}$$

$$\upsilon^i\left(\tau, \theta^i\right) \triangleq \ln\{\varrho^i\left(\tau, \theta^i\right)\}, \qquad \theta^i \in \mathbb{R}^+. \tag{24}$$

Then, the cumulant-generating function is quadratic affine

$$\psi^i\left(\tau, z_\tau^i, \theta^i\right) = (z_\tau^i)^T \Upsilon^i(\tau, \theta^i) z_\tau^i + 2(z_\tau^i)^T \ell^i(\tau, \theta^i) + \upsilon^i\left(\tau, \theta^i\right) \tag{25}$$

where the backward-in-time scalar-valued $\upsilon^i\left(\tau, \theta^i\right)$ satisfies

$$\frac{d}{d\tau}\upsilon^i\left(\tau, \theta^i\right) = -\mathrm{Tr}\left\{\Upsilon^i(\tau, \theta^i) G^i(\tau) W^i (G^i)^T(\tau)\right\}, \quad \upsilon^i\left(t_f, \theta^i\right) = 0, \tag{26}$$

whereas the backward-in-time matrix $\Upsilon^i(\tau,\theta^i)$ *and vector* $\ell^i(\tau,\theta^i)$ *solutions are satisfying*

$$\frac{d}{d\tau}\Upsilon^i(\tau,\theta^i) = -(F^i)^T(\tau)\Upsilon^i(\tau,\theta^i) - \Upsilon^i(\tau,\theta^i)F^i(\tau) - \theta^i N^i(\tau) \tag{27}$$
$$-2\Upsilon^i(\tau,\theta^i)G^i(\tau)W^i(G^i)^T(\tau)\Upsilon^i(\tau,\theta^i)\,,\quad \Upsilon^i(t_f,\theta^i) = \theta^i N_f^i$$
$$\frac{d}{d\tau}\ell^i(\tau,\theta^i) = -\Upsilon^i(\tau,\theta^i)E^i(t)u_{-ti}(\tau)\,,\quad \ell^i(t_f,\theta^i) = 0\,. \tag{28}$$

Proof. For notional simplicity, it is convenient to define

$$\varpi^i\left(\tau,z_\tau^i,\theta^i\right) \triangleq \exp\left\{\theta^i J_i\left(\tau,z_\tau^i\right)\right\}\,,\quad i\in\overline{N}$$

in which the person-by-person performance measure (22) is rewritten as the cost-to-go function from an arbitrary state z_τ^i at a running time $\tau\in[t_0,t_f]$

$$J_i(\tau,z_\tau^i) = (z^i)^T(t_f)N_f^i z^i(t_f) + \int_\tau^{t_f}(z^i)^T(t)N^i(t)z^i(t)dt \tag{29}$$

subject to

$$dz^i(t) = (F^i(t)z^i(t) + E^i(t)u_{-ti}(t))dt + G^i(t)dw^i(t)\,,\quad z^i(\tau) = z_\tau^i\,. \tag{30}$$

By definition, the moment-generating function is

$$\varphi^i(\tau,z_\tau^i,\theta^i) \triangleq E\{\varpi^i(\tau,z_\tau^i,\theta^i)\}\,.$$

Thus, the total time derivative of $\varphi^i(\tau,z_\tau^i,\theta^i)$ is obtained as

$$\frac{d}{d\tau}\varphi^i\left(\tau,z_\tau^i,\theta^i\right) = -\theta^i(z_\tau^i)^T N^i(\tau)z_\tau^i\varphi^i\left(\tau,z_\tau^i,\theta^i\right)\,.$$

Using the standard Ito's formula, it follows

$$\begin{aligned} d\varphi^i\left(\tau,z_\tau^i,\theta^i\right) &= E\left\{d\varpi^i\left(\tau,z_\tau^i,\theta^i\right)\right\} = E\left\{\varpi_\tau^i\left(\tau,z_\tau^i,\theta^i\right)d\tau + \varpi_{z_\tau^i}^i\left(\tau,z_\tau^i,\theta^i\right)dz_\tau^i\right.\\ &\quad\left.+\frac{1}{2}\mathrm{Tr}\left\{\varpi_{z_\tau^i z_\tau^i}^i(\tau,z_\tau^i,\theta^i)G^i(\tau)W^i(G^i)^T(\tau)\right\}d\tau\right\}\\ &= \varphi_\tau^i(\tau,z_\tau^i,\theta^i)d\tau + \varphi_{z_\tau^i}^i(\tau,z_\tau^i,\theta^i)(F^i(\tau)z_\tau^i + E^i(\tau)u_{-ti}(\tau))d\tau\\ &\quad+\frac{1}{2}\mathrm{Tr}\left\{\varphi_{z_\tau^i z_\tau^i}^i(\tau,z_\tau^i,\theta^i)G^i(\tau)W^i(G^i)^T(\tau)\right\}d\tau \end{aligned}$$

which under the definition of the moment-generating function or the first characteristic function

$$\varphi^i\left(\tau, z_\tau^i, \theta^i\right) = \varrho^i\left(\tau, \theta^i\right) \exp\left\{(z_\tau^i)^T \Upsilon^i(\tau, \theta^i) z_\tau^i + 2(z_\tau^i)^T \ell^i(\tau, \theta^i)\right\}$$

and its partial derivatives leads to the result

$$\begin{aligned}
-\theta^i (z_\tau^i)^T N^i(\tau) z_\tau^i \varphi^i\left(\tau, z_\tau, \theta^i\right) = \Bigg\{ & \frac{\frac{d}{d\tau}\varrho^i\left(\tau, \theta^i\right)}{\varrho^i\left(\tau, \theta^i\right)} + (z_\tau^i)^T \frac{d}{d\tau} \Upsilon^i(\tau, \theta^i) z_\tau^i \\
& + 2(z_\tau^i)^T \frac{d}{d\tau} \ell^i(\tau, \theta^i) \\
& + (z_\tau^i)^T \left[(F^i)^T(\tau)\, \Upsilon^i(\tau, \theta^i) + \Upsilon^i(\tau, \theta^i) F^i(\tau)\right] z_\tau^i \\
& + 2(z_\tau^i)^T \Upsilon^i(\tau, \theta^i) E^i(\tau) u_{-ti}(\tau) \\
& + 2(z_\tau^i)^T \Upsilon^i(\tau, \theta^i) G^i(\tau)\, W^i (G^i)^T(\tau)\, \Upsilon^i(\tau, \theta^i) z_\tau^i \\
& + \mathrm{Tr}\left\{\Upsilon^i(\tau, \theta^i) G^i(\tau) W^i (G^i)^T(\tau)\right\} \Bigg\} \varphi^i\left(\tau, z_\tau^i, \theta^i\right).
\end{aligned}$$

To have constant, linear, and quadratic terms independent of arbitrary z_τ^i, it requires that the following expressions hold true:

$$\begin{aligned}
\frac{d}{d\tau}\Upsilon^i(\tau, \theta^i) &= -(F^i)^T(\tau)\Upsilon^i(\tau, \theta^i) - \Upsilon^i(\tau, \theta^i) F^i(\tau) - \theta^i N^i(\tau) \\
&\quad - 2\Upsilon^i(\tau, \theta^i) G^i(\tau) W^i (G^i)^T(\tau) \Upsilon^i(\tau, \theta^i) \\
\frac{d}{d\tau}\ell^i(\tau, \theta^i) &= -\Upsilon^i(\tau, \theta^i) E^i(\tau) u_{-ti}(\tau) \\
\frac{d}{d\tau}\varrho^i(\tau, \theta^i) &= -\varrho^i\left(\tau, \theta^i\right) \mathrm{Tr}\left\{\Upsilon^i(\tau, \theta^i) G^i(\tau) W^i (G^i)^T(\tau)\right\}
\end{aligned}$$

where the terminal-value conditions $\Upsilon^i(t_f, \theta^i) = \theta^i N_f^i$, $\ell^i(t_f, \theta^i) = 0$, and $\varrho^i(t_f, \theta^i) = 1$. Finally, the backward-in-time differential equation satisfied by $\upsilon^i(\tau, \theta^i)$ becomes

$$\frac{d}{d\tau}\upsilon^i(\tau, \theta^i) = -\mathrm{Tr}\left\{\Upsilon^i(\tau, \theta^i) G^i(\tau) W^i (G^i)^T(\tau)\right\}, \quad \upsilon^i(t_f, \theta^i) = 0,$$

which completes the proof.

Specifically, a MacLaurin series expansion of the cumulant-generating function (25) is employed to infer behaviors regarding probabilistic distributions of (22) through the knowledge representation of the mathematical statistics

$$\psi^i\left(\tau, z_\tau^i, \theta^i\right) = \sum_{r=1}^{\infty} \frac{\partial^{(r)}}{\partial \theta^{(r)}} \psi^i(\tau, z_\tau^i, \theta^i)\Big|_{\theta^i=0} \frac{(\theta^i)^r}{r!} \tag{31}$$

where all $\kappa_r^i \triangleq \frac{\partial^{(r)}}{\partial(\theta^i)^{(r)}} \psi^i(\tau, z_\tau^i, \theta^i)\Big|_{\theta^i=0}$ are performance-measure statistics available at risk-averse decision maker or tracker i

$$\begin{aligned}\kappa_r^i = (z_\tau^i)^T \frac{\partial^{(r)}}{\partial(\theta^i)^{(r)}} \Upsilon^i(\tau, \theta^i)\Big|_{\theta^i=0} z_\tau^i \\ + 2(z_\tau^i)^T \frac{\partial^{(r)}}{\partial(\theta^i)^{(r)}} \ell^i(\tau, \theta^i)\Big|_{\theta^i=0} + \frac{\partial^{(r)}}{\partial(\theta^i)^{(r)}} \upsilon^i(\tau, \theta^i)\Big|_{\theta^i=0}.\end{aligned} \tag{32}$$

For notational convenience, the change of variables

$$\begin{aligned} H_r^i(\tau) &\triangleq \frac{\partial^{(r)} \Upsilon^i(\tau, \theta^i)}{\partial(\theta^i)^{(r)}}\Big|_{\theta^i=0}, \quad \breve{D}_r^i(\tau) \triangleq \frac{\partial^{(r)} \ell^i(\tau, \theta^i)}{\partial(\theta^i)^{(r)}}\Big|_{\theta^i=0} \\ D_r^i(\tau) &\triangleq \frac{\partial^{(r)} \upsilon^i(\tau, \theta^i)}{\partial(\theta^i)^{(r)}}\Big|_{\theta^i=0}, \quad \tau \in [t_0, t_f]; \quad r \in \mathbb{N} \end{aligned} \tag{33}$$

is introduced so that the next result will provide an effective and accurate capability for forecasting all the higher-order characteristics associated with performance uncertainty (22).

Theorem 2 (Person-by-Person Performance-Measure Statistics). *Let (A_{N_i}, B_i) and (C_{N_i}, A_{N_i}) associated with the coupling constraint (21) and the goal function (22) be stabilizable and detectable. For $k^i \in \mathbb{N}$, the k^ith performance-measure statistic of (22) concerned by risk-averse decision maker or tracker i and $i \in \overline{N}$ is given by*

$$\kappa_k^i = (z_0^i)^T H_{k^i}^i(t_0) z_0^i + 2(z_0^i)^T \breve{D}_{k^i}^i(t_0) + D_{k^i}^i(t_0) \tag{34}$$

where the supporting variables $\{H_r^i(\tau)\}_{r=1}^{k^i}$, $\{\breve{D}_r^i(\tau)\}_{r=1}^{k^i}$, and $\{D_r^i(\tau)\}_{r=1}^{k^i}$ evaluated at $\tau = t_0$ satisfy the differential equations (with the dependence of $H_r^i(\tau)$, $\breve{D}_r^i(\tau)$, and $D_r(\tau)$ upon the admissible feedback policy gain $K^i(\tau)$ and other observable policies $u_{-ti}(\tau)$ suppressed)

$$\frac{d}{d\tau} H_1^i(\tau) = -(F^i)^T(\tau) H_1^i(\tau) - H_1^i(\tau) F^i(\tau) - N^i(\tau) \tag{35}$$

$$\frac{d}{d\tau}H_r^i(\tau) = -(F^i)^T(\tau)H_r^i(\tau) - H_r^i(\tau)F^i(\tau) \tag{36}$$
$$-\sum_{s=1}^{r-1}\frac{2r!}{s!(r-s)!}H_s^i(\tau)G^i(\tau)W^i(G^i)^T(\tau)H_{r-s}^i(\tau)\,,\quad 2\le r\le k^i$$

and

$$\frac{d}{d\tau}\breve{D}_r^i(\tau) = -H_r^i(\tau)E^i(\tau)u^{-i}(\tau)\,,\qquad 1\le r\le k^i \tag{37}$$

and, finally,

$$\frac{d}{d\tau}D_r^i(\tau) = -\mathrm{Tr}\left\{H_r^i(\tau)G^i(\tau)W^i(G^i)^T(\tau)\right\}\,,\quad 1\le r\le k^i \tag{38}$$

provided that the terminal-value conditions $H_1^i(t_f) = N_f^i$, $H_r^i(t_f) = 0$ *for* $2 \le r \le k^i$, $\breve{D}_r^i(t_f) = 0$ *for* $1 \le r \le k^i$, *and* $D_r^i(t_f) = 0$ *for* $1 \le r \le k^i$.

Proof. The expression of performance-measure statistics described in (34) is readily justified by using the result (32) and the definition (33). What remains is to show that the solutions $H_r^i(\tau)$, $\breve{D}_r^i(\tau)$, and $D_r^i(\tau)$ for $1 \le r \le k^i$ indeed satisfy the dynamical equations (35)–(38). Notice that these backward-in-time equations (35)–(38) satisfied by the matrix-valued $H_r^i(\tau)$, vector-valued $\breve{D}_r^i(\tau)$, and scalar-valued $D_r^i(\tau)$ solutions are then obtained by successively taking derivatives with respect to θ of the supporting equations (26)–(28) and subject to the assumptions of (A_{N_i}, B_i) and (C_{N_i}, A_{N_i}) being uniformly stabilizable and detectable on $[t_0, t_f]$.

3 Problem Statements

In the context of risk-averse decision making, cautious decision makers or trackers who realize performance risk akin to a costly preference for certainty will have to leverage higher-order statistics of the probability distribution (22) for downside risk measures and optimizing risk-averse decisions. For such a problem it is important to have a compact statement of the risk-averse decision and control optimization so as to aid mathematical manipulations. Precisely, one may think of the k^i-tuple state variables

$$\mathcal{H}^i(\cdot) \triangleq (\mathcal{H}_1^i(\cdot), \ldots, \mathcal{H}_{k^i}^i(\cdot))\,,$$
$$\breve{\mathcal{D}}^i(\cdot) \triangleq (\breve{\mathcal{D}}_1^i(\cdot), \ldots, \breve{\mathcal{D}}_{k^i}^i(\cdot))\,,$$
$$\mathcal{D}^i(\cdot) \triangleq (\mathcal{D}_1^i(\cdot), \ldots, \mathcal{D}_{k^i}^i(\cdot))\,,$$

whose continuously differentiable state variables $\mathcal{H}_r^i \in \mathcal{C}^1([t_0, t_f]; \mathbb{R}^{2n_i \times 2n_i})$, $\breve{\mathcal{D}}_r^i \in \mathcal{C}^1([t_0, t_f]; \mathbb{R}^{2n_i \times 1})$, and $\mathcal{D}_r^i \in \mathcal{C}^1([t_0, t_f]; \mathbb{R})$ having the representations

$$\mathcal{H}_r^i(\cdot) \triangleq H_r^i(\cdot), \qquad \breve{\mathcal{D}}_r^i(\cdot) \triangleq \breve{D}_r^i(\cdot), \qquad \mathcal{D}_r^i(\cdot) \triangleq D_r^i(\cdot)$$

with the right members satisfying the dynamics (35)–(38) are defined on $[t_0, t_f]$.

In the remainder of the development, the convenient mappings are introduced as

$$\begin{aligned} \mathcal{F}_r^i &: [t_0, t_f] \times (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \times \mathbb{R}^{m_i \times n_i} \mapsto \mathbb{R}^{2n_i \times 2n_i} \\ \breve{\mathcal{G}}_r^i &: [t_0, t_f] \times (\mathbb{R}^{2n_i \times 1})^{k^i} \mapsto \mathbb{R}^{2n_i \times 1} \\ \mathcal{G}_r^i &: [t_0, t_f] \times (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \mapsto \mathbb{R} \end{aligned}$$

where the rules of action are given by

$$\begin{aligned} \mathcal{F}_1^i(\tau, \mathcal{H}^i, K^i) &\triangleq -(F^i)^T(\tau)\mathcal{H}_1^i(\tau) - \mathcal{H}_1^i(\tau)F^i(\tau) - N^i(\tau) \\ \mathcal{F}_r^i(\tau, \mathcal{H}^i, K^i) &\triangleq -(F^i)^T(\tau)\mathcal{H}_r^i(\tau) - \mathcal{H}_r^i(\tau)F^i(\tau) \\ &\quad - \sum_{s=1}^{r-1} \frac{2r!}{s!(r-s)!} \mathcal{H}_s^i(\tau)G^i(\tau)W^i(G^i)^T(\tau)\mathcal{H}_{r-s}^i(\tau), \quad 2 \le r \le k^i \\ \breve{\mathcal{G}}_r^i(\tau, \mathcal{H}^i) &\triangleq -\mathcal{H}_r^i(\tau)E^i(\tau)u^{-i}(\tau), \quad 1 \le r \le k^i \\ \mathcal{G}_r^i(\tau, \mathcal{H}^i) &\triangleq -\mathrm{Tr}\left\{\mathcal{H}_r^1(\tau)G^i(\tau)W^i(G^i)^T(\tau)\right\}, \quad 1 \le r \le k^i. \end{aligned}$$

The product mappings that follow are necessary for a compact formulation; for example,

$$\begin{aligned} \mathcal{F}_1^i \times \cdots \times \mathcal{F}_{k^i}^i &: [t_0, t_f] \times (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \times \mathbb{R}^{m_i \times n_i} \mapsto (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \\ \breve{\mathcal{G}}_1^i \times \cdots \times \breve{\mathcal{G}}_{k^i}^i &: [t_0, t_f] \times (\mathbb{R}^{2n_i \times 1})^{k^i} \mapsto (\mathbb{R}^{2n_i \times 1})^{k^i} \\ \mathcal{G}_1^i \times \cdots \times \mathcal{G}_{k^i}^i &: [t_0, t_f] \times (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \mapsto \mathbb{R}^{k^i} \end{aligned}$$

where the corresponding notations

$$\mathcal{F}^i \triangleq \mathcal{F}_1^i \times \cdots \times \mathcal{F}_{k^i}^i, \quad \breve{\mathcal{G}}^i \triangleq \breve{\mathcal{G}}_1^i \times \cdots \times \breve{\mathcal{G}}_{k^i}^i, \quad \mathcal{G}^i \triangleq \mathcal{G}_1^i \times \cdots \times \mathcal{G}_{k^i}^i$$

are used. Thus, the dynamical equations (35)–(38) can be rewritten as

$$\frac{d}{d\tau}\mathcal{H}^i(\tau) = \mathcal{F}^i(\tau, \mathcal{H}^i(\tau), K^i(\tau)), \qquad \mathcal{H}^i(t_f) \equiv \mathcal{H}_f^i \tag{39}$$

$$\frac{d}{d\tau}\breve{\mathcal{D}}^i(\tau) = \breve{\mathcal{G}}^i\left(\tau, \mathcal{H}^i(\tau)\right), \qquad \breve{\mathcal{D}}^i(t_f) \equiv \breve{\mathcal{D}}^i_f \tag{40}$$

$$\frac{d}{d\tau}\mathcal{D}^i(\tau) = \mathcal{G}^i\left(\tau, \mathcal{H}^i(\tau)\right), \qquad \mathcal{D}^i(t_f) \equiv \mathcal{D}^i_f \tag{41}$$

where the k^i-tuple terminal conditions $\mathcal{H}^i_f \triangleq (N^i_f, 0, \ldots, 0)$, $\breve{\mathcal{D}}^i_f \triangleq (0, \ldots, 0)$, and $\mathcal{D}^i_f \triangleq (0, \ldots, 0)$.

Notice that the product system (39)–(41) uniquely determines the state matrices $\mathcal{H}^i$, $\breve{\mathcal{D}}^i$, and $\mathcal{D}^i$ once the admissible feedback policy gain K^i and observable policies u_{-ti} by decision maker and/or tracker i are specified. Henceforth, these state variables are considered as

$$\mathcal{H}^i \equiv \mathcal{H}^i(\cdot, K^i, u_{-ti}), \quad \breve{\mathcal{D}}^i \equiv \breve{\mathcal{D}}^i(\cdot, K^i, u_{-ti}), \quad \mathcal{D}^i \equiv \mathcal{D}^i(\cdot, K^i, u_{-ti}).$$

Given terminal data $(t_f, \mathcal{H}^i_f, \breve{\mathcal{D}}^i_f, \mathcal{D}^i_f)$, the class of admissible person-by-person feedback decision/control gains employed by risk-averse decision maker and/or tracker i is next defined.

Definition 1 (Person-by-Person Feedback Decision/Control Gains). Let compact subset $\overline{K}^i \subset \mathbb{R}^{m_i \times n}$ be the set of allowable feedback form values. For the given $k^i \in \mathbb{N}$ and sequence $\mu^i = \{\mu^i_r \geq 0\}_{r=1}^{k^i}$ with $\mu^i_1 > 0$, the set of feedback gains $\mathcal{K}^i_{t_f, \mathcal{H}^i_f, \breve{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}$ is assumed to be the class of $\mathcal{C}([t_0, t_f]; \mathbb{R}^{m_i \times n_i})$ with values $K^i(\cdot) \in \overline{K}^i$, for which the solutions to the dynamic equations (39)–(41) with the terminal-value conditions $\mathcal{H}^i(t_f) = \mathcal{H}^i_f$, $\breve{\mathcal{D}}^i(t_f) = \breve{\mathcal{D}}^i_f$, and $\mathcal{D}^i(t_f) = \mathcal{D}^i_f$ exist on the interval of optimization $[t_0, t_f]$.

An obvious fact about the private set of design freedom $\mu^i = \{\mu^i_r \geq 0\}_{r=1}^{k^i}$ with $\mu^i_1 > 0$ is that risk sensitivity entails the lack of certainty equivalence, in the sense that any performance index formed only by the first statistic of (22) does not lead to optimal decisions. In addition, it is important to recognize that this finite set of custom weights is quite different from those of infinite sets of series expansion coefficients as in [1, 6, 17], just to name a few.

On $\mathcal{K}^i_{t_f, \mathcal{H}^i_f, \breve{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}$ the performance index with mean-risk considerations is subsequently defined as follows.

Definition 2 (Mean-Risk Aware Performance Index). Let cautious decision maker and/or tracker i select $k^i \in \mathbb{N}$ and the set of custom weights $\mu^i = \{\mu^i_r \geq 0\}_{r=1}^{k^i}$ with $\mu^i_1 > 0$. Then, for the given z^i_0, the mean-risk aware performance index

$$\phi^i_0 : \{t_0\} \times (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \times (\mathbb{R}^{2n_i})^{k^i} \times \mathbb{R}^{k^i} \mapsto \mathbb{R}^+$$

pertaining to person-by-person risk-averse decision making over $[t_0, t_f]$ is

$$\phi_0^i(t_0, \mathcal{H}^i, \breve{\mathcal{D}}^i, \mathcal{D}^i) \triangleq \underbrace{\mu_1^i \kappa_1^i}_{\text{Mean}} + \underbrace{\mu_2^i \kappa_2^i + \cdots + \mu_{k^i}^i \kappa_{k^i}^i}_{\text{Risk}}$$

$$= \sum_{r=1}^{k^i} \mu_r^i [(z_0^i)^T \mathcal{H}_r^i(t_0) z_0^i + 2(z_0^i)^T \breve{\mathcal{D}}_r^i(t_0) + \mathcal{D}_r^i(t_0)] \tag{42}$$

where additional design freedom μ_r^i's utilized by cautious and defensive decision maker and/or tracker i are tailored to meet different levels of performance-based reliability requirements, e.g., mean, variance, anti-symmetry, heavy tails of the reward/cost density (22), etc., pertaining to closed-loop performance uncertainties and whereas the supporting solutions $\{\mathcal{H}_r^i(\tau)\}_{r=1}^{k^i}$, $\{\breve{\mathcal{D}}_r^i(\tau)\}_{r=1}^{k^i}$ and $\{\mathcal{D}_r^i(\tau)\}_{r=1}^{k^i}$ evaluated at $\tau = t_0$ satisfy the dynamical equations (39)–(41).

The technical challenge faced by a cautious decision maker and/or tracker i and $i \in \overline{N}$ is that the correspondent mean-risk aware performance index (42) depends on the observable decision and/or control policies from neighboring decision makers or trackers u_{-ti}. The basic question the decision maker and/or tracker i faces is whether or not a sort of noncooperative equilibrium or Nash solution is possible at all. Defensive decision maker and/or tracker i's rationale for choosing K_*^i is to force the immediate neighbors to hold u_{-ti}^*, so as to secure the Nash payoff. Thus, it is precisely in this sense that a Nash solution is risk-averse by nature.

Definition 3 (Feedback Nash Equilibrium). Let K_*^i constitute a feedback Nash strategy such that

$$\phi_0^i(K_*^i, u_{-ti}^*) \leq \phi_0^i(K^i, u_{-ti}^*), \qquad i \in \overline{N}_i \tag{43}$$

for all admissible $K^i \in \mathcal{K}^i_{t_f, \mathcal{H}_f^i, \breve{\mathcal{D}}_f^i, \mathcal{D}_f^i; \mu^i}$, upon which the solutions to the dynamical systems (39)–(41) exist on $[t_0, t_f]$.

Then, $\left(K_*^1, \ldots, K_*^{N_i}\right)$ when restricted to $[t_0, \tau]$ is still a N_i-tuple feedback Nash equilibrium solution for the multiperson Nash decision problem with the appropriate terminal-value condition $(\tau, \mathcal{H}_*^i(\tau), \breve{\mathcal{D}}_*^i(\tau), \mathcal{D}_*^i(\tau))$ for all $\tau \in [t_0, t_f]$.

Of note, an N_i-tuple of decision and/or control policies $\left(K_*^1, \ldots, K_*^{N_i}\right)$ is said to constitute an interactive feedback Nash equilibrium solution for an N_i-agent differential graphical game if, for all $i \in \overline{N}_i$, the following Nash condition holds

$$\phi_0^i(K_*^i, u_{-ti}^*) \leq \phi_0^i(K^i, u_{-ti}^*).$$

In addition, there exist decision and/or control policies $\underline{K}^j$ and $\overline{K}^j$ such that

$$\phi_0^i(\underline{K}^j, u_{-tj}^*) \neq \phi_0^i(\overline{K}^j, u_{-tj}^*), \quad \forall\, i, j \in \overline{N}_i\,.$$

The interpretation is that the variation of person-by-person performance index pertaining to decision maker and/or tracker i is resulted while the rest of the immediate decision makers and/or trackers in the local neighborhood of decision maker and/or tracker j supported by the corresponding connectivity graph assume their optimal strategies.

Now, the objective of cautious decision maker and/or tracker i is to minimize (42) over $K^i = K^i(\cdot)$ in $\mathcal{K}^i_{t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}$ and subject to the neighboring feedback Nash policies u_{-ti}^*.

Definition 4 (Person-by-Person Optimization). Given the profile of risk-averse attitudes $\mu^i = \{\mu^i_r \geq 0\}_{r=1}^{k^i}$ with $\mu^i_1 > 0$, the decision optimization problem defined by

$$\min_{K^i(\cdot) \in \mathcal{K}_{t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}} \phi_0^i(K^i, u_{-ti}^*) \tag{44}$$

is subject to the dynamical equations (39)–(41) on $[t_0, t_f]$.

In conformity with the dynamic programming approach, the terminal time and states $(t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f)$ are parameterized as $(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i)$ whereby $\mathcal{Y}^i \triangleq \mathcal{H}^i(\varepsilon)$, $\check{\mathcal{Z}}^i \triangleq \check{\mathcal{D}}^i(\varepsilon)$, and $\mathcal{Z}^i \triangleq \mathcal{D}^i(\varepsilon)$. Thus, the value function of (44) now depends on the parameterization of the terminal-value conditions.

Definition 5 (Value Function). Let $(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) \in [t_0, t_f] \times (\mathbb{R}^{2n_i \times 2n_i})^{k^i} \times (\mathbb{R}^{2n_i \times 1})^{k^i} \times \mathbb{R}^{k^i}$. Then, the value function $\mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i)$ is defined by

$$\mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) \triangleq \inf_{K^i(\cdot) \in \mathcal{K}_{\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i; \mu^i}} \phi_0^i(K^i, u_{-ti}^*)\,.$$

For convention, $\mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) \triangleq \infty$ when $\mathcal{K}_{\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i; \mu^i}$ is empty. Some candidates for the value function are also constructed with the help of the concept of reachable set.

Definition 6 (Reachable Sets). Let a reachable set of decision maker i be defined by $\mathcal{Q}^i \triangleq \{(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) \in [t_0, t_f] \times (\mathbb{R}^{2n \times 2n})^{k^i} \times (\mathbb{R}^{2n \times 1})^{k^i} \times \mathbb{R}^{k^i} : \mathcal{K}^i_{\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i; \mu^i} \neq \emptyset\}$.

Formally, it can be shown that the value function associated with decision maker i is satisfying partial differential equation (e.g., Hamilton-Jacobi-Bellman (HJB) equation) at interior points of $\mathcal{Q}^i$, at which it is differentiable.

Theorem 3 (HJB Equation-Mayer Problem). *Let* $(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i)$ *be any interior point of* $\mathcal{Q}^i$*, at which the value function* $\mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i)$ *is differentiable. If there exists a feedback Nash strategy* $K^i_* \in \mathcal{K}^i_{t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}$*, then the partial differential equation is satisfied:*

$$
\begin{aligned}
0 = \min_{K^i \in \overline{K}^i} \Big\{ & \frac{\partial}{\partial \varepsilon} \mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) + \frac{\partial}{\partial\, vec(\mathcal{Y}^i)} \mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) vec(\mathcal{F}^i(\varepsilon, \mathcal{Y}^i, K^i)) \\
& + \frac{\partial}{\partial\, vec(\check{\mathcal{Z}}^i)} \mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) vec(\check{\mathcal{G}}^i(\varepsilon, \mathcal{Y}^i) \\
& + \frac{\partial}{\partial\, vec(\mathcal{Z}^i)} \mathcal{V}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i) vec(\mathcal{G}^i(\varepsilon, \mathcal{Y}^i) \Big\}
\end{aligned} \tag{45}
$$

and $\mathcal{V}^i(t_0, \mathcal{Y}^i(t_0), \check{\mathcal{Z}}^i(t_0), \mathcal{Z}^i(t_0)) = \phi^i_0(\mathcal{H}^i(t_0), \check{\mathcal{D}}^i(t_0), \mathcal{D}^i(t_0))$.

Proof. Similar to that of [12] and hence is omitted.

Finally, the sufficient condition for verification of a feedback Nash strategy by cautious decision maker and/or tracker i and $i \in \overline{N}$ is given below.

Theorem 4 (Verification Theorem). *Let* $\mathcal{W}^i(\varepsilon, \mathcal{Y}^i, \check{\mathcal{Z}}^i, \mathcal{Z}^i)$ *be continuously differentiable solution of the HJB equation (45) with the boundary condition*

$$
\mathcal{W}^i(t_0, \mathcal{H}^i(t_0), \check{\mathcal{D}}^i(t_0), \mathcal{D}^i(t_0)) = \phi^i_0(t_0, \mathcal{H}^i(t_0), \check{\mathcal{D}}^i(t_0), \mathcal{D}^i(t_0)) \,.
$$

Let $(t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f) \in \mathcal{Q}^i$*,* $K^i \in \mathcal{K}^i_{t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}$*,* $(\mathcal{H}^i(\cdot), \check{\mathcal{D}}^i(\cdot), \mathcal{D}^i(\cdot))$ *be the trajectory solutions of (39)–(41). Then,* $\mathcal{W}^i(\tau, \mathcal{H}^i(\tau), \check{\mathcal{D}}^i(\tau), \mathcal{D}^i(\tau))$ *is a time-backward increasing function of* $\tau \in [t_0, t_f]$.

If K^i_* *is in* $\mathcal{K}^i_{t_f, \mathcal{H}^i_f, \check{\mathcal{D}}^i_f, \mathcal{D}^i_f; \mu^i}$ *with the associative solutions* $(\mathcal{H}^i_*(\cdot), \check{\mathcal{D}}^i_*(\cdot), \mathcal{D}^i_*(\cdot))$ *of the equations (39)–(41) such that*

$$
\begin{aligned}
0 = & \frac{\partial}{\partial \varepsilon} \mathcal{W}^i(\tau, \mathcal{H}^i_*(\tau), \check{\mathcal{D}}^i_*(\tau), \mathcal{D}^i_*(\tau)) \\
& + \frac{\partial}{\partial\, vec(\mathcal{Y}^i)} \mathcal{W}^i(\tau, \mathcal{H}^i_*(\tau), \check{\mathcal{D}}^i_*(\tau), \mathcal{D}^i_*(\tau)) vec(\mathcal{F}^i(\tau, \mathcal{H}^i_*(\tau), K^i_*(\tau))) \\
& + \frac{\partial}{\partial\, vec(\check{\mathcal{Z}}^i)} \mathcal{W}^i(\tau, \mathcal{H}^i_*(\tau), \check{\mathcal{D}}^i_*(\tau), \mathcal{D}^i_*(\tau)) vec(\check{\mathcal{G}}^i(\tau, \mathcal{H}^i_*(\tau))) \\
& + \frac{\partial}{\partial\, vec(\mathcal{Z}^i)} \mathcal{W}_{is}(\tau, \mathcal{H}^i_*(\tau), \check{\mathcal{D}}^i_*(\tau), \mathcal{D}^i_*(\tau)) vec(\mathcal{G}^i(\tau, \mathcal{H}^i_*(\tau))),
\end{aligned} \tag{46}
$$

then K_*^i *is a feedback Nash strategy in* $\mathcal{K}^i_{t_f,\mathcal{H}^i_f,\check{\mathcal{D}}^i_f,\mathcal{D}^i_f;\mu^i}$,

$$\mathcal{W}^i(\varepsilon,\mathcal{Y}^i,\check{\mathcal{Z}}^i,\mathcal{Z}^i)=\mathcal{V}^i(\varepsilon,\mathcal{Y}^i,\check{\mathcal{Z}}^i,\mathcal{Z}^i)\,. \tag{47}$$

Proof. The proof follows the same manner as [12].

4 Person-By-Person Risk-Averse Strategies

To this end, the initial state z_0^i is recognized to contribute linearly and quadratically to the mean-risk performance index (42). Henceforth, it is beneficial to infer that a candidate for the value function is expected to take the form

$$\begin{aligned}\mathcal{W}^i(\varepsilon,\mathcal{Y}^i,\check{\mathcal{Z}}^i,\mathcal{Z}^i)=(z_0^i)^T\sum_{r=1}^{k^i}\mu_r^i(\mathcal{Y}_r^i+\mathcal{E}_r^i(\varepsilon))\,z_0^i\\ +2(z_0^i)^T\sum_{r=1}^{k^i}\mu_r^i(\check{\mathcal{Z}}_r^i+\check{\mathcal{T}}_r^i(\varepsilon))+\sum_{r=1}^{k^i}\mu_r^i(\mathcal{Z}_r^i+\mathcal{T}_r^i(\varepsilon))\end{aligned} \tag{48}$$

where the functions $\mathcal{E}_r^i\in\mathcal{C}^1([t_0,t_f];\mathbb{R}^{2n_i\times 2n_i})$, $\check{\mathcal{T}}_r^i\in\mathcal{C}^1([t_0,t_f];\mathbb{R}^{2n_i\times 1})$, and $\mathcal{T}_r^i\in\mathcal{C}^1([t_0,t_f];\mathbb{R})$ are time parameterized and yet to be determined.

As reported in [13], the time derivative of $\mathcal{W}^i(\varepsilon,\mathcal{Y}^i,\check{\mathcal{Z}}^i,\mathcal{Z}^i)$ can be shown as follows:

$$\begin{aligned}\frac{d}{d\varepsilon}\mathcal{W}^i(\varepsilon,\mathcal{Y}^i,\check{\mathcal{Z}}^i,\mathcal{Z}^i)=(z_0^i)^T\sum_{r=1}^{k^i}\mu_r^i[\mathcal{F}_r^i(\varepsilon,\mathcal{Y}^i,K^i)+\frac{d}{d\varepsilon}\mathcal{E}_r^i(\varepsilon)]z_0^i\\ +2(z_0^i)^T\sum_{r=1}^{k^i}\mu_r^i[\check{\mathcal{G}}_r^i(\varepsilon,\mathcal{Y}^i)+\frac{d}{d\varepsilon}\check{\mathcal{T}}_r^i(\varepsilon)]\\ +\sum_{r=1}^{k^i}\mu_r^i[\mathcal{G}_r^i(\varepsilon,\mathcal{Y}^i)+\frac{d}{d\varepsilon}\mathcal{T}_r^i(\varepsilon)]\,.\end{aligned} \tag{49}$$

The substitution of this candidate (48) for the value function into the HJB equation (45) and making use of (49) yield

$$\begin{aligned}0=\min_{K^i\in\overline{K}^i}\Big\{(z_0^i)^T\sum_{r=1}^{k^i}\mu_r^i[\mathcal{F}_r^i(\varepsilon,\mathcal{Y}^i,K^i)+\frac{d}{d\varepsilon}\mathcal{E}_r^i(\varepsilon)]z_0^i\\ +2(z_0^i)^T\sum_{r=1}^{k^i}\mu_r^i[\check{\mathcal{G}}_r^i(\varepsilon,\mathcal{Y}^i)+\frac{d}{d\varepsilon}\check{\mathcal{T}}_r^i(\varepsilon)]+\sum_{r=1}^{k^i}\mu_r^i[\mathcal{G}_r^i(\varepsilon,\mathcal{Y}^i)+\frac{d}{d\varepsilon}\mathcal{T}_r^i(\varepsilon)]\Big\}.\end{aligned} \tag{50}$$

Taking the gradient with respect to K^i of the expression within the bracket of (50) yields the necessary conditions for an extremum of (42) on $[t_0, \varepsilon]$ where $I_0^T \triangleq \left[I_{n_i \times n_i} \; 0 \right]$

$$K^i = -R_i^{-1}(B_i)^T I_0^T \sum_{r=1}^{k^i} \hat{\mu}_r^i \mathcal{Y}_r^i \, I_0((I_0^T I_0)^{-1})^T \tag{51}$$

in which $\hat{\mu}_r^i \triangleq \mu_r^i / \mu_1^i$ for $\mu_1^i > 0$. With the feedback Nash strategy (51) replaced in the expression of the bracket (50) and having $\{\mathcal{Y}_r^i\}_{r=1}^{k^i}$ evaluated on the optimal solution trajectories (39)–(41), the time parametric functions $\mathcal{E}_r^i(\varepsilon)$, $\breve{\mathcal{T}}_r^i(\varepsilon)$, and $\mathcal{T}_r^i(\varepsilon)$ are thus chosen so that the sufficient condition (46) in the verification theorem is satisfied in spite of the arbitrary values z_0^i; for example,

$$\dot{\mathcal{E}}_1^i(\varepsilon) = (F_*^i)^T(\varepsilon)\mathcal{H}_{1*}^i(\varepsilon) + \mathcal{H}_{1*}^i(\varepsilon)F_*^i(\varepsilon) + N_*^i(\varepsilon)\,, \quad \mathcal{E}_1^i(t_0) = 0$$

$$\begin{aligned}\dot{\mathcal{E}}_r^i(\varepsilon) = {} & (F_*^i)^T(\varepsilon)\mathcal{H}_{r*}^i(\varepsilon) + \mathcal{H}_{r*}^i(\varepsilon)F_*^i(\varepsilon) \\ & + \sum_{s=1}^{r-1} \frac{2r!}{s!(r-s)!} \mathcal{H}_{s*}^i(\varepsilon)G^i(\varepsilon)W^i(\varepsilon)(G^i)^T(\varepsilon)\mathcal{H}_{r-s*}^i(\varepsilon), \quad \mathcal{E}_r^i(t_0) = 0\,, \quad r \geq 2\end{aligned}$$

$$\dot{\breve{\mathcal{T}}}_r^i(\varepsilon) = \mathcal{H}_{r*}^i(\varepsilon)E^i(\varepsilon)u_{-ti}^*(\varepsilon), \quad \breve{\mathcal{T}}_r^i(t_0) = 0, \quad 1 \leq r \leq k^i$$

$$\dot{\mathcal{T}}_r^i(\varepsilon) = \mathrm{Tr}\left\{\mathcal{H}_{1*}^i(\varepsilon)G^i(\varepsilon)W^i G^{iT}(\varepsilon)\right\}\,, \quad \mathcal{T}_r^i(t_0) = 0\,, \quad 1 \leq r \leq k^i\,.$$

Before closing the section, it is important to note that the sufficient condition (46) of the verification theorem is satisfied. Hence, the extremizing feedback strategy (51) associated with cautious decision maker i becomes optimal.

Theorem 5 (Person-by-Person Risk-Averse Decision/Control Strategies). *Consider the multi-agent tracking and synchronization supported by connectivity graphs wherein cautious and defensive decision maker and/or tracker i and $i \in \overline{N}$ have complete knowledge of the coupling constraints (6), (7), (18), and mean-risk aware performance index (42). When all decision makers i have the similar risk attitudes, an imitative or Nash equilibrium exists and is enabled by the risk-averse class of feedback strategies*

$$u_{ti}^*(t) = K_*^i(t)\hat{z}_i^*(t), \quad t \triangleq t_0 + t_f - \tau \tag{52}$$

$$K_*^i(\tau) = -R_i^{-1}(B_i)^T I_0^T \sum_{r=1}^{k^i} \hat{\mu}_r^i \mathcal{H}_{r*}^i(\tau) I_0((I_0^T I_0)^{-1})^T$$

where all the parametric design freedom through $\hat{\mu}_r^i$ represent the risk-averse preferences toward performance distributions; the optimal trajectory solutions $\mathcal{H}_{r*}^i(\cdot)$ are satisfying (39).

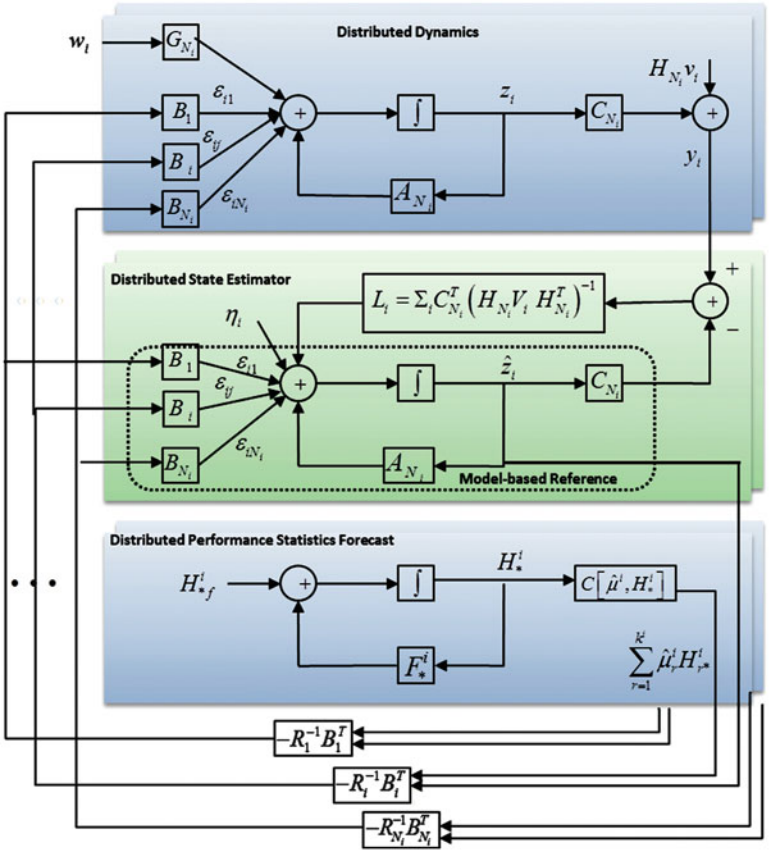

Fig. 1 Distributed decision architecture for performance risk aversion

Notice that, to have the distributed feedback Nash policy (52) of decision maker i be defined and continuous for all $\tau \in [t_0, t_f]$, the solutions $\mathcal{H}^i_{r*}(\tau)$ to the equations (39) when evaluated at $\tau = t_0$ must also exist. Therefore, it is necessary that $\mathcal{H}^i_{r*}(\tau)$ are finite for all $\tau \in [t_0, t_f)$. Moreover, the solutions of (39) exist and are continuously differentiable in a neighborhood of t_f. Under the assumption of (A_{N_i}, B_i) and (C_{N_i}, A_{N_i}) being stabilizable and detectable, the result from [3] concludes that these solutions can further be extended to the left of t_f as long as $\mathcal{H}^i_{r*}(\tau)$ remain finite. Hence, the existence of unique and continuously differentiable solutions to the equations (39) is certain if $\mathcal{H}^i_{r*}(\tau)$ are bounded for all $\tau \in [t_0, t_f)$. As the result, the candidate value functions $\mathcal{W}^i(\tau, \mathcal{H}^i, \breve{\mathcal{D}}^i, \mathcal{D}^i)$ are continuously differentiable as well.

As illustrated by Fig. 1, the person-by-person decisions and controls, generated by risk-averse policies $u^*_{ti}(t)$ for $r = 1, \ldots, N$, not only depend on the basis of information $\hat{z}^*_i(t)$ about the conditional probability distribution for coupling

interactions from the local neighborhood supported by connectivity graphs, but also rely on the robust prediction of higher-order characteristics for person-by-person performance uncertainty, e.g., mean, variance, skewness, etc. The need for the control decision laws to take into account accurate estimations of performance uncertainty is one form of interaction between two interdependent functions of a decision and/or control strategy: i) anticipation of performance uncertainty and ii) proactive decisions for mitigating downside performance risk measures. This form of interaction between these two decision and/or control strategy functions gives rise to what are now termed as *performance probing* and *performance cautioning* and thus are of particular importance in the newly developed theory of statistical optimal control.

5 Conclusions

In this chapter, the research emphasis and contributions have been the generalization of the results known for linear-quadratic classes of noncooperative stochastic games and distributed controls. Specifically, under risk attitudes toward performance uncertainties, the risk-averse feedback decision laws are not only the functions of higher-order statistics of the chi-squared rewards or costs but also dependent of a priori knowledge of common process noises as well as subjective observation noises. Thus, both certainty equivalence and separation principle do not hold. Also important is that the existence of the Nash equilibrium as proposed herein is conditional upon the custom sets of selective weights, which in turn relate to risk parameters residing at cautious decision makers or controllers. An extension of the results obtained in this exposition may be worthy of future investigation, when there are presence of mistrust and excessive risk aversion; such results could constitute fundamentals and principles in adversarial systems sciences and flexibly survivable decision making.

References

1. Basar, T.: Nash Equilibria of Risk-Sensitive Nonlinear Stochastic Differential Games. J. Optimizat. Theor. Appl. **100**(3), 479–498 (1999)
2. Basar, T., Olsder, G.J.: Dynamic Noncooperative Game Theory. 2nd Edn., Society for Industrial and Applied Mathematics (1999)
3. Dieudonne, J.: Foundations of Modern Analysis. Academic Press, New York and London (1960)
4. Engwerda, J.C.: LQ Dynamic Optimization and Differential Games. Wiley, NewYork (2005)
5. Gu, D.: A differential game approach to formation control. IEEE Trans. Contr. Syst. Technol. **16**(1), 85–93, (2008)
6. Jacobson, D.H. : Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic games. IEEE Trans. Autom. Contr. **18**, 124–131 (1973)

7. Kang, W., Ross, M.I., Pham, K.D., Gong, Q.: Observability of networked multi-satellite systems J. Guidance, Contr. Dynam. **32**(3), 869–877 (2009)
8. Langbort, C., Chandra, R., D'Andrea, R.: distributed control design for systems interconnected over an arbitrary graph. IEEE Trans. Automat. Contr. **49**(9), 1502–1519 (2004)
9. Luce, R.D.: Several possible measures of risk. Theor. Decis. **12**, 217–228 (1980)
10. Martinez, S., Bullo, F.: Optimal sensor placement and motion coordination for target tracking. Automatica **42**(4), 661–668 (2006)
11. Pham, K.D.: Cooperative Outcomes for Stochastic Nash Games: Decision Strategies towards Multi-Attribute Performance Robustness, The 17th International Federation of Automatic Control World Congress, pp. 11750–11756, Seoul, Korea (2008)
12. Pham, K.D.: Performance-reliability-aided decision-making in multiperson quadratic decision games against jamming and estimation confrontations. J. Optim. Theo. Appl. Edited by F. Giannessi, **149**(1), 599–629 (2011)
13. Pham, K.D.: Statistical Control Paradigms for Structural Vibration Suppression, PhD Dissertation (2004). Department of Electrical Engineering, University of Notre Dame, Indiana U.S.A. Available via the URL http://etd.nd.edu/ETD-db/theses/available/etd-04152004-121926/unrestricted/PhamKD052004.pdf. Cited 12 May 2014
14. Pollatsek, A., Tversky, A.: Theory of risk. J. Math. Psychol. **7**(3), 540–53 (1970)
15. Shen, D., Chen, G., Pham, K.D., Blasch, E.P.: A Trust-Based Sensor Allocation Algorithm in Cooperative Space Tracking Problems, SPIE Defense and Security 2011: Sensors and Systems for Space Applications IV, Proceedings of SPIE, Vol. 8044, Orlando, FL, 2011
16. Wang, J., Xin, M.: Multi-agent consensus algorithm with obstacle avoidance via optimal control approach. Int. J. Contr. **83**(12), 2606–2621 (2010)
17. Whittle, P.: Risk Sensitive Optimal Control. Wiley, Newyork (1990)

Informational Issues in Decentralized Control

Meir Pachter and Khanh Pham

Abstract The long-standing decentralized optimal control problem posed by Witsenhausen is analyzed and the underlying modeling issues are discussed. The strong connection to communication theory is highlighted. Informational aspects are emphasized and it is shown that, to some extent, Witsenhausen's decentralized optimal control problem is somewhat contrived.

Keywords Decentralized Control • Game Theory • Witsenhausen's Problem

1 Introduction

Informational issues in decentralized control are discussed. In this regard, Witsenhausen's problem [1] is the simplest decentralized optimal control problem. Indeed, the "simple" LQG control problem posed by Witsenhausen in his seminal paper made a great splash when it first appeared in 1968 because the optimal linear strategy is not optimal. This is caused by the nonclassical information pattern. Since then numerous attempts have been made in the intervening 45 years to obtain a better estimate of the minimal cost. A special session devoted to the Witsenhausen problem was held at the 2008 CDC and this stimulated renewed interest in the subject matter. Recently numerous additional papers on Witsenhausen's counterexample have appeared, and this problem was also extensively addressed in the most recent CDCs [2–6]. There is a fascination with Witsenhausen's counterexample in control circles and for good reason: It touches on the dual control issue where one needs to strike a balance between exploration and exploitation—this, exclusively due to the information pattern, which is non-nested. However, the objective of this article is not to survey the field, nor is it to further slightly improve the estimate of the minimal cost—the current best estimate of the minimal cost for

M. Pachter (✉)
Air Force Institute of Technology, Wright-Patterson A.F.B., OH 45433, USA
e-mail: meir.pachter@afit.edu

K. Pham
Air Force Research Laboratory, Kirtland A.F.B., NM 87117, USA
e-mail: khanh.pham@kirtland.af.mil

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_3

the "canonical" problem parameters currently stands at 0.1670790. The objective is to gain an understanding of the underlying engineering or physical problem that Witsenhausen's mathematical model is addressing. In this respect, the coupled communications and control aspects of Witsenhausen's problem are discussed and the attendant informational issues are carefully examined. Also, our aim is to gain physical insight into a range of methods for obtaining suboptimal solutions and, by doing so, dispel some of Witsenhausen's counterexample mystique. The strong connection to communication theory is emphasized and the informational aspects of the problem are highlighted. The latter seem to direct one to the conclusion that Witsenhausen's decentralized optimal control problem is to some extent contrived.

The paper is organized as follows. In Sect. 2 Witsenhausen's problem is carefully stated. In Sect. 3 the communications aspect of Witsenhausen's decentralized LQG optimal control problem are analyzed and the connection to detection theory is elucidated in Appendix A. The special case where the "receiver's" noise floor is high is analyzed in Sect. 4 and the optimal modulation and detection scheme is shown to be linear. Concluding remarks are made in Sect. 5.

2 Witsenhausen's Problem Statement

In Witsenhausen's paper [1] the following decentralized LQG optimal control problem is considered.

1. *Dynamics*: The discrete-time dynamics are linear and scalar, and the planning horizon is $N = 2$. There are two "players," Player 1 and Player 2. Player 1 acts at decision time $k = 0$ and his input is u_0:

$$x_1 = x_0 + u_0, \; x_0 \sim \mathcal{N}(0, \sigma_0^2). \tag{1}$$

   Player 2 acts at decision time $k = 1$ and his input is u_1:

$$x_2 = x_1 - u_1. \tag{2}$$

2. *Information*: The information of Player 1 at his decision time $k = 0$ is the initial state x_0. The information available to Player 2 at his decision time $k = 1$ is a noise corrupted measurement of the state at time $k = 1$,

$$z_1 = x_1 + v_1, \;\; v_1 \sim \mathcal{N}(0, \sigma^2). \tag{3}$$

   Thus, the initial state x_0 is the private information of Player 1 which is not shared with Player 2. Player 1 has perfect information at his decision time $k = 0$, whereas Player 2 which acts at time $k = 1$ has access to the noise corrupted

measurement z_1 of the state x_1. Thus, at his decision time $k = 1$, Player 2 has imperfect information on the state x_1. In addition, Player 2 does not know the control u_0 of Player 1.

However, note that in [1], and in Eq. (1), it is also stated that the initial state x_0 is a random variable which is Gaussian distributed. This important point will be further discussed in the sequel.

3. *Strategy*: The so far specified information pattern mandates that the strategy of Player 1 is

$$u_0 = \gamma_0(x_0) \tag{4}$$

and the strategy of Player 2 is

$$u_1 = \gamma_1(z_1). \tag{5}$$

4. *Payoff*: The cost function, which both players strive to minimize, is

$$J(u_0, u_1; x_0) = K^2 u_0^2 + x_2^2. \tag{6}$$

Player 1, who has perfect state information, has a penalty K^2 on his control effort, whereas the control effort of Player 2, whose states' measurements are corrupted by noise, is free. Both players want the terminal state x_2 to be "small" while at the same time, the control effort, exclusively expanded by Player 1, also to be "small." Evidently, Player 2 could effortlessly make the terminal state $x_2 \approx 0$, if only he knew the state x_1 with a high degree of precision.

As far as Player 1 is concerned, the random variable in the problem statement is the measurement error v_1 of Player 2. As far as Player 2 is concerned, the random variables in the problem statement are the initial state x_0 and his measurement error v_1. The players are interested in the expectation of the cost function (6), which requires Player 1 to take the expectation on the random variable v_1 and Player 2 takes the expectation on the random variables v_1 and x_0. Since the Players 1 and 2 have private information, x_0 and z_1, respectively, their expected costs will be conditional on their private information. This brings us into the realm of nonzero-sum games.

Witsenhausen's decentralized control problem is schematically illustrated in Fig. 1. In Fig. 1 we have allowed for a more general initial state specification,

$$x_0 \sim \mathcal{N}(\overline{x}_0, \sigma_0^2) \tag{7}$$

and, without loss of generality, have set the parameter $\sigma = 1$; if $\sigma = 0$, the optimization problem is trivial—obviously, the optimal controls are $u_0^* = 0$ and $u_1^* = z_1$. The other extreme case where $\sigma \to \infty$ will be addressed in Sect. 4. Thus, the problem parameters are K and σ_0.

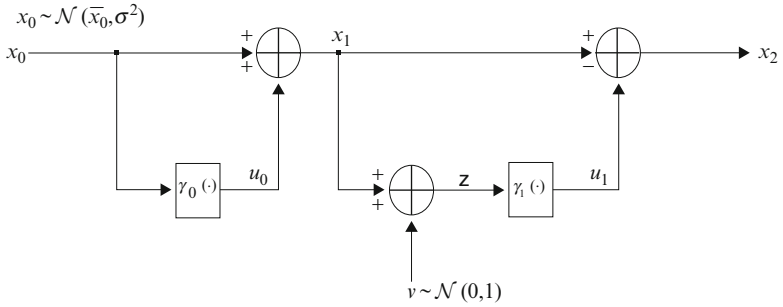

Fig. 1 Witsenhausen's decentralized control problem

2.1 Initial State Information

In Witsenhausen's paper it is stated that the initial state x_0 is *random*, Gaussian distributed, and its statistics are specified according to Eq. (1), namely,

$$x_0 \sim \mathcal{N}(0, \sigma_0^2).$$

The informational aspect of the specification of the initial state's statistics merits special attention.

In Witsenhausen's paper [1], attention is confined to the special case where the statistic $\overline{x}_0 = 0$. It stands to reason that the consideration of a more general initial state is warranted, s.t. its statistics are specified by Eq. (7)—see also Fig. 1. Indeed, including in the problem statement the initial state's statistics immediately begs the question: why not allow for a more generic form of the initial state information, as specified in Eq. (7), whereupon the issue of who knows what, when, is immediately brought up; exclusively considering the special case where the initial state's statistic $\overline{x}_0 = 0$ tends to hide the importance of the $\overline{x}_0$ information. This brings to the forefront the question of whether Player 1, or Player 2, or both players are privy to the $\overline{x}_0$ information.

Consider Eqs. (7) or (1). This can be construed to mean that at time $k = 0$, Player 2 took a measurement of the initial state x_0 so that at his decision time $k = 1$, Player 2, in addition to the measurement z_1, also has imperfect information about the state x_0. Thus, the strategy (5) of Player 2 should in fact be replaced by the strategy

$$u_1 = \gamma_1(\overline{x}_0, z_1). \tag{8}$$

Whether or not Player 1 is informed on the initial state's measurement taken by Player 2 is an important question. If indeed the measurement information $\overline{x}_0$ is shared with Player 1 then his strategy (4) should be replaced by the strategy

$$u_0 = \gamma_0(x_0, \overline{x}_0). \tag{9}$$

Clearly, should at time $k = 0$ the information on the measurement $\overline{x}_0$ of Player 2 be shared with Player 1, Witsenhausen's control problem would be somewhat less decentralized than it appears to be at first glance.

Alternatively, one might argue that the specifications in Eqs. (1) or (7) mean that although it is so that at his decision time $k = 0$, Player 1 has perfect state information, he might be aware of the fact that the initial state x_0 presented to him at time $k = 0$ will actually be drawn from the distributions (1) or (7). This is critical information if a degree of information sharing among the players before the kickoff of this seemingly decentralized control problem is in fact allowed, or is required, to take place—in which case Player 1 employs a *prior commitment* strategy and the information $\overline{x}_0$ plays a crucial role in its synthesis. In this case, the strategy of Player 1 will take the somewhat unconventional form of Eq. (9). Concerning Eq. (9), one could then wonder why would Player 1 want to know the statistic of the initial state, since he already knows the initial state proper, but as already stated above, this additional information is required in order for Player 1 to be able to synthesize his prior commitment strategy. In this instance, whether or not also Player 2 is informed on the initial state's statistics will be discussed in the sequel. Suffice it to say that if Player 2 is informed about the statistic $\overline{x}_0$, the strategy (5) of Player 2 will be replaced by the strategy (8). In this case the strategy (9) of Player 1, which acts at time $k = 0$, could be viewed as a delayed commitment strategy which encodes the fact that at his decision time $k = 0$, Player 1 knows that Player 2, who's turn will come at time $k = 1$, knows the statistic $\overline{x}_0$ of the initial state x_0. In other words, both players know that the initial state x_0 will be drawn from a distribution (1) or (7), except that, prior to his move at time $k = 0$, Player 1 is given the initial state information.

When the initial state's measurement/information $\overline{x}_0$ is available to Player 2 and at time $k = 0$ this information is shared with Player 1, then Player 1, who has perfect information on the initial state x_0, also knows that Player 2 knows that the initial state is distributed according to Eqs. (1) or (7). By virtue of this fact, the strategy of Player 1, which incorporates all the information available to him, takes the rather unusual form (9). This point, whether or not at his decision time $k = 0$ Player 1 knows that Player 2 knows that the initial state is distributed according to Eqs. (1) or (7), will be emphasized in Sect. 4.

Suppose the measurement (7) of the initial state taken by Player 2 is shared with Player 1, which now uses the strategy (9). In other words, the result of the initial state measurement of Player 2 is communicated to Player 1. Alternatively, suppose the information (1), or (7), of Player 1 concerning the p.d.f. wherefrom the initial state will be drawn is communicated to Player 2 before the start of the game. In both cases the information $\overline{x}_0$ is shared at time $k = 0$ and the respective strategies of Players 1 and 2 will be specified by Eqs. (9) and (8). This is the operational information pattern in Witsenhausen's paper. The control problem at hand now assumes a less decentralized character.

2.2 Cost Function

The ramifications as far as the cost functional is concerned are as follows. Since $\overline{x}_0$ is public information, the strategies' dependence on $\overline{x}_0$ is suppressed. The fact that the players have private information would naturally lead to a formulation where each player strives to minimize his own cost function: in the case of Player 1 it would be the expectation on v_1 of the cost function (6), conditional on his private information x_0, and in the case of Player 2 it would be the expectation on x_0 and v_1 of the cost function (6), conditional on his private information z_1. The players' strategies would be delayed commitment strategies, which means that the optimization of their respective cost functions would be performed in the Euclidean space R^1.

Although the players have private information, it is nevertheless stipulated in [1] that both players are minimizing a common cost functional, namely the expectation on x_0 and v_1 of the cost function (6): according to Witsenhausen [1] and the many papers written on Witsenhausen's problem, the cost functional is

$$J(\gamma_0, \gamma_1; \overline{x}_0) = E_{x_0,v_1}(K^2(\gamma_0(x_0))^2 + (x_0 + \gamma_0(x_0) - \gamma_1(x_0 + \gamma_0(x_0) + v_1))^2). \tag{10}$$

This definition of the common cost functional is made possible by the fact that both players share the information on the statistics (1), or (7), of the initial state x_0. Now, the players' strategies $\gamma_0(x_0)$ and $\gamma_1(z_1)$ are prior commitment strategies. Indeed, Player 2 now employs a prior commitment strategy—he decides on his optimal strategy $\gamma_1(\cdot)$ ahead of time, before ever receiving the measurement z_1, which does not feature in Witsenhausen's cost functional. In order for the cost of Player 2 which employs a prior commitment strategy to equal the average realized cost of Player 1, the latter is minimizing the expectation of the cost (6), taken not just over the random variable v_1 but also over the initial state x_0. This gives the appearance of Player 1 playing a *prior commitment* game, that is, instead of determining his control u_0 upon receiving his initial state information x_0, Player 1 determines his *strategy* function $\gamma_0(\cdot)$ ahead of time. Indeed, his private information x_0 does not feature in Witsenhausen's cost functional. The "prior commitment" aspect of the strategy of Player 1 in this inherently cooperative game further manifests itself in the scenario discussed in Sect. 3 where *prior communication* is allowed, that is, the players are allowed to communicate before kickoff time, in which case Player 1 informs Player 2 about his optimal or suboptimal *strategy* prior to time $k = 0$. This requires Player 1 to be privy to the statistic $\overline{x}_0$ of the initial state x_0, as indeed he must be if he is to minimize Witsenhausen's cost functional. Evidently, the requirement of prior communication further diminishes the much touted decentralized control character of Witsenhausen's problem.

The dynamic "game" is now in *normal form*, is static, and is not in *extensive form*. The price to pay for transforming the dynamic game into normal form is that prior commitment strategies are used. When the game is in extensive form and delayed

commitment strategies are used, part of the optimization is carried out in Euclidean space. Now that prior commitment strategies are used, the optimization must be performed in function space.

3 Communication and Control

To gain an understanding of, and insight into, the decentralized optimal control problem at hand, the communications context of Witsenhausen's decentralized control problem is now discussed. By directly viewing the decentralized optimal control problem (1)–(5), (7)–(10) as a communications problem, which indeed it is, and was originally perceived by Witsenhausen, one realizes that notwithstanding the fact that the cost function is quadratic in the controls, the problem at hand entails the minimization of a non-convex and complex cost functional, which is a hard problem. At the same time, a hierarchy of suboptimal solutions readily suggests itself. This must have been the motivation for Witsenhausen's original counterexample in the first place [1]. However, rather than directly viewing Witsenhausen's problem as an optimization problem in function space, we shall dwell on the physical meaning of the mathematical problem posed by Witsenhausen. By emphasizing the communications context of Witsenhausen's decentralized optimal control problem one opens wide the door to the synthesis of a family of suboptimal solutions of Witsenhausen's decentralized optimal control problem, as evidenced by the rich current literature. In doing so we hope to dispel some of Witsenhausen's counterexample mystique.

A communication problem is considered where both Player 1, the transmitter, and Player 2, the receiver, know that the "message" x_0 will be selected according to the probability distribution specified by Eq. (7). Player 1 will encode the information x_0 according to

$$x_1 = f(x_0)$$

before sending x_1 over a Gaussian communication channel whose statistics are specified by Eq. (3). The optimal function $f(\cdot)$ must be determined. It is also specified that the cost to Player 1 of encoding the information x_0 is $K^2(f(x_0)-x_0)^2$. Player 2, using his measurement z_1, estimates the received signal x_1. Player 2 strives to reduce the variance of the estimation error of x_1. Both players want to minimize the average expected total cost of encoding and decoding the transmitted signal. As such, this formulation models a communication problem, but not a decentralized optimal control problem or a dynamic game.

1. In the context of a communication problem, we need to assume that the initial state's information/measurement (1), or (7) of Player 2, is shared with Player 1.

Thus, both players know that the initial state will be drawn from the Gaussian distribution (7), that is, the initial state's p.d.f. is

$$\phi(x_0) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\overline{x}_0)^2}{2\sigma_0^2}}.$$

2. Furthermore, suppose that before the "game" Player 1 and Player 2 come together and agree that upon receiving his x_0 information, Player 1 will choose his control u_0 to make sure that the state at time $k = 1$ is either $x_1 = \overline{x}_0 + b$ or $x_1 = \overline{x}_0 - b$. Thus, the state x_1 is quantized; the amplitude $b \geq 0$ will jointly be determined by the players before the start of the game. Now, in order to keep the control cost low and thus keep the cost low, the choice of Player 1's control u_0 will be dictated by the distance of x_0 from the two "guideposts" $\overline{x}_0 + b$ and $\overline{x}_0 - b$. In other words, the strategy of Player 1 will be

$$\gamma_0(x_0, \overline{x}_0) = \begin{cases} \overline{x}_0 + b - x_0 \; if \; x_0 \geq \overline{x}_0, \\ \overline{x}_0 - b - x_0 \; if \; x_0 < \overline{x}_0. \end{cases} \tag{11}$$

The strategy of Player 1 is illustrated in Fig. 2, where the function

$$f(x_0, \overline{x}_0) \equiv x_1 = x_0 + \gamma_0(x_0, \overline{x}_0)$$

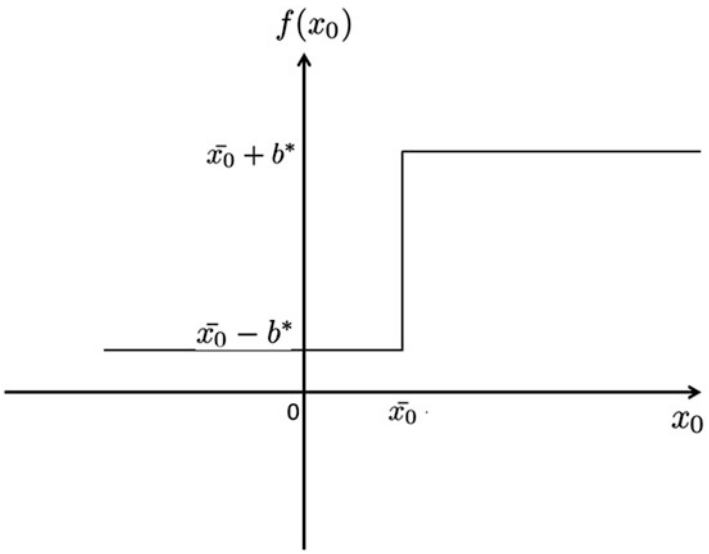

Fig. 2 Strategy of Player 1

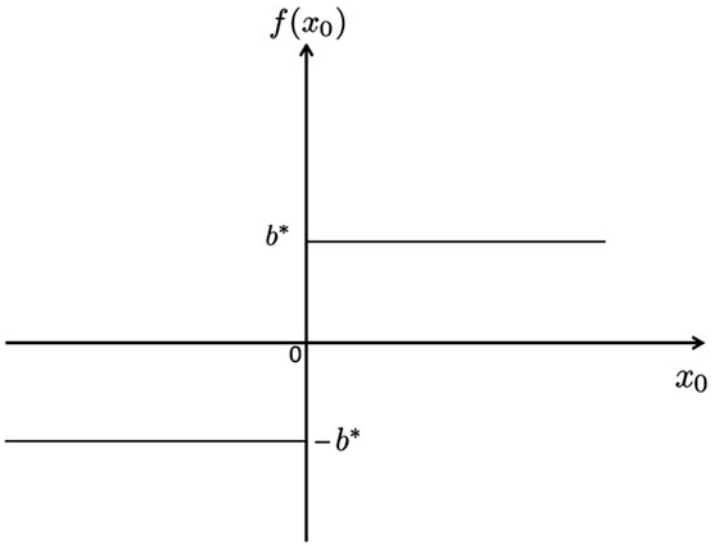

Fig. 3 Strategy of Player 1, $\overline{x}_0 = 0$

and in the special case where $\overline{x}_0 = 0$, as in Witsenhausen's counterexample, the strategy of Player 1 is shown in Fig. 3. The dependence of the strategy of Player 1 on the $\overline{x}_0$ information is here suppressed.

Since the players come together before the game starts and a cooperative game is played, Player 2 knows the strategy (11) of Player 1. Consequently, at decision time $k = 1$, Player 2 knows that the true state at time $k = 1$ is either $x_1 = \overline{x}_0 + b$ or $x_1 = \overline{x}_0 - b$. Moreover, Player 2 calculates the prior probabilities

$$\mathcal{P}(x_1 = \overline{x}_0 + b) = \mathcal{P}(x_1 = \overline{x}_0 - b) = \frac{1}{2}.$$

Hence, the decision process of Player 2 is now greatly simplified. Player 2 is faced with a binary choice: based on his measurement z_1, he will have to decide whether the true state x_1 is $\overline{x}_0 + b$ or $\overline{x}_0 - b$ whereupon he will apply his costless control to hopefully set the state $x_2 = 0$ and thus, to the best of his ability, reduce the cumulative cost. Indeed, a typical *communications* scenario is at hand where Player 1 sends one of two possible letters, $\overline{x}_0 + b$ or $\overline{x}_0 - b$, over a Gaussian channel and the job of Player 2 is to *detect* the transmitted letter. To minimize his control effort, Player 1 will decide on which letter to transmit according to its distance from the random initial state x_0 which is known to him.

In view of the strategy (11) employed by Player 1, his expected share of the incurred cost, J_1, will be

$$J_1 = K^2\left[\int_{\overline{x}_0}^{\infty}(\overline{x}_0 + b - x_0)^2\phi(x_0)dx_0 + \int_{-\infty}^{\overline{x}_0}(\overline{x}_0 - b - x_0)^2\phi(x_0)dx_0\right]$$
$$= K^2\left[\int_{\overline{x}_0}^{\infty}(x_0 - \overline{x}_0)^2\phi(x_0)dx_0 - 2b\int_{\overline{x}_0}^{\infty}(x_0 - \overline{x}_0)\phi(x_0)dx_0\right.$$
$$+b^2\int_{\overline{x}_0}^{\infty}\phi(x_0)dx_0 + \int_{-\infty}^{\overline{x}_0}(x_0 - \overline{x}_0)^2\phi(x_0)dx_0$$
$$\left.+2b\int_{-\infty}^{\overline{x}_0}(x_0 - \overline{x}_0)\phi(x_0)dx_0 + b^2\int_{-\infty}^{\overline{x}_0}\phi(x_0)dx_0\right],$$

that is,

$$J_1 = K^2\left[\int_{-\infty}^{\infty}(x_0 - \overline{x}_0)^2\phi(x_0)dx_0 + b^2\int_{-\infty}^{\infty}\phi(x_0)dx_0\right.$$
$$\left.+ 2b\left(\int_{-\infty}^{0}x\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{x^2}{2\sigma_0^2}}dx - \int_{0}^{\infty}x\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{x^2}{2\sigma_0^2}}dx\right)\right]$$
$$= K^2\left[\sigma_0^2 + b^2 - 4b\int_{0}^{\infty}x\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{x^2}{2\sigma_0^2}}dx\right]. \tag{12}$$

Now,

$$\int xe^{-\frac{x^2}{2\sigma_0^2}}dx = -\sigma_0^2 e^{-\frac{x^2}{2\sigma_0^2}}$$

and therefore

$$\int_{0}^{\infty}xe^{-\frac{x^2}{2\sigma_0^2}}dx = \sigma_0^2.$$

Inserting this expression into Eq. (12) yields the cost component

$$J_1(b) = K^2\left(b^2 - \frac{4}{\sqrt{2\pi}}\sigma_0 b + \sigma_0^2\right), \quad b \geq 0.$$

The expected contribution of the actions of Player 1 to the cost functional (10), J_1, is parameterized by his choice of the amplitude/signalling level b. In [1], Witsenhausen chose the amplitude $b = \sigma_0$.

Remark. The expected contribution of Player 1 to the cost functional (10) is minimized when the signalling level $b^* = \sqrt{\frac{2}{\pi}}\sigma_0$, whereupon

$$J_1^* = \left(1 - \frac{2}{\pi}\right)(K\sigma_0)^2. \tag{13}$$

The strategy of Player 2 is the following detection algorithm:

$$\gamma_1(\overline{x}_0, z_1) = \begin{cases} \overline{x}_0 + b \text{ if } z_1 \geq \overline{x}_0, \\ \overline{x}_0 - b \text{ if } z_1 < \overline{x}_0. \end{cases} \tag{14}$$

We now draw the reader's attention to the fact that for Witsenhausen's formulation of the control problem to model a communications scenario and be consistent, one must tacitly assume that before the kickoff the initial state's statistic information $\overline{x}_0$ is shared with Player 2, as is evident in Eq. (14) above. This, of course, would draw less attention and would appear to be less of an issue if $\overline{x}_0 = 0$, whereupon the strategy of Player 2

$$\gamma_1(z_1) = \begin{cases} b \text{ if } z_1 \geq 0 \\ -b \text{ if } z_1 < 0 \end{cases}$$

would somewhat misleadingly look like Eq. (5).

Concerning the contribution of Player 2 to the cost functional (9), we calculate the probabilities of the possible outcomes:

$$\begin{aligned} \mathcal{P}(x_2 = 0) &= \mathcal{P}(x_0 \geq \overline{x}_0, z_1 \geq \overline{x}_0) + \mathcal{P}(x_0 < \overline{x}_0, z_1 < \overline{x}_0) \\ &= \mathcal{P}(x_0 \geq \overline{x}_0, \overline{x}_0 + b + v_1 \geq \overline{x}_0) + \mathcal{P}(x_0 < \overline{x}_0, \overline{x}_0 - b + v_1 < \overline{x}_0) \\ &= \mathcal{P}(x_0 \geq \overline{x}_0, v_1 \geq -b) + \mathcal{P}(x_0 < \overline{x}_0, v_1 < b) \\ &= \mathcal{P}(x_0 \geq \overline{x}_0)\mathcal{P}(v_1 \geq -b) + \mathcal{P}(x_0 < \overline{x}_0)\mathcal{P}(v_1 < b) \\ &= \frac{1}{2}\mathcal{P}(v_1 \geq -b) + \frac{1}{2}\mathcal{P}(v_1 < b) \\ &= \frac{1}{2}[\mathcal{P}(v_1 \geq -b) + \mathcal{P}(v_1 < b)] \\ &= \frac{1}{2}[\mathcal{P}(v_1 < b) + \mathcal{P}(v_1 < b)] \\ &= \mathcal{P}(v_1 < b) \\ &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right], \end{aligned}$$

where

$$\text{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Now,

$$\overline{x}_0 + b - (\overline{x}_0 - b) = 2b$$

and we calculate

$$\begin{aligned}\mathcal{P}(x_2 = 2b) &= \mathcal{P}(x_0 \geq \overline{x}_0, z_1 < \overline{x}_0)\\ &= \mathcal{P}(x_0 \geq \overline{x}_0, \overline{x}_0 + b + v_1 < \overline{x}_0)\\ &= \mathcal{P}(x_0 \geq \overline{x}_0, v_1 < -b)\\ &= \mathcal{P}(x_0 \geq \overline{x}_0)\mathcal{P}(v_1 < -b)\\ &= \frac{1}{2}\mathcal{P}(v_1 < -b)\\ &= \frac{1}{4}\left[1 - \operatorname{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right].\end{aligned}$$

Similarly,

$$\overline{x}_0 - b - (\overline{x}_0 + b) = -2b$$

and we calculate

$$\begin{aligned}\mathcal{P}(x_2 = -2b) &= \mathcal{P}(x_0 < \overline{x}_0, z_1 > x_0)\\ &= \mathcal{P}(x_0 < \overline{x}_0, \overline{x}_0 - b + v_1 > x_0)\\ &= \mathcal{P}(x_0 < \overline{x}_0, v_1 > b)\\ &= \mathcal{P}(x_0 < \overline{x}_0)\mathcal{P}(v_1 > b)\\ &= \frac{1}{2}\mathcal{P}(v_1 > b)\\ &= \frac{1}{4}\left[1 - \operatorname{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right].\end{aligned}$$

This allows us to calculate the expected contribution of Player 2 to the cost functional (10),

$$\begin{aligned}J_2(b) &= 0 \cdot \mathcal{P}(x_2 = 0) + 4b^2 \cdot \mathcal{P}(x_2 = 2b) + 4b^2 \cdot \mathcal{P}(x_2 = -2b)\\ &= 2\left[1 - \operatorname{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right]b^2.\end{aligned}$$

The total cost, $J_1 + J_2$, is

$$J(b) = \left[K^2 + 2\left(1 - \operatorname{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right)\right]b^2 - 2\sqrt{\frac{2}{\pi}}K^2\sigma_0 b + (K\sigma_0)^2.$$

Without loss of generality assume $\sigma = 1$. Also set

$$b := \frac{1}{\sqrt{2}} b$$

so that as a function of the yet to be determined signalling level b, the cumulative cost (10) is

$$J(b) = 2[K^2 + 2(1 - \text{erf}(b))]b^2 - \frac{4}{\sqrt{\pi}} K^2 \sigma_0 b + (K\sigma_0)^2. \tag{15}$$

When Player 1 uses a binary signalling level of $\pm b$—see Eq. (11)—and Player 2 uses the detection strategy (14), the cost function is $J(b)$. It can be further reduced if Player 2 modifies his strategy as follows. Since he cannot be absolutely sure about the correct outcome of the detection step as specified by the strategy (14), Player 2 hedges his bet, does not go all the way, and uses a modified strategy, parameterized by $\alpha \in R^1, \mid \alpha \mid << 1$:

$$u_1 = \gamma_1(\overline{x}_0, z_1) = \begin{cases} (1-\alpha)(\overline{x}_0 + b) \; if \; z_1 \geq \overline{x}_0, \\ (1-\alpha)(\overline{x}_0 - b) \; if \; z_1 < \overline{x}_0. \end{cases}$$

As a result, the contribution of Player 2 to the expected cost is

$$\begin{aligned} J_2(\alpha, b; \overline{x}_0) &= \frac{1}{4}\left[1 + \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right][\alpha^2(\overline{x}_0 + b)^2 + \alpha^2(\overline{x}_0 - b)^2] \\ &\quad + \frac{1}{4}\left[1 - \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right][(\overline{x}_0 + b - (1-\alpha)(\overline{x}_0 - b))^2 \\ &\quad + (\overline{x}_0 - b - (1-\alpha)(\overline{x}_0 + b))^2] \\ &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2 + b^2)\alpha^2 \\ &\quad + \frac{1}{2}\left[1 - \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2 + b^2) \\ &\quad + \frac{1}{2}\left[1 - \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2 + b^2)(1-\alpha)^2 \\ &\quad - \left[1 - \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2 - b^2)(1-\alpha) \\ &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2 + b^2)\alpha^2 \\ &\quad + \frac{1}{2}\left[1 - \text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2 + b^2)(2 - 2\alpha + \alpha^2) \end{aligned}$$

$$-\left[1-\text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right](\overline{x}_0^2-b^2)(1-\alpha)$$
$$=\alpha^2\overline{x}_0^2+(2-2\alpha+\alpha^2)b^2-2(1-\alpha)b^2\text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right).$$

The minimum of J_2 is attained when the parameter

$$\alpha^*=\frac{b^2}{\overline{x}_0^2+b^2}\left[1-\text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right]\ (\geq 0)$$

When $\overline{x}_0=0$, as in Witsenhausen's counterexample, the optimal parameter

$$\alpha^*=1-\text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)$$

and the expected contribution of Player 2 to the cost is reduced to

$$J_2^*(b)=\left[1-\text{erf}^2\left(\frac{b}{\sqrt{2}\sigma}\right)\right]b^2.$$

By optimally hedging his bets, Player 2 has reduced by 50 % his expected contribution to the cost (10). Thus, when $\overline{x}_0=0$, as in Witsenhausen's counterexample, the expected cumulative cost $J=J_1+J_2^*$ is

$$J(b)=(K\sigma_0)^2+K^2\left(b^2-\frac{4}{\sqrt{2\pi}}\sigma_0 b\right)+\left[1-\left(\text{erf}\left(\frac{b}{\sqrt{2}\sigma}\right)\right)^2\right]b^2.$$

Without loss of generality, assume $\sigma=1$. Also set

$$b:=\frac{1}{\sqrt{2}}b$$

so that the expected cumulative cost as a function of the amplitude/signalling level b is

$$J(b)=2[K^2+1-(\text{erf}(b))^2]b^2-\frac{4}{\sqrt{\pi}}K^2\sigma_0 b+(K\sigma_0)^2\ ,\ \ b>0. \tag{16}$$

Assume the parameter $\sigma_0>>1$. The cumulative cost (10) then attains a local minimum at $b^*\approx\frac{1}{\sqrt{\pi}}\sigma_0$ and consequently the value of the functional (10) is

$$J^*\approx\left(1-\frac{2}{\pi}\right)(K\sigma_0)^2. \tag{17}$$

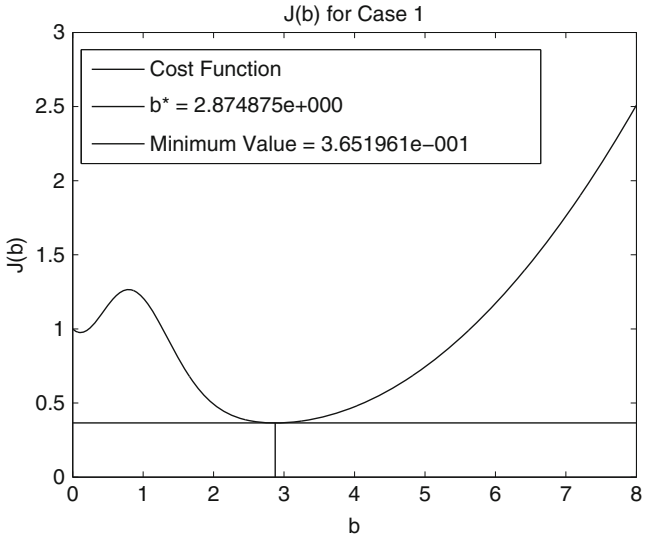

Fig. 4 Expected cost as a function of signalling level

This is also the minimal contribution of Player 1 to the expected cost—see Eq. (13). Hence, we conclude that when $\sigma_0 >> 1$ and the binary signalling communications protocol is used, the global minimum is given by Eq. (13) and it is attained when Player 1 uses the amplitude/signalling level

$$b^* = \sqrt{\frac{2}{\pi}}\sigma_0.$$

We will focus on the benchmark/canonical scenario where the parameters $K = 0.2$ and $\sigma_0 = 5$ (> 1). The cumulative expected cost function (16) is depicted in Fig. 4. The optimal binary signalling level is

$$b^* = 2.874875$$

and the expected minimal cumulative cost

$$J^* = 0.3651961.$$

As expected, and by design, in this scenario, the contribution of Player 2 to the cumulative cost is small. From Fig. 4 it is also plainly evident that the cost function is not convex and it has a local minimum.

Binary signalling was also used in Witsenhausen's original counterexample [1] where the signalling level is

$$b = \sigma_0$$

but in [1] the strategy $\gamma_1^*(z_1)$ of Player 2 did not entail a detection protocol and instead a *continuous* function of his measurement z_1 is used. The expected cumulative cost in Witsenhausen's paper is

$$J_W^* = 0.404253$$

and we see that

$$J_W^* > J^* = 0.3651961.$$

Players 1 and 2 could agree on a multilevel signalling/communications protocol—a staircase-like rendition of a multilevel signalling protocol [6, 7], is shown in Fig. 5. This allows Player 1 to choose the signalling level which is closest to the randomly selected initial state x_0, thereby reducing his control effort. At the same time, the use of multiple signalling levels complicates the detection task of Player 2 and the probability of him erring increases. This, in turn, increases the expected contribution

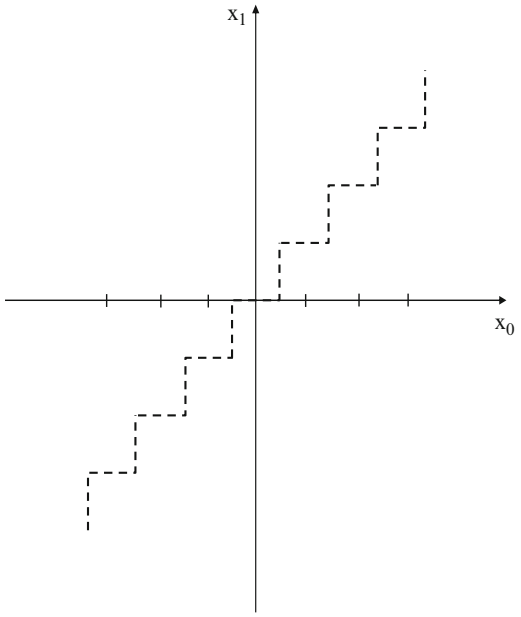

Fig. 5 Multilevel threshold strategy of Player 1

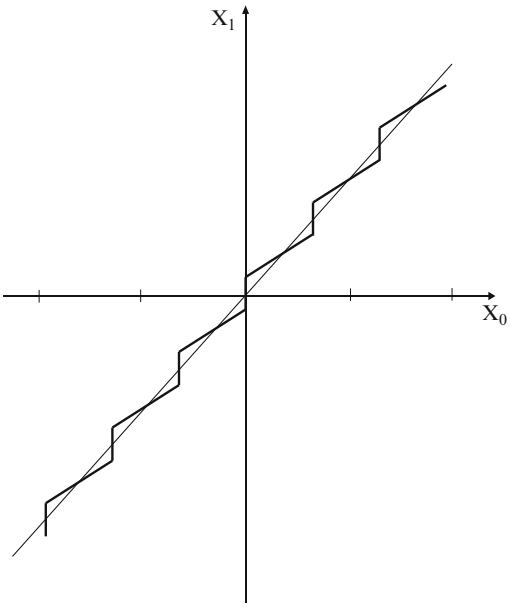

Fig. 6 Multilevel graded strategy of Player 1

of Player 2 to the cumulative cost and a trade-off is needed. A $3\frac{1}{2}$-level signalling protocol, as in [3], reduces the expected cost to $J^* = 0.1673132$.

A piecewise constant staircase signalling protocol—see Fig. 5—is not optimal and the horizontal rungs should be slightly slanted upward. The cost can be further reduced if Player 1 uses a continuous signalling protocol as illustrated in Fig. 6.

The best result so far, using a continuous, monotonically increasing, signalling function and a continuous "detection" algorithm, is reported in [4]. In [4] the parameters are $K = 0.5$, $\sigma_0 = 10$, and $K = 0.25$, $\sigma_0 = 10$. When our suboptimal binary signalling protocol is employed and, as in [4], the "non-canonical" parameters are $K = 0.5$ and $\sigma_0 = 10$, the cost function $J(b)$ is shown in Fig. 7. The optimal binary signalling level is then

$$b^* = 5.645646$$

and the expected minimal cumulative cost

$$J^* = 9.084513.$$

When, as in [4], the problem parameters are $K = 0.25$, $\sigma_0 = 10$ and our suboptimal binary signalling protocol is employed, the cost function $J(b)$ is shown in Fig. 8. The optimal signalling level is

$$b^* = 5.645646$$

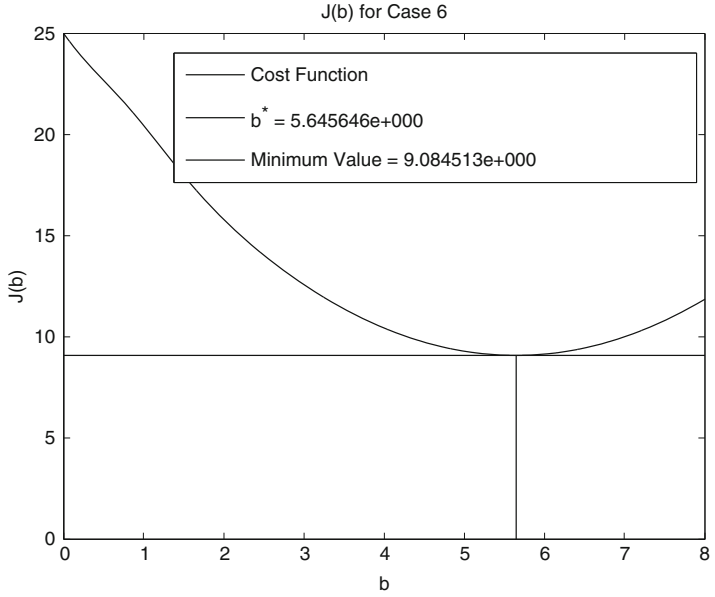

Fig. 7 Expected cost as a function of signalling level

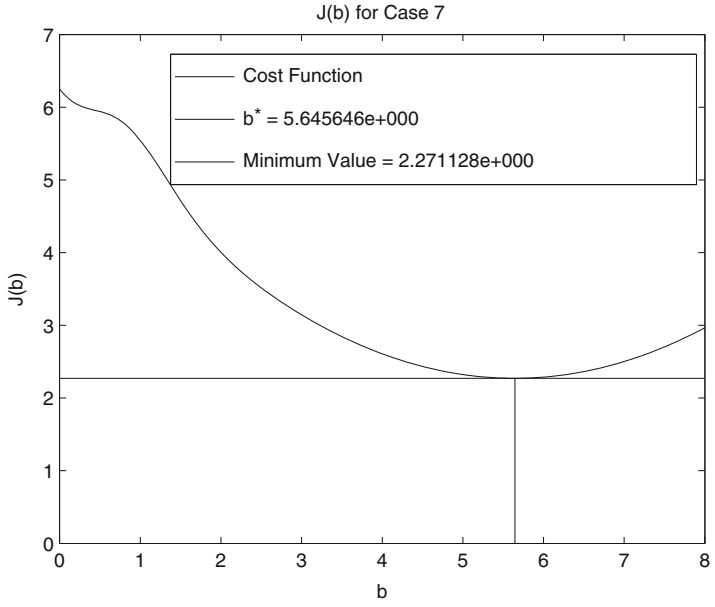

Fig. 8 Expected cost as a function of signalling level

and the expected minimal cumulative cost

$$J^* = 2.271128.$$

In summary, in decentralized control, the players are not supposed to communicate, and yet in Witsenhausen's communication and control scenario, the players share the initial state's statistic $\overline{x}_0$ and, moreover, are tacitly allowed to communicate before the kickoff of the cooperative "game" and establish a communication protocol—one thus refers to signalling. Indeed, in Witsenhausen's original counterexample, two letters are transmitted over a Gaussian channel, that is, the signalling strategy of Player 1 is

$$\gamma_0(x_0) = \begin{cases} \sigma_0 & \text{if } x_0 \geq 0, \\ -\sigma_0 & \text{if } x_0 < 0. \end{cases}$$

The job of Player 2, the receiver, is reduced to detection; however in Witsenhausen's counterexample the conventional thresholding—based detection scheme is not used, and instead an optimal continuous decision function akin to a squashing function is used. We have shown that using a detection strategy reduces the cost compared to Witsenhausen's continuous decision function. The reader is referred to Appendix A where the basics of detection theory are outlined.

Thus, while not explicitly presented as such, Witsenhausen's counterexample centers on the design of a suboptimal communications protocol. Note however that communication before the kickoff of the game would not be needed and the optimal control problem would be a bit more decentralized if the players could independently arrive at their respective optimal strategies.

4 No Communication During the Game

Since communicating/signalling over a noise-corrupted channel is so integral to Witsenhausen's decentralized control problem suboptimal solutions, it is instructive to take things to an extreme and consider the following decentralized optimal control problem where during run time no communication is to take place: the decentralized control problem is specified by Eqs. (1), (2), (4), (6) and the initial state information is specified in (1), or (7), as before, but now, during runtime, Player 2 receives no information whatsoever, that is, at time $k = 1$ a measurement of the state x_1 is *not* taken by Player 2. In other words, the extreme case is now considered where the parameter $\sigma \to \infty$. It is also assumed that over time the random initial state x_0 of the dynamical system presented to Player 1 will be chosen according to the probability distributions (1) or (7), and this is known to both Players 1 and 2. Thus, the respective strategies of Players 1 and 2 are of the form (9) and

$$u_1 = \gamma_1(\overline{x}_0).$$

We intentionally now take a "wrong turn" and initially analyze this special case from the vantage point of the widely used communications paradigm discussed in Sect. 3, where the players come together and establish a communication protocol before kickoff time.

4.1 The Players Communicate Before the Game

Although Player 2 operates in an open-loop mode and no signalling will take place, not everything is lost: since a cooperative control scenario is considered where the players are allowed to communicate *prior* to the start of the game, they could as well agree that Player 1 will always see to it that, irrespective of the realized initial state x_0, at decision time $k = 1$ the state $x_1 = b$, always; the optimal amplitude b is yet to be determined. Player 1 is now committed to an affine strategy

$$\gamma_0(x_0, \overline{x}_0) = b - x_0 \tag{18}$$

and Player 2 knows this. Hence, Player 2 is absolved of even taking a measurement of x_1—he does not need the measurement z_1 and he will always apply the control

$$u_1 = b,$$

driving the state x_2 to 0.

The amplitude b must be decided on ahead of time. In order to minimize his average cost, Player 1 will calculate the optimal amplitude b^* by solving the optimization problem

$$J_1^* = \min_b E_{x_0}(K^2(b - x_0)^2)$$

which yields

$$b^* = \overline{x}_0$$

so that the optimal strategy of Player 1 is affine and is

$$\gamma_0^*(x_0, \overline{x}_0) = \overline{x}_0 - x_0. \tag{19}$$

Hence, the realized minimal cost is

$$J_1^*(x_0) = K^2(x_0 - \overline{x}_0)^2 \tag{20}$$

and consequently, the average minimal cost of Player 1 is,

$$J_1^* = K^2\sigma_0^2. \tag{21}$$

The minimal expected cost (21) is *not* dependent on the initial state's statistic $\overline{x}_0$. Also,

$$J_2^* = J_1^*.$$

Note: Similar to the stochastic games paradigm, it is assumed throughout that the σ_0 statistic is shared by the players.

Indeed, Player 2 does not need to know the initial state's statistic $\overline{x}_0$: Since Player 1 communicates the amplitude b^* to Player 2 before the start of the game, then, similar to Player 1, upon analyzing the optimization problem at hand, Player 2 will independently arrive at the conclusion that the initial state's statistic is in fact $\overline{x}_0 = b^*$ and he will apply the control

$$u_1^* = \overline{x}_0. \tag{22}$$

Example Assume that prior communication is allowed and $\sigma \to \infty$. When the problem parameters are the canonical parameters $K = 0.2$ and $\sigma_0 = 5$, the "minimal" average cost J_1^* and the expected "minimal" cost J_2^* are

$$J_1^* = J_2^* = 1.$$

In particular, if $\overline{x}_0 = 0$ then the "optimal" strategies are $\gamma_0^*(x_0) = -x_0$ and $u_1^* = 0$.

In conclusion, the strategy (19) of Player 1 and the open-loop optimal control (22) of Player 2 are commensurate with the herein stipulated information pattern, irrespective of whether Player 2 is privy to the initial state's statistic $\overline{x}_0$. For the "minimal" average cost (21) to be realizable when prior communication is allowed, Player 2 does not need to know the initial state's statistic $\overline{x}_0$. This is so because prior communication takes place and the "game" is cooperative. Indeed, prior communication allows Player 2 to infer the initial state's statistic $\overline{x}_0$ on his own.

However, should the initial state's statistic $\overline{x}_0$ be known to both Players 1 and 2, and if, in addition, the strategy (19) of Player 1 and the open-loop control (22) of Player 2 were indeed optimal, that is, $K^2\sigma_0^2$ is *the* minimal expected cost, then no prior communication would be required: based on the public information $\overline{x}_0$ available to them, both players would independently solve the decentralized optimal control problem, arrive at their respective optimal strategies, and calculate their expected costs—this being predicated on the assumption that Player 1 is after minimizing his average cost. Strictly speaking, a unique Nash equilibrium would have been obtained. In this instance, and courtesy of the solution of the optimal control problem, implicit communication would automatically materialize. It however turns out that the modulation strategy (19) of Player 1 which was derived using the communications model from Sect. 3 is *not* optimal, as will be shown in the next section where the optimal solution is derived.

4.2 The Players Do Not Communicate Before the Game

The information pattern is s.t. Player 2 knows that Player 1 knows that his measurement of the initial state x_0 is $\overline{x}_0$, according to the distribution (7), or, alternatively, Player 1 knows that the initial state x_0 presented to him will be drawn from the distribution (7) and Player 2 knows this as well. Thus, the strategy of Player 1 is given by Eq. (9) and the strategy of Player 2 is

$$u_1 = \gamma_1(\overline{x}_0).$$

Since the statistic $\overline{x}_0$ is public information and it is not a random variable, the strategies' dependence on $\overline{x}_0$ is temporarily suppressed and we shall refer to the strategy

$$u_0 = \gamma_0(x_0)$$

of Player 1 and the control

$$u_1 \in R^1$$

of Player 2.

Concerning the correct analysis of the optimization process:

1. First, take the point of view of Player 1, who has been provided the initial state information x_0: Player 1 is playing against the optimal input $u_1^* \in R^1$ of Player 2, and since his strategy is a delayed commitment strategy, no random variables feature in his optimization. Thus, his cost function

   $$J^{(1)}(u_1^*) \equiv \min_{u_0 \in R^1}[K^2 u_0^2 + (x_0 + u_0 - u_1^*)^2]$$

   and consequently his optimal control must satisfy the relationship

   $$u_0^* = -\frac{1}{K^2+1}(x_0 - u_1^*).$$

   Hence, the optimal strategy of Player 1 and the optimal control of Player 2 must satisfy the relationship

   $$\gamma_0^*(x_0) = -\frac{1}{K^2+1}(x_0 - u_1^*). \tag{23}$$

2. Next, take the point of view of Player 2, who is playing against the optimal strategy $\gamma_0^*(x_0)$ of Player 1 and as far as he is concerned, the initial state is a random variable whose p.d.f. is specified by Eq. (7). Thus, his cost functional

   $$J^{(2)}(\gamma_0^*(\cdot)) \equiv \min_{u_1 \in R^1}\{E_{x_0}(\, K^2(\gamma_0^*(x_0))^2 + [x_0 + \gamma_0^*(x_0) - u_1]^2 \mid \overline{x}_0\,)\}$$

and consequently his optimal control must satisfy the relationship

$$u_1^* = E_{x_0}(x_0 + \gamma_0^*(x_0) \mid \overline{x}_0)$$

that is,

$$u_1^* = \overline{x}_0 + E_{x_0}(\gamma_0^*(x_0) \mid \overline{x}_0). \tag{24}$$

Both Players 1 and 2 calculate the expectations of the L.H.S. and R.H.S. of Eq. (23) and obtain the equation

$$E_{x_0}(\gamma_0^*(x_0) \mid \overline{x}_0) = -\frac{1}{K^2+1}(\overline{x}_0 - u_1^*). \tag{25}$$

Inserting Eq. (25) into Eq. (24) yields

$$u_1^* = \overline{x}_0. \tag{26}$$

In other words, the optimal strategy of Player 2 is

$$\gamma_1^*(\overline{x}_0) = \overline{x}_0 \tag{27}$$

and inserting Eq. (26) into Eq. (23) yields the optimal strategy of Player 1

$$\gamma_0^*(x_0, \overline{x}_0) = \frac{1}{K^2+1}(\overline{x}_0 - x_0). \tag{28}$$

Having obtained the optimal strategies, the respective value functions of Players 1 and 2 are calculated as follows:

$$V_0^{(1)}(x_0, \overline{x}_0) = \frac{1}{K^2+1}(x_0 - \overline{x}_0)^2 \tag{29}$$

and

$$V_0^{(2)}(\overline{x}_0) = \frac{1}{K^2+1}\sigma_0^2. \tag{30}$$

The analysis from above is summarized in

Theorem 1. *The special case of Witsenhausen's decentralized optimal control problem (1), (2), (4), and (6), where the parameter $\sigma \to \infty$, but with the slightly more general initial state information specified by Eq. (7), is considered. Thus, the case is considered where at time $k = 1$ a measurement of the state x_1 is not taken by Player 2. The respective optimal strategies of Players 1 and 2 are linear and are given by Eqs. (28) and (27) and their value functions are given by Eqs. (29) and (30).*

Remark. The value of Player 2 is equal to the average cost/value of Player 1, that is,

$$V_0^{(2)}(\overline{x}_0) = E_{x_0}(\, V_0^{(1)}(x_0, \overline{x}_0) \mid \overline{x}_0 \,). \tag{31}$$

Corollary 2. *In the special case where, as in Witsenhausen's paper, the initial state's statistic* $\overline{x}_0 = 0$*, the optimal strategies are*

$$\gamma_0^*(x_0) = -\frac{1}{1+K^2}x_0, \tag{32}$$

and

$$\gamma_1^* = 0 \tag{33}$$

and the players' value functions are

$$V_0^{(1)}(x_0) = \frac{1}{1+K^2}x_0^2 \tag{34}$$

and

$$V_0^{(2)} = \frac{1}{1+K^2}\sigma_0^2. \tag{35}$$

When the parameter (the canonical parameter)

$$K = 0.2$$

the slope of the linear control law of Player 1 is ≈ -1 *and the function*

$$\begin{aligned} f^*(x_0) &\equiv x_0 + \gamma_0^*(x_0) \\ &= \frac{K^2}{1+K^2}x_0 \\ &\approx 0.04x_0. \end{aligned}$$

In summary, in our analysis the somewhat unconventional information pattern germane to Witsenhausen's problem where at time $k = 0$ the information about the initial state's statistic $\overline{x}_0$ is shared among the players and which therefore somewhat detracts from its perceived degree of decentralization, has been retained. It is however important to realize that in this special case of Witsenhausen's problem, where the variance of the measurement error of Player 2 is very large, the *optimal* solution has been obtained. As a result, no additional communication among the players before kickoff time is needed in order to establish a communications protocol, as is tacitly assumed in the case where the parameter σ is finite, the optimization problem is much harder and has not yet been solved, and one must

fall back on suboptimal solutions derived using the communications/signalling paradigm discussed in Sect. 3. Since in the special case treated herein where the parameter $\sigma \rightarrow \infty$ the optimal solution can be independently obtained by both Players 1 and 2, *no* communication prior to kickoff is needed and therefore one is entitled to say that the attendant optimal control problem has been solved in a more decentralized manner.

An additional instance where an *optimal* solution can be easily obtained entails the symmetric information pattern where Player 1 does not have access to the initial state information and both players' information is shared and is specified by Eq. (7); in addition, Player 1 knows that Player 2 has the public information (7) and, vice versa, Player 2 knows that Player 1 has the public information (7). In this case, the strategy of Player 1 is the control $u_0 \in R^1$ and the strategy of Player 2 is the control $u_1 \in R^1$.

1. First, take the point of view of Player 1: Player 1 is playing against the optimal input $u_1^* \in R^1$ of Player 2 and now the random variable x_0 features in his optimization. Thus, his cost function

$$J^{(1)}(u_1^*) \equiv \min_{u_0 \in R^1} E_{x_0}([K^2 u_0^2 + (x_0 + u_0 - u_1^*)^2])$$

and consequently his optimal control must satisfy the relationship

$$u_0^* = -\frac{1}{K^2 + 1}(\overline{x}_0 - u_1^*).$$

2. Next, take the point of view of Player 2, who is playing against the optimal strategy/control u_0^* of Player 1 and as far as he is concerned, the initial state is a random variable whose p.d.f. is specified by Eq. (7). Thus, his cost functional

$$J^{(2)}(u_0^*) \equiv \min_{u_1 \in R^1} \{E_{x_0}(K^2 (u_0^*)^2 + [x_0 + u_0^* - u_1]^2 \mid \overline{x}_0)\}$$

and consequently his optimal control must satisfy the relationship

$$u_1^* = \overline{x}_0 + u_0^*. \tag{36}$$

This yields the optimal controls/strategies

$$u_0^* = 0, \tag{37}$$

$$u_1^* = \overline{x}_0 \tag{38}$$

and the optimal/minimal expected cost

$$J^* = \sigma_0^2 \tag{39}$$

The analysis from about is summarized in

Theorem 2. *The special case of Witsenhausen's decentralized optimal control problem (1), (2), (4), and (6), where the parameter* $\sigma \to \infty$*, but with the slightly more general initial state information specified by Eq. (7), is considered. Thus, the case is now considered where at time* $k = 1$ *a measurement of the state* x_1 *is not taken by Player 2 and, in addition, Player 1 does not have access to the initial state information* x_0*. The respective optimal strategies of Players 1 and 2 are given by Eqs. (36) and (37) and the minimal cost is given by Eq. (38). The minimal cost is not dependent on the initial state's statistic* $\overline{x}_0$*.*

5 Conclusion

In Witsenhausen's problem statement the following must be made clear. The decision problem is only partially decentralized. At time $k = 0$ the information on the initial state's statistic $\overline{x}_0$ is exchanged among the players. In addition, the synthesis of suboptimal solutions rests on the assumption that before the game starts, during "foreplay," the players are allowed to come together and establish a communication protocol. This entails allowing Player 1 to convey the information on the initial state's statistic to Player 2. Thus, a communications problem using a Gaussian communications channel is modeled. Now, communication can be referred to as signalling, although, in the informational economics literature [9] the term *signalling* assumes a somewhat different meaning. Alternatively, it is tacitly assumed that at decision time $k = 0$, Player 2 takes a measurement of the initial state and communicates his measurement to Player 1. In this case Player 1 knows that Player 2 thinks that the initial state is distributed according to Eqs. (1) or (7). Evidently, the control problem is not completely decentralized and the strategy of Player 1, which naturally incorporates all the information available to him at decision time $k = 0$, has the somewhat unconventional form (9). This state of affairs is masked if, as in Witsenhausen's problem statement, it is assumed that the statistic $\overline{x}_0 = 0$.

The problem with Witsenhausen's problem goes beyond the somewhat hidden requirement that the initial state's statistics information be shared by Players 1 and 2, which immediately detracts from the decentralized aspect of the control problem: the suboptimal solutions are based on the perception that a cooperative communication problem is at hand and this requires the Players to come together and agree on a communications protocol prior to the kickoff of the game. Obviously, the better the communications protocol is, the lower will be the expected cost, and so, when viewing Witsenhausen's problem in the context of a mathematical model of a communications scenario, suboptimal solution methods readily suggest themselves. Now, the fact that the initial state's statistics information is shared is perhaps OK, but the additional requirement that the players come together before the kickoff of the game and agree on a communications protocol, as is indeed the case

in the classical communications paradigm is giving one pause for thought. No such thing would be required if the *optimal* solution of Witsenhausen's problem would be known, in which case the two intelligent Players 1 and 2 could independently figure out their respective modulation and detection strategies. This unfortunately is not the case, because the *optimal* solution of Witsenhausen's problem is not yet fully known and therefore the game is not playable without the artificial preliminary step of setting up a (suboptimal) communications protocol. Knowledge of the optimal solution would obviate the need for this preliminary step. There is one exception: in the special case investigated in Sect. 4 where the variance of the measurement error of Player 2 is very big, the optimal strategies are linear and are known—we refer to the players' optimal strategies (23) and (27); the point is that both players can *independently* derive their optimal strategies.

In summary, Witsenhausen's problem is not fully decentralized to start with and in the absence of an optimal solution the players must establish a somewhat artificial communication protocol before kickoff time. As such, Witsenhausen's problem is somewhat contrived.

Appendix A: Witsenhausen's Counterexample and Detection Theory

An information theoretic analysis of Witsenhausen's binary signalling protocol follows. The case where a system could be in either one of two states is considered.

Specifically, when the binary signalling protocol is invoked in Sects. 3 and 4, the state could be $x_1 = b$ or $x_1 = -b$. Suppose the true state x_1 is not known; however a measurement of the state x_1 is taken, whereupon, based on the measurement's outcome, a declaration concerning the state x_1 is made, namely, x_1 is declared to be b or, alternatively, the state x_1 is declared to be $-b$. Let the performance of the said classifier be quantified by the receiver operating characteristic (ROC) which is a *confusion matrix*

Confusion matrix

True/Rep.	$x_1 = b$	$x_1 = -b$
b	P_{TR}	$1 - P_{\mathrm{FTR}}$
$-b$	$1 - P_{\mathrm{TR}}$	P_{FTR}

Here, the probabilities

$$P_{\mathrm{TR}} \equiv \mathcal{P}(x_1 = b \mid b)$$

and

$$P_{\text{FTR}} \equiv \mathcal{P}(x_1 = -b \mid -b).$$

Concerning the detection task, consider the typical situation where prior information on the state x_1 of the system is available and, based on a received measurement of the state x_1, one is about to decide whether the state $x_1 = b$ or $x_1 = -b$. The performance of the classifier, that is, its ROC, is quantified by the above specified confusion matrix. The following holds.

Theorem. *The state x_1 is known to be either $x_1 = b$ or $x_1 = -b$ and the prior information on x_1 is*

$$\mathcal{P}(x_1 = b) = p.$$

Suppose a measurement is about to be taken whereupon, based on the recorded measurement, a classifier will either declare the state $x_1 = b$ or, alternatively, the state $x_1 = -b$. The performance of the classifier is characterized by its confusion matrix which is parametrized by P_{TR} and P_{FTR}. Then the expected information gain will be

$$\begin{aligned} I(P_{\text{TR}}, P_{\text{FTR}}, p) = {} & pP_{\text{TR}} \log\left(\frac{P_{\text{TR}}}{pP_{\text{TR}} + (1-p)(1-P_{\text{FTR}})}\right) \\ & + p(1-P_{\text{TR}}) \log\left(\frac{1-P_{\text{TR}}}{p(1-P_{\text{TR}}) + (1-p)P_{\text{FTR}}}\right) \\ & + (1-p)(1-P_{\text{FTR}}) \log\left(\frac{1-P_{\text{FTR}}}{pP_{\text{TR}} + (1-p)(1-P_{\text{FTR}})}\right) \\ & + (1-p)P_{\text{FTR}} \log\left(\frac{P_{\text{FTR}}}{p(1-P_{\text{TR}}) + (1-p)P_{\text{FTR}}}\right). \end{aligned} \tag{40}$$

The expected information gain function (40) is depicted in Fig. 9.

Concerning the classification algorithm proper, the probabilities P_{TR} and P_{FTR}, which characterize the classifier's performance, are obtained as follows.

The performance of the sensor used to measure the state x_1 is specified by Eq. (41): if the classifier's threshold is set to t, the detection algorithm yields the estimate $\hat{x}_1$ of the state x_1 according to

$$\hat{x}_1 = \begin{cases} b & \text{if } z_1 \geq t, \\ -b & \text{if } z_1 < t. \end{cases} \tag{41}$$

Hence, from Fig. 10—see, e.g., [8]—one concludes that the detection probability

$$P_{\text{TR}} \equiv \mathcal{P}(\hat{x}_1 = b \mid x_1 = b)$$

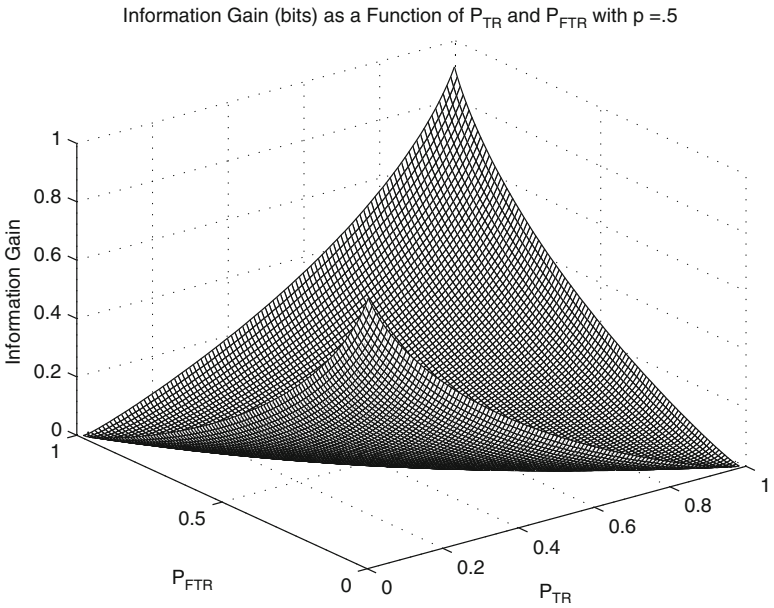

Fig. 9 Expected information gain

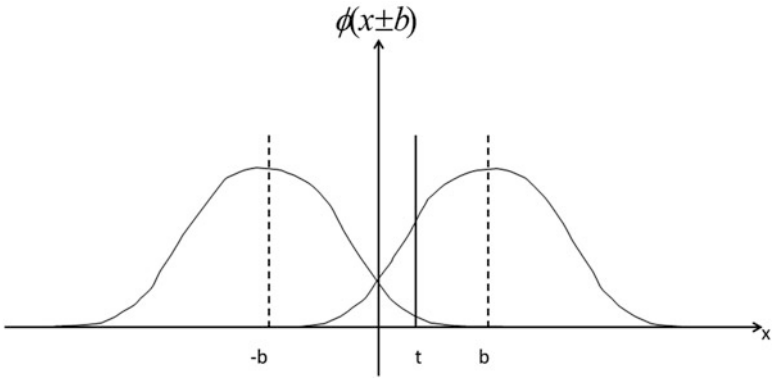

Fig. 10 Calculation of the probabilities of detection and false alarm

$$= \mathcal{P}(z_1 \geq t)$$
$$= 1 - \phi\left(\frac{t-b}{\sigma}\right).$$

Similarly, with reference to Fig. 10, the "False Alarm" probability P_{FA} is calculated as follows:

$$P_{\text{FA}} \equiv \mathcal{P}(\hat{x}_1 = b \mid x_1 = -b)$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sigma} \int_t^{\infty} e^{-\frac{(x-(-b))^2}{2\sigma^2}} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_t^{\infty} e^{-\frac{(x+b)^2}{2\sigma^2}} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{t+b}^{\infty} e^{-\frac{y^2}{2\sigma^2}} dy \\
&= 1 - \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t+b} e^{-\frac{y^2}{2\sigma^2}} dy
\end{aligned}
$$

that is,

$$P_{\text{FA}} = 1 - \phi\left(\frac{t+b}{\sigma}\right).$$

Hence,

$$
\begin{aligned}
P_{\text{FTR}} &= 1 - P_{\text{FA}} \\
&= \phi\left(\frac{t+b}{\sigma}\right).
\end{aligned}
$$

Now,

$$\phi(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right].$$

Using the signal to noise ratio (SNR) definition

$$\text{SNR} \equiv \frac{b}{\sigma}$$

and non-dimensionalizing the threshold

$$t := \frac{t}{\sigma}$$

we finally obtain

$$P_{\text{TR}}(t, \text{SNR}) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{t - \text{SNR}}{\sqrt{2}}\right)\right] \tag{42}$$

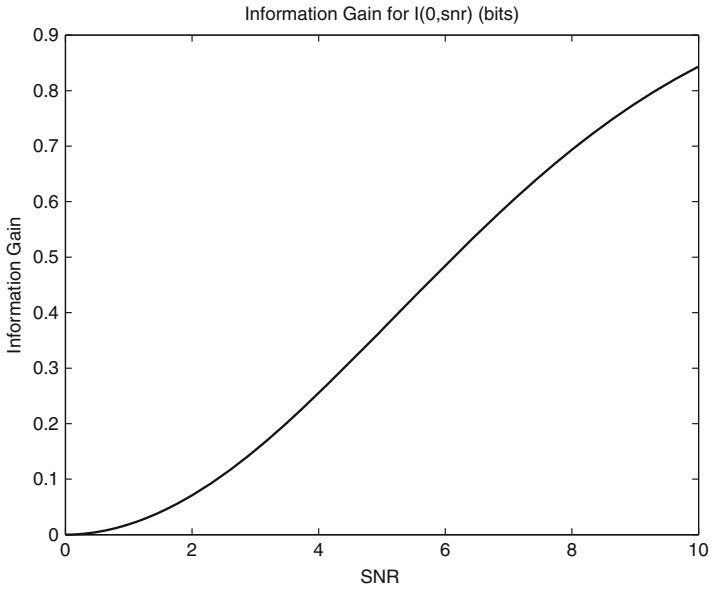

Fig. 11 Expected information gain as a function of SNR

and

$$P_{\mathrm{FTR}}(t, \mathrm{SNR}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{t + \mathrm{SNR}}{\sqrt{2}}\right)\right]. \tag{43}$$

Due to symmetry, the prior probability that the state $x_1 = b$ is $p = \frac{1}{2}$. Inserting the expressions (42) and (43) into Eq. (40) with $p = \frac{1}{2}$ directly yields the expected information gain as a function of the threshold setting t and the SNR. Due to symmetry, the threshold will be set to $t = 0$. The expected information gain as a function of the SNR is shown in Fig. 11. For the optimal b^* calculated in Sect. 3 the attendant SNR yields the expected maximal information gain: it is 1 bit.

References

1. Witsenhausen, H.S.: A counterexample in stochastic optimal control. SIAM J. Control **6**(1), 131–147 (1968)
2. Li, N., Marden, J.R., Shamma, J.S.: Learning approaches to the witsenhausen counterexample from a view of potential games. In: Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, Shanghai, PR China, pp. 157–162, 16–18 Dec 2009
3. Lee, J.T., Lau, E., Ho, Y.-C.: The Witsenhausen counterexample: a hierarchical search approach for non-convex optimization problems. IEEE Trans. Automatic Control **46**(3), 382–397 (2001)

4. McEneaney, W.M., Han, S.H., Liu, A.: An optimization approach to the Witsenhausen counterexample. In: IEEE Conference on Decision and Control, Orlando, FL, Dec 2011
5. Basar, T.: Variations on the theme of the Witsenhausen counterexample. In: Proceedings of the 47th IEEE Conference on Decision and Control Cancun, Mexico, pp. 1614–1619, 9–11 Dec 2008
6. Park, S.Y., Grover, P., Sahai, A.: A constant-factor approximately optimal solution to the Witsenhausen Counterexample. In: Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference, Shanghai, P.R. China, pp. 2882–2886, 16–18 Dec 2009
7. Sahai, A., Grover, P.: Demystifying the Witsenhausen counterexample. IEEE Control Syst. Mag. **30**, 20–24 (2010)
8. Selin, I.: Detection Theory. Academic, New York (1965)
9. Spence, M.: Job market signaling. Q. J. Econ. **87**(3), 355–374 (1973)

Sparse Signal Reconstruction: LASSO and Cardinality Approaches

Nikita Boyko, Gulver Karamemis, Viktor Kuzmenko, and Stan Uryasev

Abstract The paper considers several optimization problem statements for sparse signal reconstruction problems. We tested the performance of AORDA portfolio safeguard (PSG) package with different problem formulations. We solved several medium-size test problems with cardinality functions: (a) minimize L1-error of regression subject to a constraint on cardinality of the solution vector; (b) minimize cardinality of the solution vector subject to a constraint on L1-error of regression. We compared performance of PSG and IBM CPLEX solvers on these problems. Although cardinality formulations are very appealing because of the direct control of the number of nonzero variables, large problems are beyond the reach of the tested commercial solvers. Step-down from the cardinality formulations is the formulation with the constraint on the sum of absolute values of the solution vector. This constraint is a relaxation of the cardinality constraint. Medium and large problems (from SPARCO toolbox for testing sparse reconstruction algorithms) were solved with PSG in the following formulation: minimize L1-error subject to a constraint on the sum of absolute values of the solution vector. The further step-down in the sparse reconstruction problem formulations is the LASSO approach which does not have any constraints on functions. With the LASSO approach you do not know in advance the cardinality of the solution vector and you solve many problems with different regularization parameters. Then you select a solution with appropriate regression error and cardinality. Definitely, it is a time-consuming process, but an advantage of LASSO approach is that optimization problems can be solved quite

N. Boyko (✉) • S. Uryasev
Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
e-mail: nikita@ufl.edu; uryasev@ufl.edu

G. Karamemis
Department of Information Systems and Operations Management, University of Florida, Gainesville, FL 32611, USA
e-mail: gkaramemis@ufl.edu

V. Kuzmenko
V.M. Glushkov Institute of Cybernetics, Kyiv, Ukraine
e-mail: kvnu@mail.ru

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_4

quickly even for very large problems. We have solved with PSG several medium and large problems from the SPARCO toolbox in LASSO formulation (minimize L2-error plus the weighted sum of absolute values of the solution vector).

1 Introduction

Problems considered in this paper are special cases of a broad family of approaches known as compressive sensing. The goal of compressive sensing is to reconstruct a sparse signal from a small number of observations. The theory of compressed sensing was first introduced by [14]. Donoho [15] defines a vector $y = Ax$ in $\mathbb{R}^m$, where they reconstruct vector x given the $m \times n$ matrix A with $m < n \leq \phi m$. Compressed sensing is further discussed in many other papers including [9, 10, 17]. More recent work on compressed sensing is presented and summarized in [24]. Furthermore, many resources including tutorials and papers about compressive sensing can be downloaded from the website [12]. Another resource about compressive sensing, involving mathematical programming formulations and codes, can be found in the website [2]. This website provides "L1-MAGIC" collection of MATLAB routines for solving the convex optimization programs central to compressive sensing. The algorithms are based on interior-point methods and are suitable for large-scale problems.

Let us consider a problem of reconstructing an n-dimensional vector x given the $m \times n$ matrix A and the m-dimensional vector y, which is an observation (possibly noisy) of the product Ax. The most common approach for restoration is to minimize the difference between observation y and estimation Ax:

$$\min_x \frac{1}{2}\|y - Ax\|_2^2. \tag{1}$$

This problem has the analytical solution

$$x = (A'A)^{-1}A'y, \tag{2}$$

if $n \leq m$ and is degenerate if $n > m$. Here $\|v\|_2 = \sqrt{\sum_i v_i^2}$ denotes the Euclidean norm. If the noise is not normally distributed (especially if distribution has fat tails) the estimated vector x can be quite sensitive to tail observations (outliers) of y. In this case, robust statistical approaches can be used. In particular, the L1-error is much less sensitive to outliers than the standardly used L2-error. Therefore the following formulation of the regression problem is a good alternative to the formulation (1):

$$\min_x \|y - Ax\|_1, \tag{3}$$

where L1 norm is defined as $\|v\|_1 = \sum_i |v_i|$. The objective is a convex piece-wise linear function of x and the problem can be reduced to a linear program. This fact is important because of the availability of efficient large-scale linear programming solvers and relevant numerical technologies based on linear programming.

If the modeling of the phenomenon assumes that x vector is sparse (i.e., the vector has many zero components) a regularization is used to "suppress" the irrelevant components of x. The regularization part $\tau\|x\|$ is added to the objective. The problem is formulated as follows:

LASSO-O

$$\min_x \frac{1}{2}\|y - Ax\|_2^2 + \tau\|x\|, \tag{4}$$

where τ sets a trade-off between error and sparsity and is chosen based on empirical considerations. The latter case allows to handle the ill-conditioned formulation, where $n > m$ and x are sparse. Using L1 norm for the regularization part is known as LASSO (least absolute shrinkage and selection operator) technique in the literature [8, 26, 27]. LASSO approach gained popularity because it provides sparse solutions due to properties of L1 norm.

The following two problems can be used instead of problem (4):

LASSO-I

$$\min_x \|y - Ax\|_2, \text{ s.t.} \|x\|_1 \leq t. \tag{5}$$

LASSO-II

$$\min_x \|x\|_1, \text{ s.t.} \|y - Ax\|_2 \leq \epsilon. \tag{6}$$

Here t and ϵ are some predefined parameters. By varying parameter t in problem **LASSO-I** and parameter ϵ in problem **LASSO-II** we can get the same solution vectors x as by variation of parameter τ in problem **LASSO-O**.

The considered minimization problems are convex which makes them computationally attractive. Paper [19] efficiently solves a large-size optimization model (4) with a gradient projection algorithm.

A natural approach to sparse signal reconstructing would be to bound the number of nonzero components (spikes) of vector x. Such a cardinality constraint results in nonconvex formulations. Therefore appropriate numerical tools are needed to solve such problems.

The goal of this paper is to formulate a signal reconstruction problem with cardinality constraints and test the performance of existing commercial solvers on such formulations. In our numerical experiments we generated the data set according to the procedure described in [19].

2 Problem Formulation

First, let us define cardinality function as the number of nonzero components of a vector, i.e.,

$$\mathrm{card}(x) = \sum_{i=1}^{n} I(x_i),$$

where I is an indicator function defined as

$$I(z) = \begin{cases} 1, z \neq 0; \\ 0, \text{otherwise.} \end{cases} \tag{7}$$

Next let us formulate optimization problems similar to **LASSO-I** and **LASSO-II**. We will use the robust L1 norm to quantify the difference between the observed and reconstructed outputs. The first formulation minimizes the L1-error of regression.

Cardinality-I

$$\min_x \|y - Ax\|_1 \tag{8}$$

$$\text{s.t. } \mathrm{card}(x) \leq S, \tag{9}$$

$$L \leq x_i \leq U, \forall i = 1, \ldots, n, \tag{10}$$

$$x \in \mathbb{R}^n, \tag{11}$$

where S is a threshold on "sparsity," i.e., we have an a priori knowledge that no more than S spikes is possible in the initial signal x. L and U are upper and lower bounds on vector x components (in the considered case, $L = -1$ and $U = 1$).

The alternative formulation is as follows:

Cardinality-II

$$\min_x \mathrm{card}(x) \tag{12}$$

$$\text{s.t. } \|y - Ax\|_1 \leq \epsilon, \tag{13}$$

$$L \leq x_i \leq U, \forall i = 1, \ldots, n, \tag{14}$$

$$x \in \mathbb{R}^n. \tag{15}$$

Here we want to determine the signal with the minimal number of spikes that provides us with the output within predefined accuracy ϵ. These two problems, **Cardinality-I** and **Cardinality-II**, can be directly optimized without any additional programming by AORDA portfolio safeguard (PSG) solver. PSG includes the L1-error, L1 norm, and cardinality functions in the list of standard functions.

In order to use integer linear solvers such as IBM CPLEX, we rewrite the problems **Cardinality-I** and **Cardinality-II** as mixed-integer linear programs.

The problem **Cardinality-I**, minimizing the error, can be equivalently rewritten as follows:

MILP I

$$\min_{x,z,t} \sum_{j=1}^{m} z_j \tag{16}$$

$$\text{s.t. } -z_j \le y_j - \sum_{i=1}^{n} a_{ji} x_i \le z_j, \ \forall j = 1, \ldots, m, \tag{17}$$

$$Lt_i \le x_i \le Ut_i, \ \forall i = 1, \ldots, n, \tag{18}$$

$$\sum_{i=1}^{n} t_i \le S, \tag{19}$$

$$z \ge 0, \tag{20}$$

$$x \in \mathbb{R}^n, \ z \in \mathbb{R}^m, \ t \in \{0,1\}^n \ \forall i = 1, \ldots, n. \tag{21}$$

The problem **Cardinality-II**, minimizing the number of spikes in the initial signal, can be equivalently reformulated as follows:

MILP II

$$\min_{x,z,t} \sum_{i=1}^{n} t_i \tag{22}$$

$$\text{s.t. } -z_j \le y_j - \sum_{i=1}^{n} a_{ji} x_i \le z_j, \ \forall j = 1, \ldots, m, \tag{23}$$

$$Lt_i \le x_i \le Ut_i, \ \forall i = 1, \ldots, n, \tag{24}$$

$$\sum_{j=1}^{m} z_j \le \epsilon, \tag{25}$$

$$z \ge 0, \tag{26}$$

$$x \in \mathbb{R}^n, \ z \in \mathbb{R}^m, \ t \in \{0,1\}^n \ \forall i = 1, \ldots, n. \tag{27}$$

We also consider the relaxed version with L1-norm instead of the cardinality function, which is similar to the problem **LASSO-I** :

Relaxed III

$$\min_{x} \|y - Ax\|_1 \tag{28}$$

$$\text{s.t. } \|x\|_1 \le S, \tag{29}$$

$$L \le x_i \le U, \forall i = 1, \ldots, n, \tag{30}$$

$$x \in \mathbb{R}^n. \tag{31}$$

The last problem can be written as a linear program:

LP III

$$\min_{x,z,t} \sum_{j=1}^{m} z_j \tag{32}$$

$$\text{s.t. } -z_j \le y_j - \sum_{i=1}^{n} a_{ji} x_i \le z_j, \ \forall j = 1, \ldots, m, \tag{33}$$

$$L \le x_i \le U, \ \forall i = 1, \ldots, n, \tag{34}$$

$$-t_i \le x_i \le t_i, \ t_i \ge 0 \ \forall i = 1, \ldots, n, \tag{35}$$

$$\sum_{i=1}^{n} t_i \le S, \tag{36}$$

$$z \ge 0, \tag{37}$$

$$x, t \in \mathbb{R}^n, \ z \in \mathbb{R}^m, \ \forall i = 1, \ldots, n. \tag{38}$$

The next section compares computational performance of AORDA [1] nonlinear PSG solver and MILP solver IBM CPLEX [23].

We also consider problem formulation **Relaxed III D** which is equivalent to **Relaxed III**, but all variables have lower bounds equal to 0. This fact is very important for linear programming approach because the sparse solution vector has almost all components equal to zero.

Relaxed III D formulation has the double set of variables. The positive additional variables are used instead of original variables having negative values.

Relaxed III D

$$\min_{x,z} \|y - Ax + Az\|_1 \tag{39}$$

$$\text{s.t. } \sum_{i=1}^{n} (x_i + z_i) \le S, \tag{40}$$

$$0 \le x_i \le U, \ 0 \le z_i \le -L, \ \forall i = 1, \ldots, n, \tag{41}$$

$$x \in \mathbb{R}^n, \ z \in \mathbb{R}^n. \tag{42}$$

The **Relaxed III D** formulation was used to test Car and Tank solvers of PSG on sparse problems provided by [3,4].

3 Computational Experiments with Normally Distributed Data

This section presents numerical experiments for relatively simple test problems with matrices of samples from normal distribution. To test the performance of several optimization solvers based on quite different principals we have generated the sparse reconstruction problem described in [19]. The matrix A, $1{,}024 \times 4{,}096$, is initially filled with independent standard Gaussian samples and then the rows were orthogonalized. Vector x contains 160 randomly placed ± 1 spikes and the observation $y = Ax + d$.

Input data and solutions for **Cardinality-I**, **Cardinality-II**, and **Relaxed-III** problems are available in PSG format. We also provide **MILP-I** and **MILP-II** problems in CPLEX LP format. These PSG and CPLEX files can be downloaded from www.ise.ufl.edu/uryasev/testproblems/case_studies/CS_Sparse_Reconstruction/CS_Sparse_Reconstruction.

We report performance of AORDA PSG 64-bit version (PSG 64-bit version runs about 50 % faster than PSG 32-bit version) and IBM CPLEX 32-bit version. Computations were conducted on PC with 2.66 MHz processor.

We have considered two cases. In the first case, we assumed the absence of noise (i.e., $d = 0$). The calculation results are presented in Table 1. For Problem Type 1 and Problem Type 3 the PSG and CPLEX solvers have comparable performances (PSG solves these problems faster than CPLEX, but we used 64-bit version of PSG and 32-bit version of CPLEX).

In the second case we considered Gaussian noise $d \sim N(0, 10^{-4})$. The result shown in Table 2 demonstrates that presence of noise significantly complicates computations for both PSG and CPLEX.

Tables 1 and 2 show that **Relaxed III** problem is solved quite fast by PSG in no-noise and in noisy cases (solving time is below 13 s). Therefore, Problem **Relaxed III** is a good alternative for Problem **LASSO-O**, which is standardly considered in the literature. CPLEX solver showed good performance (solving time about 19 s) for the equivalent problem **LP III** in no-noise case (see Table 1). In the noisy case it takes CPLEX quite long to solve (2,129 s) (see Table 2). However, a faster

Table 1 Noise $d = 0$

CPLEX			PSG		
Problem	Time	Objective	Problem	Time	Objective
MILP I	76.64	0	**Cardinality I**	42.5	3.067e−13
MILP II	a	44	**Cardinality II**	3,234.24	160
LP III	18.74	0	**Relaxed III**	12.38	1.063e−12

Computation time in seconds for CPLEX and PSG (solver TANK). Problem Types: I = error minimization, II = cardinality minimization, and III = relaxed version of I

[a]CPLEX did not find the exact solution within 24 h

Table 2 Noise $d \sim N(0, 10^{-4})$

CPLEX			PSG		
Problem	Time	Objective	Problem	Time	Objective
MILP I	a	7.91e−5	**Cardinality I**	34.09	1.311e−4
MILP II	10,793.9	160	**Cardinality II**	1,925.96	160
LP III	2,129.48	3.848e−5	**Relaxed III**	10.66	1.315e−4

Computation time in seconds for CPLEX and PSG (solver TANK). Problem Types: I = error minimization, II = cardinality minimization, and III = relaxed version of I

[a]CPLEX crashed because of out-of-memory status

performance of CPLEX could be achieved if we stop it when a lower precision is achieved; CPLEX minimized regression error with precision 3.848e−5, but PSG reported precision error 1.315e−4 (see Table 1).

If the number of spikes or upper bound on the number of spikes is known, then Problem **Cardinality-I** is a good alternative to Problems **LASSO-O** and **Relaxed III**. We knew (by construction of the problem) that number of spikes equals 160. Therefore, the constraint on the number of spikes was set to 160. PSG solves Problem **Cardinality-I** quite fast (solving time is below 43 s) (see Tables 1 and 2). Moreover, the solution can be immediately used "as it is" without eliminating small nonzero terms, as it is necessary in Problems **LASSO-O** and **Relaxed III**. The reason for this is that Problem **Cardinality-I** can be considered as the "correct" formulation of the original sparse reconstruction problem, compared to the approximate regularized Problems **LASSO-O** and **Relaxed III** which are computationally efficient "surrogates" of the original problem. CPLEX demonstrated good performance for the equivalent formulation **MILP I** in no-noise case (solving time 77 s) (see Table 1); however CPLEX crashed in noisy case. The crash can probably be prevented by solving the problem with a 64-bit version of CPLEX or on some UNIX machine.

If the precision of solution of the regression problem is specified in advance, then Problem **Cardinality-II** and the equivalent Problem **MILP II** can be considered. PSG solving times for Problem **Cardinality-II** are quite large (3,234 s in no-noise case and 1,926 in noisy case) (see Tables 1 and 2). CPLEX performed poorly for the Problem **MILP II**; it did not find a solution during 24 h for the no-noise case (see Table 1), and it has found solution during 1,079 s for the noisy case (see Table 2).

4 Computational Experiments with SPARCO Problems

This section presents performance of AORDA PSG solvers (64 bit) for a set of real-life problems from the SPARCO website [3, 4]. Computations were conducted on PC with 2.83 MHz processor.

Compared to the previous section, mostly, these are real-life problems in imaging, compressed sensing, geophysics, information compressing, etc. References to the problem sources are included in the calculation results tables.

To access initial data we used software from [3, 4]. This site provides also a set of operators to deal with data. For the relatively small problems we converted all data to PSG format. The problems in PSG can be solved in PSG MATLAB or PSG RunFile environments. Large problems were solved with the External Function tool of PSG in MATLAB environment to avoid generating full matrix for the regressions problem (and to save computational time and memory).

We think that the cardinality problem formulations, such as **Cardinality-I** or **Cardinality-II**, are much better formulations than **LASSO**-type formulations which do not directly control the cardinality of the decision vector. With the non-**Cardinality** formulations we need to solve the problem many times and adjust parameters of the problem until we find the solution with acceptable cardinality. Although the **Cardinality**-type formulations are preferable, it may be quite difficult to solve them for very large dimensions.

This section solves the problems with the **Relaxed III D** and **LASSO-O** formulations. Results of solving problems in **Relaxed III D** formulation with different values of upper bound S are shown in the Table 3. For some instances we used linearization of objective (39) to obtain fully linearized form of **Relaxed III D**. The table shows that the problems with several thousands of variables and several thousands of scenarios can be solved quite fast when the parameter S bounding the total absolute value of nonzero variables is small. However, for large values of S, the solving time may be quite large. The data for the problems and codes, solutions in PSG format, and references to the original sources of information are placed in this website www.ise.ufl.edu/uryasev/testproblems/case_studies/CS_Sparse_Reconstruction_SPARCO/CS_Sparse_Reconstruction_SPARCO.

For the really large problems, we have found that **LASSO-O** is the most preferable problem formulation, at least with the solvers included in PSG. Table 4 shows the calculation results for the large problems using the External Function tool of PSG in MATLAB environment. This table presents the runs with different value of the regularization coefficient in objective. MATLAB *.m files to prepare data, convert them to PSG format, and solve problems using either fully converting to PSG format or External Function tool of PSG can be downloaded from www.ise.ufl.edu/uryasev/testproblems/case_studies/CS_Sparse_Reconstruction_SPARCO_matlab/CS_Sparse_Reconstruction_SPARCO_matlab.

5 Conclusion

We have proposed an approach for solving signal sparse reconstruction problems using nonconvex formulations with cardinality functions. In the first group of experiments, we have used test problems with matrices of samples from normal distribution. In order to test the performance of several optimization solvers,

Table 3 Calculation results for medium-size SPARCO problems in **Relaxed III D** form

Problem	S	Objective	Card	Card	Max	Time	Solver	References
$m \times n$			$\lvert x_i \rvert \geq 1$	$\lvert x_i \rvert \geq 1e{-}3$	$\lvert x_i \rvert$	s		
Problem 2	100	1.22E+00	9	9	29.4	0.5	C	[5, 11, 16]
1,024 × 1,024	200	6.63E−01	16	16	29.4	0.6	C	
	400	4.42E−02	42	44	29.4	4.5	C	
	500	1.24E−14	67	71	43.1	2.8	C	
Problem 3	100	6.38E−01	2	2	69.0	0.5	T	[3, 4]
1,024 × 2,048	140	8.03E−02	6	6	90.5	4.1	T	
	220	2.27E−03	35	119	91.5	130.3	T	
	240	4.58E−14	34	1,004	90.5	460.3	Cfl	
Problem 5	100	1.00E+00	3	3	56.4	0.5	C	[3, 4]
300 × 2,048	140	2.71E−01	4	39	66.5	8.0	T	
	170	5.79E−02	15	127	67.6	41.1	Cfl	
	200	1.91E−14	15	299	68.1	3.6	Cfl	
Problem 6	170	1.41E+02	14	14	35.1	1.4	T	[6]
600 × 2,048	800	6.62E+01	64	85	57.5	56.5	T	
	1,700	1.59E+00	115	423	76.6	185.9	Cfl	
	2,000	3.77E−12	285	601	78.2	30.9	Cfl	
Problem 7	3	5.81E−02	0	13	0.65	2.3	C	[7]
600 × 2,560	10	3.39E−02	0	20	0.79	7.9	C	
	17	1.02E−02	0	20	0.94	12.8	C	
	20	1.37E−13	9	20	1.00	1.6	C	
Problem 8	5	4.94E−02	0	19	0.63	6.0	T	[7]
600 × 2,560	15	1.64E−02	0	20	0.87	11.9	T	
	19	3.29E−03	0	20	0.97	11.6	T	
	20	1.36E−13	12	20	1.00	1.2	T	
Problem 9	15	3.14E−01	5	7	4.3	0.01	C	[11, 16]
128 × 128	30	9.30E−02	9	10	5.0	0.02	C	
	40	7.81E−03	12	12	5.0	0.02	C	
	45	2.92E−13	12	12	5.0	0.04	C	
Problem 10	600	1.32E−01	10	10	121.5	0.5	C	[11, 16]
1,024 × 1,024	800	6.10E−02	11	11	121.5	0.6	C	
	950	1.15E−02	12	12	131.7	0.7	C	
	1,000	1.64E−12	14	14	149.2	143.1	C	
Problem 11	15	1.22E+00	8	28	1.94	3.1	C	[3, 4]
256 × 1,024	20	5.18E−01	10	45	2.10	6.5	T	
	23	1.29E−01	11	58	2.17	35.5	Cfl	
	25	2.53E−14	11	221	2.18	31.6	Cfl	
Problem 601	100	6.48E+00	15	26	23.2	81.4	C	[25]
3,200 × 4,096	200	2.78E+00	46	173	23.2	1,965.1	V	
	300	8.42E−01	63	836	23.2	2,894.0	V	
	400	9.18E−08	7	3,478	23.2	509.2	V	

(continued)

Table 3 (continued)

Problem	S	Objective	Card	Card	Max	Time	Solver	References
$m \times n$			$\lvert x_i \rvert \geq 1$	$\lvert x_i \rvert \geq$1e−3	$\lvert x_i \rvert$	s		
Problem 602	200	6.80E+00	40	209	15.9	1,874.2	V	[25]
3,200 × 4,096	400	2.99E+00	85	728	16.1	2,906.6	V	
	600	3.60E−01	130	1,060	16.3	2,982.0	V	
	700	9.71E−05	142	3,964	16.2	272.6	V	
Problem 603	50	3.35E−01	4	4	24.4	2.2	C	[19]
1,024 × 4,096	100	1.75E−01	12	16	30.9	10.7	C	
	200	4.16E−02	30	294	31.0	1,150.4	Cfl	
	300	2.32E−14	39	1,019	32.4	391.0	Cfl	
Problem 902	0.2	2.32E−02	0	1	0.20	0.07	C	[20–22]
200 × 1,000	0.5	1.86E−02	0	3	0.40	0.12	C	
	1.5	3.60E−03	0	3	0.78	0.12	C	
	2.0	3.12E−14	0	5	0.87	26.30	C	
Problem 903	5	6.17E−01	3	6	1.18	0.6	C	[18]
1,024 × 1,024	7	4.06E−01	3	11	1.47	1.4	C	
	12	9.92E−02	6	48	2.19	29.7	T	
	13	9.18E−05	6	17	2.19	10.7	C	

S = upper bound; Card $\lvert x_i \rvert \geq 1$ = number of variables with absolute value greater than 1; Card $\lvert x_i \rvert \geq$1e−3 = number of variables with absolute value greater than 1e−3; Max $\lvert x_i \rvert$ = maximum value of $\lvert x_i \rvert$; Solver: C = Car PSG solver, T = Tank PSG solver, V = Van PSG solver; Cfl = Car PSG solver with full linearization; Tfl = Tank PSG solver with full linearization

the Sparse Reconstruction Problem was generated which was described in [19]. Using different principals, the problems have been solved with AORDA PSG optimization package. Correspondent MILP formulations have been used to solve equivalent problems with IBM CPLEX optimization package and compare solvers' performance and results were obtained. Computational experiments show that the optimization package specialized for dealing with cardinality functions (PSG) has better performance than general MILP package on the considered test problems.

Second group of experiments were conducted using real-life medium and large-scale problems that were downloaded from the SPARCO website [3, 4]. The problems have been solved with PSG optimization package. For the really large problems, the most preferable formulation was founded to be **LASSO-O**. Based on the computational experiments, we can conclude that PSG is a reasonable alternative to the specially developed algorithms for the sparse reconstruction problems, considered, for instance in [19]. The main advantages of using PSG versus other softwares are the simplicity of the PSG codes which have only several lines for even large-scale problems and the transparency of formulations where pre-coded mathematical functions are used to solve the problems.

Table 4 Calculation results for medium- and large-size SPARCO problems in **LASSO-O** form using external function tool of PSG

Problem	τ	Objective	Card	Card	Max	Time	References
$m \times n$			$\lvert x_i \rvert \geq 1$	$\lvert x_i \rvert \geq 1e{-}3$	$\lvert x_i \rvert$	s	
Problem 3	10	1.308E+03	2	2	80.2	0.02	[3,4]
1,024 × 2,048	1	1.780E+02	5	38	89.3	0.05	
	0.1	2.164E+01	29	111	90.3	0.18	
	0.01	2.217E+00	35	121	90.5	2.57	
Problem 5	10	1.226E+03	3	3	57.2	0.04	[3,4]
300 × 2,048	1	1.580E+02	4	33	66.8	0.17	
	0.1	1.778E+01	18	109	67.8	1.67	
	0.01	1.810E+00	23	156	67.9	19.24	
Problem 6	10,000	9.652E+06	35	37	57.1	0.8	[6]
600 × 2,048	1,000	1.586E+06	102	175	74.3	3.6	
	100	1.722E+05	118	405	77.0	36.8	
	10	1.745E+04	121	555	77.6	385.6	
Problem 7	0.2	2.252E+00	0	14	0.64	0.07	[7]
600 × 2,560	0.1	1.562E+00	0	20	0.82	0.10	
	0.05	8.905E−01	0	20	0.91	0.11	
	0.02	3.825E−01	0	20	0.96	0.15	
Problem 8	0.2	4.381E+00	0	13	0.37	0.09	[7]
600 × 2,560	0.1	3.799E+00	0	20	0.68	0.11	
	0.05	3.156E+00	0	20	0.84	0.13	
	0.01	2.470E+00	0	20	0.97	0.23	
Problem 401	2	2.431E+02	1	9	1.45	1.2	[13]
29,166 × 57,344	1	2.273E+02	6	121	2.43	3.1	
	0.5	1.849E+02	25	497	2.93	8.2	
	0.2	1.152E+02	37	1,713	3.40	23.5	
Problem 402	2	2.643E+02	1	10	1.45	1.8	[13]
29,166 × 86,016	1	2.473E+02	6	137	2.43	4.4	
	0.5	2.007E+02	26	545	2.93	13.9	
	0.2	1.248E+02	37	1,898	3.40	32.1	
Problem 403	10	1.124E+04	57	65	10.6	2.7	[3,4]
196,608 × 196,608	3.3	6.174E+03	119	122	20.2	3.0	
	0.33	1.136E+03	277	970	25.5	10.0	
	0.1	4.802E+02	449	5,591	25.8	35.2	
Problem 601	10,000	8.474E+05	4	4	22.7	420.7	[25]
3,200 × 4,096	1,000	1.948E+05	26	60	23.1	598.0	
	500	1.187E+05	38	124	23.1	808.1	
	200	5.741E+04	54	288	23.1	1,460.1	
Problem 602	1,000	3.165E+05	37	132	16.1	543.6	[25]
3,200 × 4,096	500	2.086E+05	68	335	16.2	848.9	
	200	1.049E+05	95	717	16.2	1,117.9	
	100	5.745E+04	110	859	16.2	2,038.8	

(continued)

Table 4 (continued)

Problem	τ	Objective	Card	Card	Max	Time	References
$m \times n$			$\lvert x_i \rvert \geq 1$	$\lvert x_i \rvert \geq$1e−3	$\lvert x_i \rvert$	s	
Problem 603	2	1.893E+02	5	5	25.6	0.20	[19]
1,024 × 4,096	1	1.210E+02	7	9	29.3	0.20	
	0.1	2.145E+01	28	169	32.0	0.79	
	0.01	2.544E+00	38	605	32.2	6.64	
Problem 701	10	6.911E+03	48	49	13.5	0.59	[19]
65,536 × 65,536	5	4.336E+03	55	55	18.6	0.67	
	2	2.110E+03	68	87	21.5	0.75	
	1	1.198E+03	107	144	22.3	0.88	
Problem 702	0.05	2.484E+00	0	4	0.23	0.30	[19]
16,384 × 16,384	0.04	2.470E+00	0	26	0.47	0.55	
	0.02	2.087E+00	0	169	0.76	1.30	
	0.01	1.332E+00	0	194	0.88	1.88	

τ = regularization coefficient in objective; Card $\lvert x_i \rvert \geq 1$ = number of variables with absolute value greater than 1; Card $\lvert x_i \rvert \geq$1e−3 = number of variables with absolute value greater than 1e−3; Max $\lvert x_i \rvert$ = maximum value of $\lvert x_i \rvert$

References

1. American Optimal Decisions (AORDA), Portfolio Safeguard (PSG) (2009). http://www.aorda.com
2. L1-magic (2009). http://www.acm.caltech.edu/l1magic/
3. Berg, E.v., Friedlander, M.P.: SPARCO: a toolbox for testing sparse reconstruction algorithms (2008). http://www.cs.ubc.ca/labs/scl/sparco/
4. Berg, E.v., Friedlander, M.P., Hennenfent, G., Herrmann, F., Saab, R., Yılmaz, Ö.: Sparco: a testing framework for sparse reconstruction. Technical Report TR-2007-20, Department of Computer Science, University of British Columbia, Vancouver (2007)
5. Buckheit, J., Donoho, D.L.: Wavelab and reproducible research. In: Wavelets and Statistics. Springer, Berlin (1995). http://citeseer.ist.psu.edu/article/buckheit95wavelab.html
6. Candés, E.J., Romberg, J.: Practical signal recovery from random projections. In: Proceedings of SPIE Conference Wavelet Applications in Signal and Image Processing XI, vol. 5914, San Diego (2004)
7. Candès, E., Romberg, J.: L1-magic (2007). http://www.l1-magic.org/
8. Candes, E., Tao, T.: The dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313 (2007). doi:10.1214/009053606000001523
9. Candes, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
10. Candes, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
11. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998). http://epubs.siam.org/SISC/volume-20/art30401.html
12. Compressed Sensing Resources (2009). http://www.dsp.ece.rice.edu/cs/
13. Database of Creative Commons licensed sounds (2007). http://freesound.iua.upf.edu/
14. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
15. Donoho, D.L.: For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. Commun. Pure Appl. Math. **59**(7), 907–934 (2006)

16. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994). http://citeseer.ist.psu.edu/donoho93ideal.html
17. Donoho, D.L., Tsaig, Y.: Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. Technical Report, Institute for Computational and Mathematical Engineering, Stanford University (2006). http://www-stat.stanford.edu/~donoho/
18. Dossal, C., Mallat, S.: Sparse spike deconvolution with minimum scale. In: Proceedings of Signal Processing with Adaptive Sparse Structured Representations, pp. 123–126, Rennes, France (2005). http://spars05.irisa.fr/ACTES/PS2-11.pdf
19. Figueiredo, M., Nowak, R., Wright, S.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Topics Signal Process. **1**(4), 586–597 (2007). DOI 10.1109/JSTSP.2007.910281. http://www.lx.it.pt/~mtf/GPSR
20. Hennenfent, G., Herrmann, F.J.: Sparseness-constrained data continuation with frames: applications to missing traces and aliased signals in 2/3-D. In: SEG International Exposition and 75th Annual Meeting. Tulsa, OK, USA (2005). http://slim.eos.ubc.ca/Publications/Public/Conferences/SEG/hennenfent05seg.pdf
21. Hennenfent, G., Herrmann, F.J.: Simply denoise: wavefield reconstruction via coarse nonuniform sampling. Technical Report, UBC Earth & Ocean Sciences (2007)
22. Herrmann, F.J., Hennenfent, G.: Non-parametric seismic data recovery with curvelet frames. Technical Report, UBC Earth & Ocean Sciences Department (2007). http://slim.eos.ubc.ca/Publications/Public/Journals/CRSI.pdf. [TR-2007-1]
23. IBM CPLEX 11.2 (2009). http://www.ilog.com/products/cplex/
24. Eldar, Y.C., Kutyniok, G.: Compressed sensing: Theory and applications. Cambridge University Press (2012)
25. Takhar, D., Laska, J.N., Wakin, M., Duarte, M., Baron, D., Sarvotham, S., Kelly, K.K., Baraniuk, R.G.: A new camera architecture based on optical-domain compression. In: Proceedings of the IS&T/SPIE Symposium on Electronic Imaging: Computational Imaging, vol. 6065, Springfield, VA, USA and Bellingham, WA, USA (2006)
26. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1994)
27. Wang, L., Gordon, M.D., Zhu, J.: Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In: ICDM '06: Proceedings of the 6th International Conference on Data Mining, pp. 690–700, IEEE Computer Society, Washington, DC, USA (2006). DOI http://dx.doi.org/10.1109/ICDM.2006.134

Evaluation of the Copycat Model for Predicting Complex Network Growth

Tiago Alves Schieber, Laura C. Carpi, and Martín Gómez Ravetti

Abstract We deal here with the issue of complex network evolution. In particular, we propose the use of the Copycat Model as a framework to predict the dynamic behavior of networks. This model has the ability to dynamically adjust the topological properties step by step during the network's growth. We test the methodology with three networks, an artificial net called popularity vs. similarity and two real ones Manufacturing Emails Network and Slashdot threads Network. The results show that the methodology is able to correctly predict network's evolution reproducing several network properties.

1 Introduction

The analysis of networks has received, in the past few years, a central role within the research interests. The increasing availability of network data resources and the possibility of representing many natural and artificial systems make them really attractive. The real fact is that we live in a world of networks, and what initially belonged to the field of physics nowadays belongs to a new interdisciplinary field. This explosion of interest in networks can be considered a social and cultural phenomenon by itself [1].

Real networks display nontrivial structures, deviating from traditional assumptions of network theory. Real networks were supposed to posses links homogeneously distributed and were typically modeled by random graphs [2]. However, it was later demonstrated that the number of connections of the majority of nodes is very small, existing only a reduced number of nodes highly connected. These

T.A. Schieber (✉) • M.G. Ravetti
Departamento de Engenharia de Produção, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
e-mail: tiagoschieber@eng-pro.dout.ufmg.br; martin.ravetti@dep.ufmg.br

L.C. Carpi
LaCCAN/CPMAT - Instituto de Computação, Universidade Federal de Alagoas, Maceió, Brazil
e-mail: lauracarpi@gmail.com

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_5

sparsely connected networks posses a node degree distribution that follows a power-law form, presenting, in general, large clustering coefficient and small average path length values. These properties are the ones reflecting the mechanisms driving their growth. Several attempts were trying to reproduce the evolution of real networks preserving their properties. However, none of them was able to dynamically keep their individual behavior during the whole growing process.

The first attempt to understand the way networks with scale-free features grow was done by Barabási and Albert in 1999 [3]. They observed that the scale-free characteristics present in real networks are originated by a mechanism known as preferential attachment. It is based on the fact that new nodes in the network tend to connect to those that are already more connected. The Barabási and Albert model (BA) is able to generate networks with power-law distributions; however, the obtained clustering coefficient is usually lower than the ones present in real networks. They presented a model that, at each step, a new node is included and it is connected to $m > 0$ other nodes with probability:

$$p_i = \frac{k_i}{\sum k_i}, \tag{1}$$

where k_i is the degree of the node i and the summation is over all nodes of the network. It can be proved that this network possesses an average degree equals $2m$ and a degree distribution following $P(k) \sim k^{-3}$ [4,5].

In 2000, Dorogotsev and Mendes [6] proposed a modification in the rules governing the preferential attachment mechanism. They have found that in some systems, the probability of a connection is not only proportional to the node degree but also to its age. Since then, several changes in preferential attachment have been performed. In 2001, Barabasi et al. used direct measurements on the available data of the social network of scientific collaborations [7] and constructed a model that allows the investigation of the large scale topology of the network and its dynamical features. After that, increasingly sophisticated models have been created (see [8] for a deeper discussion on the topic). In 2007, Goshal and Newman proposed a model that grows a network with any desirable degree distribution [9]. This model has some drawbacks because two networks with the same degree distribution are not necessarily isomorphic [10]. Also in 2007, Leskovec et al. created a model, called Forest Fire model, based on the observation that densification power laws and shrinking effective diameters are properties that hold across a range of diverse networks [11]. In 2008, Leskovec et al. also created a model based on the generation of triangles in the network (triangle closure) using the maximum-likelihood principle [12]. In 2010, Barthelemy proposed the tree growth model with local optimization [13] where spatial networks were considered. In 2011, Herrera and Zufiria proposed a model that creates networks with adjustable clustering coefficient via random walks [14]. In 2012 [15, 16], the discussion about the processes underlying the preferential attachment is revisited and a new model based on the popularity versus similarity was presented by Papadopoulos et al. [15]. They developed a framework where new connections, instead of preferring popular nodes,

optimize certain trade-offs between popularity and similarity [15]. This model finds an interesting geometric interpretation: each node of the network is represented by a single point in the plane and identified by its polar coordinates (r, θ). Two nodes are said *similar* if they possess the same θ. Thus we can define the angular distance between two nodes, θ_{xy} as the *similarity distance* between the nodes x and y. The *popularity* is viewed as the radial position.

In 2013, Schieber and Ravetti proposed the Copycat Model (CP) that reproduces network growths maintaining their main property values [17]. By using the mean degree, the global clustering coefficient, the transitive clustering coefficient, and the distance to the uniform distribution, the CP model is able to mimic real networks. These properties dynamically adjusted each node inclusion by solving an optimization problem. This automatic self-adjustment is a good framework for prediction purposes. In this chapter we use the CP model to predict the evolution of network behavior. Two real networks and one artificial network are tested and analyzed.

The remainder of this chapter is organized as follows: in Sect. 1 notation and complex network measures are introduced. The Copycat Model is explained in Sect. 2. In Sect. 3 the computational experiments are described and the databases used to test our methodology are introduced. Finally Sect. 4 concludes the chapter.

2 Preliminaries and Notation

In this section we introduce basic notions of complex networks and define some properties that will be used later on.

A complex network is a structure that can be represented as a graph with nontrivial topological features. Most natural and artificial networks have a specific architecture based on a fat-tailed distribution of the number of connections, high global clustering coefficient (C), and small average path length (l). They are not static, but evolving systems far from equilibrium.

Formally, a network G is a pair $(\mathcal{N}, \mathbb{E})$, where $\mathcal{N}$ is a set of points, which we call nodes or vertices, and $\mathbb{E}$ is a set of pairs of distinct nodes, which we call edges. We say that a network is undirected if $\mathbb{E}$ is formed by nonordered pairs of nodes and directed otherwise.

In order to simplify, we consider only finite and undirected networks. The size of the network is set as N. An edge $e \in \mathbb{E}$ is denoted by $\langle x, y \rangle$, where x and y are the endpoints. The *degree* of a vertex x, represented by $k(x)$, is the number of edges that have x as an endpoint. As not all nodes have the same number of edges, the spread in the number of edges a node has, is characterized by a distribution function $P(k)$, which gives the probability that a randomly selected node has exactly k edges.

A *path* is a finite set of edges $\{e_1, e_2, \ldots, e_n\}, e_i = \langle x_i, x_{i+1} \rangle$ such that all nodes $x_1, x_2, \ldots, x_{n+1}$ are distinct from one another. The size of the path is the number of edges presented in the path. We say that two nodes, z and w, are connected if there is a path $\{e_1, e_2, \ldots, e_n\}$ such that $x_1 = z$ and $x_{n+1} = w$. The *distance* between the

nodes z and w, represented by $d(z, w)$, is the smallest size of all paths between them. The *average path length* (l) is the average of all possible distances in the network and the *diameter* of a network is the maximum distance of any two vertices in the network.

The *closeness centrality* measure of a node $x \in \mathcal{N}$ is given by

$$\mathrm{cc}(x) = \frac{1}{\sum_{y \in \mathcal{N} - \{x\}} d(x, y)}.$$

It can be viewed as the efficiency of each vertex in spreading information to all other vertices. The larger the closeness centrality of a vertex, the shorter the average distance from the vertex to any other, and thus the better positioned the vertex is on the network for spreading information to other vertices.

The *betweenness centrality* measure of a node x quantifies its importance in terms of interactions of nodes in the network. It is defined by

$$\mathrm{btwc}(x) = \sum_{y,z \in \mathcal{N} - \{x\},\ y \neq z} \frac{\sigma(y, x, z)}{\sigma(y, z)},$$

where $\sigma(y, x, z)$ is the number of shortest paths between y and z passing through x and $\sigma(y, z)$ is the number of shortest paths between y and z.

Let V be a normalized eigenvector with respect to the greater eigenvalue λ of the adjacency matrix A; we define the *eigenvector centrality* of a node x simply by

$$\mathrm{evc}(x) = V(x).$$

The geometrical idea behind the definition of eigenvector centrality is that a node is central to the extent that the node is connected to others that are central.

There are two other interesting measures related to community in a network: the clustering coefficient, C, and the transitive clustering coefficient, C^T. The concept community is inspired by the idea that all people belonging to a certain group know each other. In graphs we denote G_x as subgraph of G formed by the first neighbors of x and $|E_x|$ the number of edges in G_x. A measurement of community structure of a node x is defined as the fraction between $|E_x|$ and the total number of possible edges of G_x, $(k(x) \cdot (k(x) - 1)/2)$. This measure is called *local clustering coefficient* of node x, and it can be mathematically represented as

$$C(x) = \begin{cases} \dfrac{2|E_x|}{k(x) \cdot (k(x) - 1)} & \text{if } k(x) > 1 \\ 0 & \text{otherwise} \end{cases}.$$

The *clustering coefficient*, C, is the average of all clustering coefficient on the network:

$$C = \frac{\sum_{x \in \mathcal{N}} C(x)}{N}. \tag{2}$$

$C(x)$ could also be written as the fraction between three times the number of triangles involving node x and the number of connected triple having x as a central node. The *global clustering coefficient* also known as the *transitive clustering coefficient* of the network is defined as the fraction between the number of triangles present in the network, $\#_\Delta$, and the number of connected triples, $\#_3$:

$$C^T = \frac{3 \cdot \#_\Delta}{\#_3}.$$

Readers can refer to [18] for a deeper discussion on the topic.

3 Copycat Model

The Copycat Model (CP) [17] is a methodology to dynamically mimic complex network's growth. It uses the square root of Jensen–Shannon divergence ($\mathcal{J}^{1/2}$), a metric between two probability distributions [19,20], to map the network's evolution by the pair $(N, \mathcal{J}^{1/2}(P, P_U))$, where P is the degree distribution and P_U is the uniform distribution. Although different networks could exhibit same values of the Jensen–Shannon divergence, the CP model provides a remarkable help when it is used in combination with other metrics, such as the average degree and the clustering coefficient.

By giving the $g(x)$, $c(x)$, $c^T(x)$, and $\hat{\mathcal{J}}(x)$ functions that respectively represent the mean degree, global clustering coefficient, transitive clustering coefficient, and distance to the uniform distribution of a network with size x, the algorithm creates a new random network that preserves the values of these properties.

As input data, the algorithm also receives the initial network G_0, and the final number of nodes N which is also the highest stage the network has in its evolution, considering the removal of nodes not allowed.

At each iteration one node is added and, by using a simple computation procedure, it is decided how many links this new node will gain to maintain the mean average degree. To choose the nodes to which the new one is connected, the difference between the distance $\mathcal{J}^{1/2}(P, P_U)$ (new network to reference) and the distance $\hat{\mathcal{J}}(x)$ (copied network to reference) is minimized. The model chooses the node from a restricted candidate list (RCL). This procedure is similar to a classic heuristic procedure to solve combinatorial optimization problem called GRASP (greedy randomized adaptive search procedure) [21]. As the algorithm randomly

chooses a node to create the link, by controlling the random number generator seed, it is possible to create as many different topologies which preserve the values of main properties, as desired (see Algorithm 1). A detailed explanation and earlier tests of the CP model can be found in [17].

Data: G^0, N, $g(x)$, $c(x)$, $c^T(x)$ $\hat{\mathcal{J}}(x)$
Result: Network topology with N nodes and the desired metric values
for $i = |G^0|$ *to* N **do**
  Compute the mean degree of the network, (G_i);
  Add a new node;
  $m := max\left\{1, \left\lfloor \frac{(i+1)\cdot g(i+1) - i \cdot G_i}{2} \right\rfloor\right\}$;
  while $m > 0$ **do**
    Create a list of candidates (RLC) with of nodes v_j, such that if a connection between v_{i+1} and v_j is performed, the difference between the distances of their degree distributions to the reference is minimized;
    Randomly choose a node from RCL;
    m=m-1;
    Compute the clusterings coefficients of the resulting network (C_{i+1} and C^T_{i+1});
    Define $\delta = |C_{i+1} - c(i+1)|$ and $\delta^T = |C^T_{i+1} - c^T(i+1)|$;
    if $\delta \geq \delta^T$ **then**
      $C^* = C$ and $c^* = c$;
    end
    if $\delta < \delta^T$ **then**
      $C^* = C^T$ and $c^* = c^T$;
    end
    if $m > 0$ *and* $C^*_{i+1} - c^*(i+1) < 0$ **then**
      Create a list of candidates (RLC) with of nodes v_j such that the distance between v_{i+1} and v_j equals 2, and the clustering coefficient of the resulting network, after connection v_{i+1} and v_j is performed, possess its maximum value;
      Randomly choose a node from RCL;
      m=m-1;
    end
  end
end

Algorithm 1: Pseudo-code of the Copycat Model. The algorithm receives as input parameters, an initial graph, the number of nodes, the mean degree ($g(x)$), the average clustering coefficient ($c(x)$), the transitive clustering coefficient ($c^T(x)$), and its distance to the reference to copy the evolution of the network ($\hat{\mathcal{J}}(x)$). It is important to notice that the fastest convergence of the mean degree and the average clustering coefficient to specific values allows us to use constant values instead of functions when analyzing bigger networks. For each node insertion the algorithm has a complexity of $O(m * N * d^2)$, where N is the number of nodes in the network, d is the average node degree, and m is an integer value that depends on the difference between the network and the reference PDF.

The main difference between the CP and other models is the ability to capture oscillations during its evolution. Most models must be previously adjusted to create a network with fixed properties. The proposed model has the ability to dynamically adjust the topological properties step by step during the network's creation. Its main drawback when compared with the abovementioned models is that the decisions at each node inclusion are computationally more expensive: at each node inclusion the model has to solve an optimization problem that consists in determining how many and which links are necessary to reach the stage of the copied network. To improve its computational time, a heuristic-like procedure was included.

To better understand the consequences of optimizing the different input functions of the model, Figs. 1 and 2 show the effect of network creation by one ensemble of the Copycat Model when considering G^0 as a single node and $g(x) = 4$ for

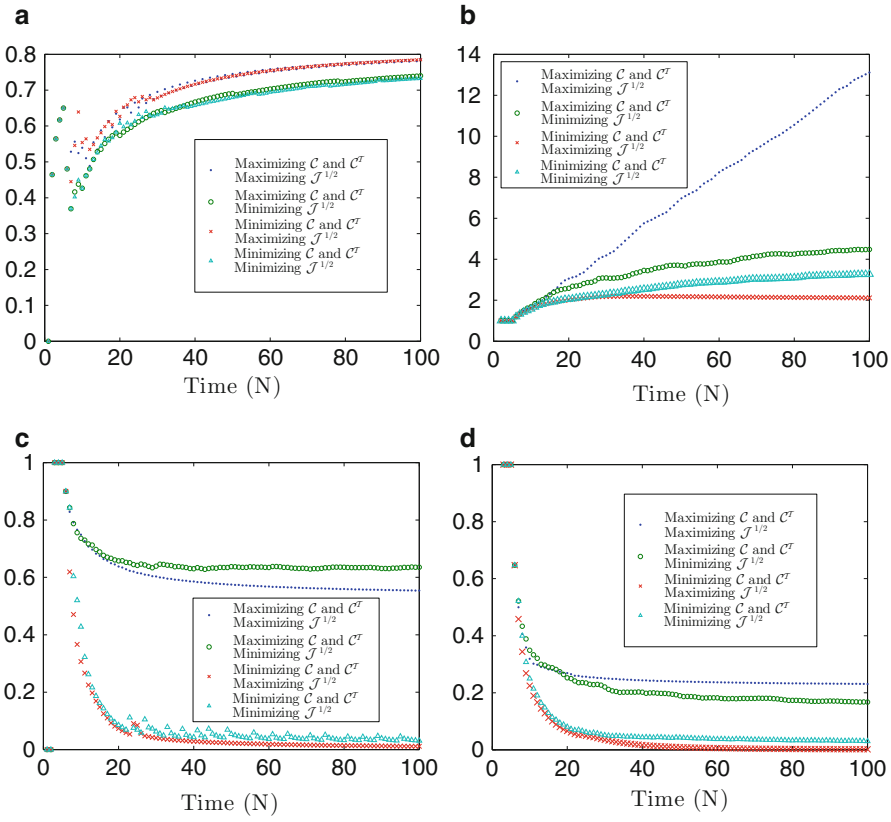

Fig. 1 Copycat model evolution for different values of the input functions C, C^T and $\mathcal{J}^{1/2}$ considering the average degree constant (equals 4) at each iteration (**a**) $\mathcal{J}^{1/2}$ (**b**) APL (**c**) C (**d**) C^T

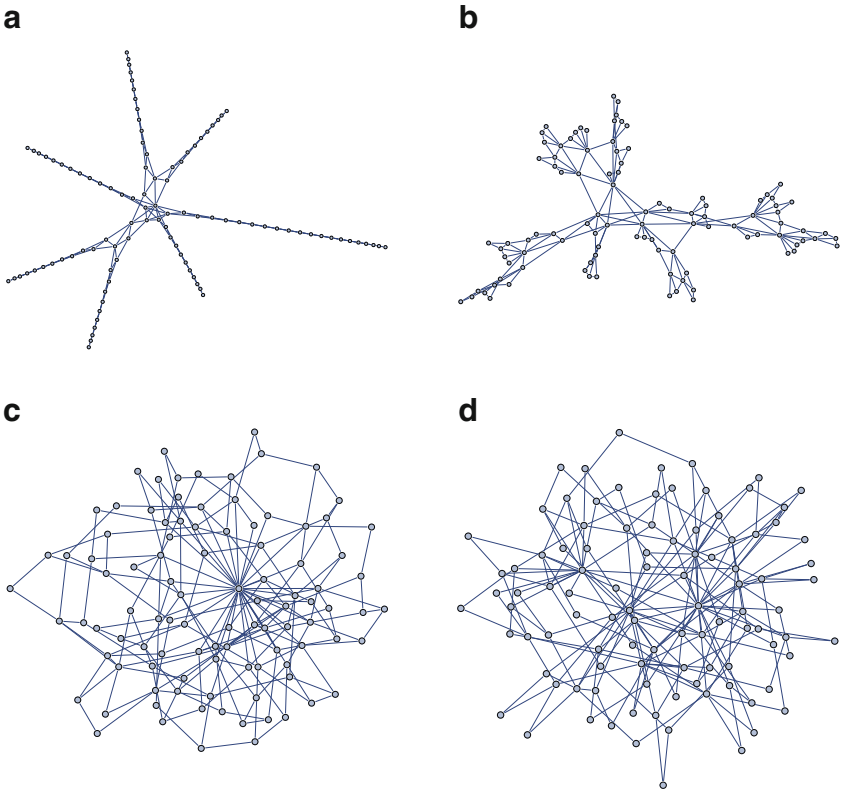

Fig. 2 Copycat model evolution for different values of the input functions C, C^T, and $\mathcal{J}^{1/2}$ considering the average degree constant (equals 4) at each iteration (see Fig. 1) (**a**) Maximizing C, C^T and $\mathcal{J}^{1/2}$ (**b**) Maximizing C and C^T. Minimizing $\mathcal{J}^{1/2}$ (**c**) Minimizing C and C^T. Maximizing $\mathcal{J}^{1/2}$ (**d**) Minimizing C, C^T and $\mathcal{J}^{1/2}$

all $x = 1, \ldots, 100$. At each time step, a new node is linked to the network by minimizing or maximizing the clustering coefficient functions and square root of the Jensen-Shannon divergence.

4 Predicting Network Growth

One way to mimic the evolution of complex networks was by looking at the time series generated by the evolution of some of its properties. In [17] we showed how the Copycat Model was able to reproduce the network's growth by looking at the time series associated to the clustering coefficient, the average degree, and

the square root of the Jensen-Shannon divergence if, at each time step, these values are known. In this section we show that we can use a simple predicting method for the input functions of the CP model. For this purpose, we choose to use the moving average method (MA) to predict the behavior of the time series. This well-known methodology was chosen because we are interested in the design of a framework as a general tool. However, when analyzing particular cases, other prediction methods could improve the results. The simple moving average (SMA) works by averaging the last n datum points; thus, $\text{SMA}_t = \frac{d_{t-1}+d_{t-2}+\cdots+d_{t-(n-1)}}{n}$. Different versions of this methodology can be obtained by applying weighting factors, e.g., weighted moving average or exponential moving average.

For each network tested we compute approximations of the functions $g(x)$, $c(x)$, $c^T(x)$, and $\hat{\mathcal{J}}(x)$ via SMA by averaging the last ten datum points. Let G^n be the evolution of the network G until it reaches size n and $\text{CP}(G^n, N, g, c, c^T, \hat{\mathcal{J}})$ the random network of size $N \geq n$ generated by Copycat Model. Then, given a $n, n_{\text{step}} \in \mathbb{N}$ we can define

$$\tilde{G}(n) = \begin{cases} \text{CP}(G^{n_0+\alpha n_{\text{step}}}, n, g, c, c^T, \hat{\mathcal{J}}), \; if \;\; n \in (n_0 + \alpha n_{\text{step}}, n_0 + (\alpha+1) n_{\text{step}}) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for some } \alpha \in \mathbb{N} \cup \{0\} \\ G^n, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$$

The random experiment consists in generating for each n 30 ensembles of $\tilde{G}(n)$ (different seeds). For each ensemble, we compute the square root of the Jensen-Shannon divergence, the average degree, the clustering coefficient, the transitive clustering coefficient, the average path length, and the diameter. For $n = |G|$, we compute the betweenness centrality, the closeness centrality, and the eigenvector centrality for each node in the network. For each n, we take the average of each measure computed and compare with the real ones.

We analyze the evolution of three networks, the Popularity vs Similarity Model, the Manufacturing Emails Network, and the Slashdot threads Network.

Popularity vs Similarity

The Popularity vs Similarity Model [15] describes the way new nodes in a scale-free network connect to others. In this model, a different dimension of attractiveness is proposed, and the mechanism of preferential attachment, called in this framework popularity, is no longer the only driver in the evolution of scale-free networks. The authors develop a framework where new connections, instead of preferring popular nodes, optimize certain trade-offs between popularity and similarity. This model finds an interesting geometric interpretation: each node of the network is represented by a single point in the plane and identified by its polar coordinates (r, θ). Two nodes

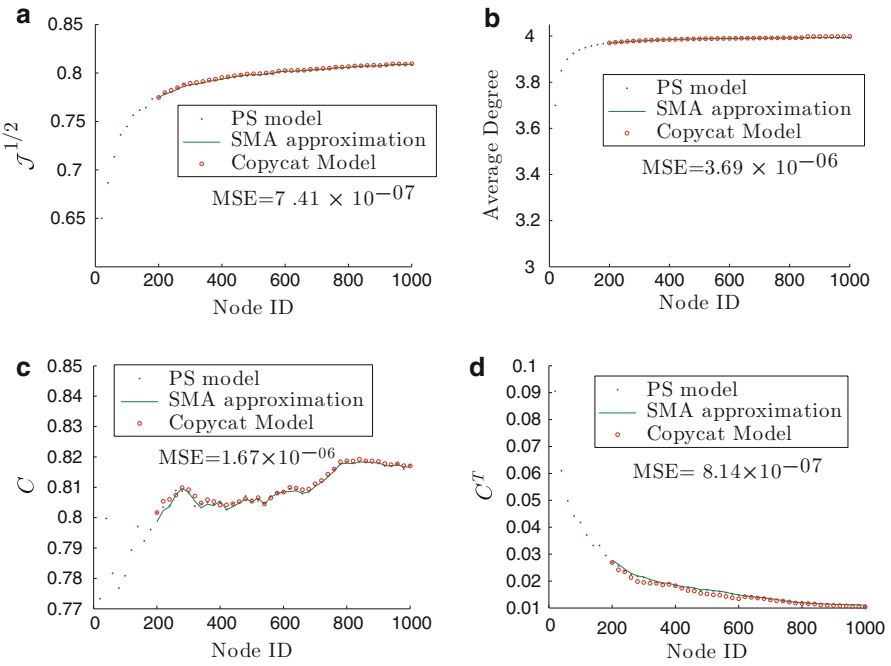

Fig. 3 Input functions of the CP model for the PS network: the mean squared error was performed for $n \geq 200$. (**a**) Evolution of the square root of the Jensen–Shannon divergence values; (**b**) evolution of the average degree; (**c**) evolution of the clustering coefficient; (**d**) evolution of the transitive clustering coefficient values

are said *similar* if they possess the same θ. Thus, we can define θ_{xy}, the angular distance between two nodes as the *similarity distance* between the nodes x and y. We simulate the Copycat Model for the PS model with parameter $m = 2$ and 1,000 nodes. We ran the computational experiment considering $n_0 = 200$ and $n_{\text{step}} = 200$. The results are shown in Figs. 3, 4, and 5. It is possible to see from these figures that all properties generated by the Copycat Model are predicted with a very good accuracy when compared with the original one, even those not considered in the design of the methodology.

Manufacturing Emails Network

The Manufacturing Emails Network is the internal email communication network between employees of a midsized manufacturing company [22–24]. The network is directed and the nodes represent employees. The left node represents the sender and the right node represents the recipient. Edges between two nodes are individual emails. For simplicity, for each of the directed graphs, we create their undirected

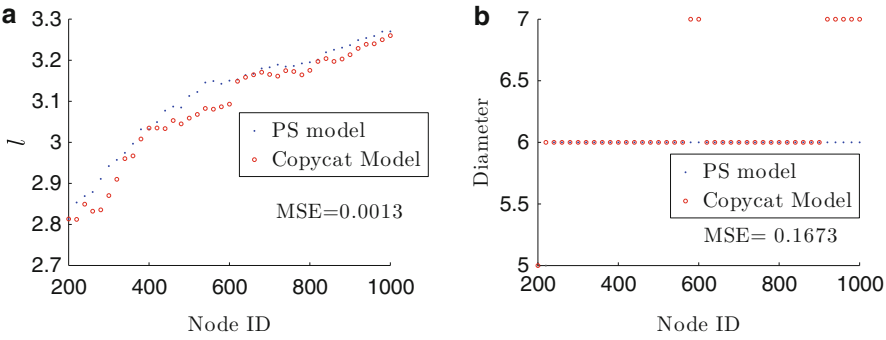

Fig. 4 Distance measures for the PS network: the mean squared error was performed for $n \geq 200$. (**a**) Average path length; (**b**) diameter

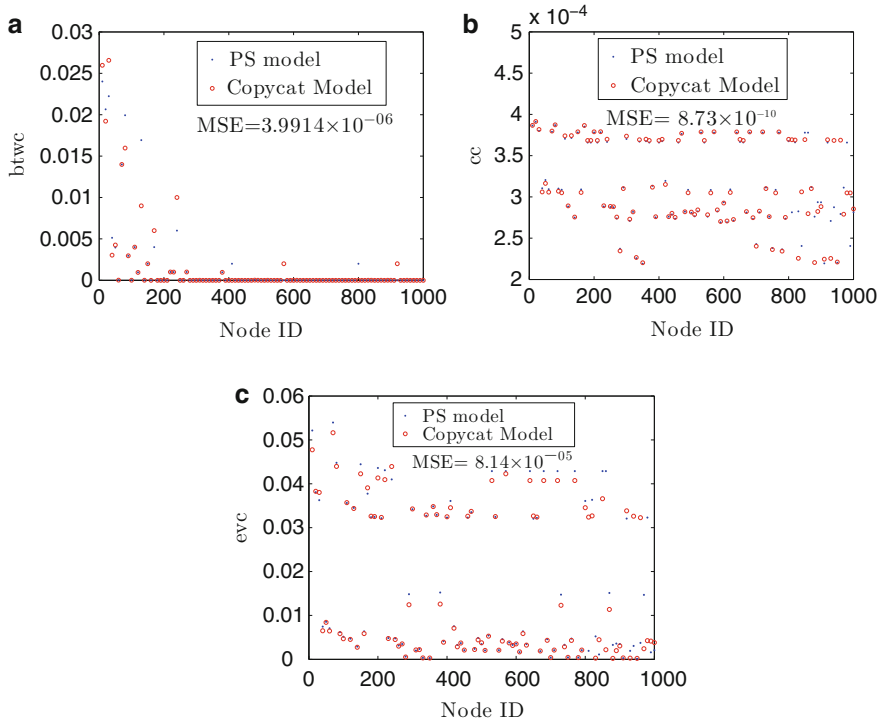

Fig. 5 Centrality measures for PS network: the mean squared error was performed for $n \geq 200$. (**a**) Betweenness; (**b**) closeness; and (**c**) eigenvector

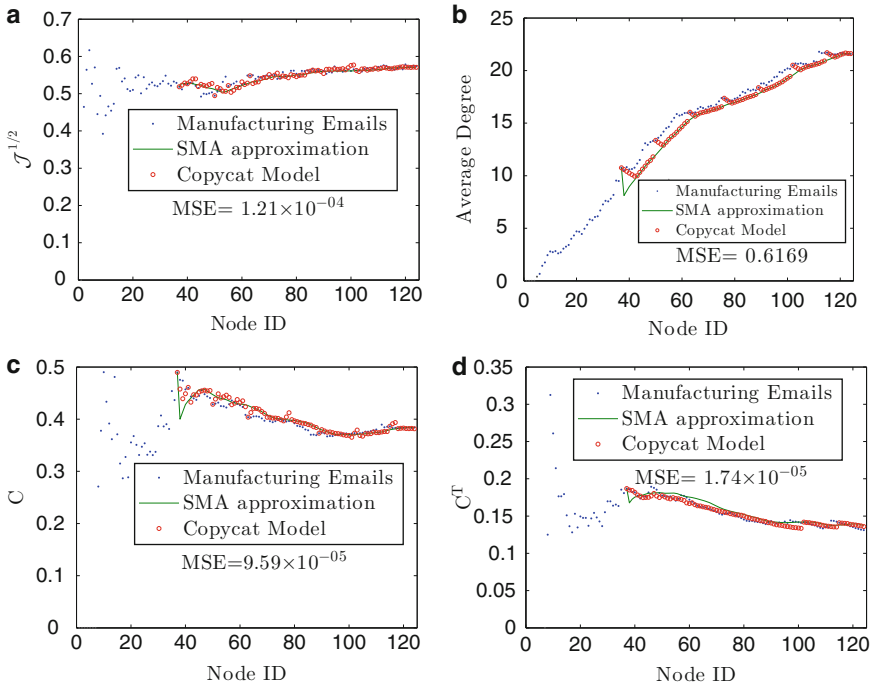

Fig. 6 Input functions of the CP model for the manufacturing emails network: the mean squared error was performed for $n \geq 37$. (**a**) Evolution of the square root of the Jensen–Shannon divergence values; (**b**) evolution of the average degree; (**c**) evolution of the clustering coefficient; (**d**) evolution of the transitive clustering coefficient

counterparts by taking into account only bidirectional links between the users resulting in an undirected network with 124 nodes. We also considered that the network evolves by the increase of node ID. We consider $n_0 = 37$ and $n_{\text{step}} = 12$. The results are presented in Figs. 6, 7, and 8. These figures show that all properties generated by the Copycat Model are predicted with a very good accuracy when compared with the original network, with the exception of the average degree and the diameter that show higher error values. It is well known that this network does not follow a scale-free behavior; this can be seen in the crescent pattern of the average degree. The use of SMA as a predicting methodology in a crescent time series provides a serious underestimation of the real values, causing this error on the framework. The use of a different methodology will certainly improve the results.

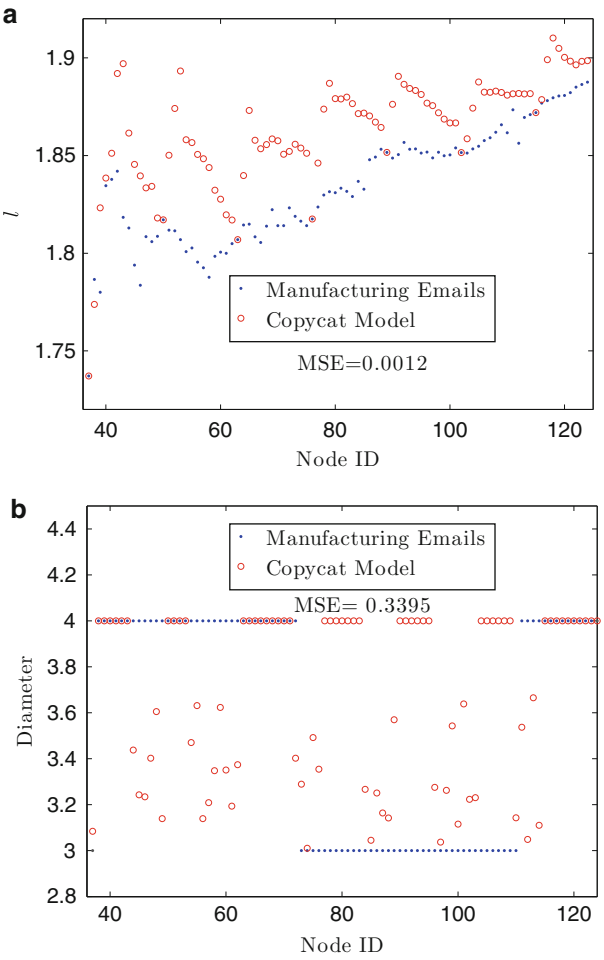

Fig. 7 Distance measures for manufacturing emails network: the mean squared error was performed for $n \geq 37$. (**a**) Average path length; (**b**) diameter

Slashdot Threads Network

The Slashdot Threads Network is the reply network of technology website Slashdot [24–26]. Nodes are users and edges are replies. The edges are directed and start from the responding user. Edges are annotated with the timestamp of the reply. For simplicity, for each of the directed graphs, we create their undirected counterparts by taking into account only bidirectional links between the users resulting in an undirected network with 10,707 nodes. We also considered that the network evolves

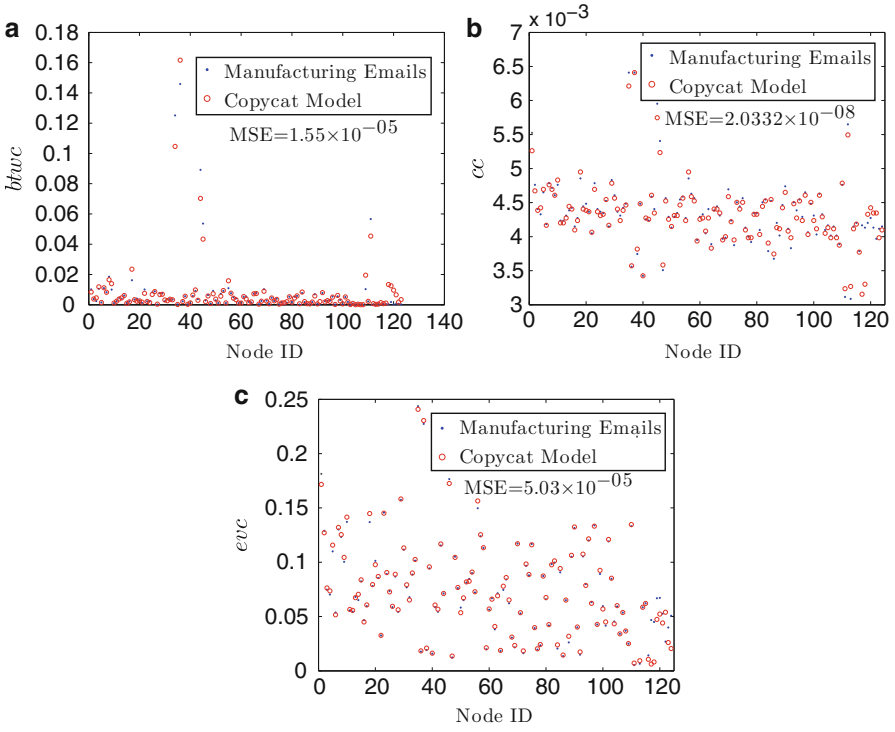

Fig. 8 Centrality measures for manufacturing emails network. (**a**) Betweenness; (**b**) closeness; and (**c**) eigenvector

by the increase of node ID. Here we consider $n_0 = 3{,}207$ and $n_{\text{step}} = 1{,}500$. The results show that all properties generated by the Copycat Model are predicted with a very good accuracy when compared with the original one. See Figs. 8, 9, 10 and 11.

5 Discussions

In this chapter we propose the use of the Copycat Model to capture and predict the dynamic behavior of networks. The methodology is based on Information Theory quantifiers that, when embedded in an optimization algorithm, creates a model able to reproduce the behavior of network's evolution. By using the SMA method for predicting the mean degree, the clustering coefficient, and the Jensen–Shannon divergence time series, the model is able to predict the main drivers of the network's growth.

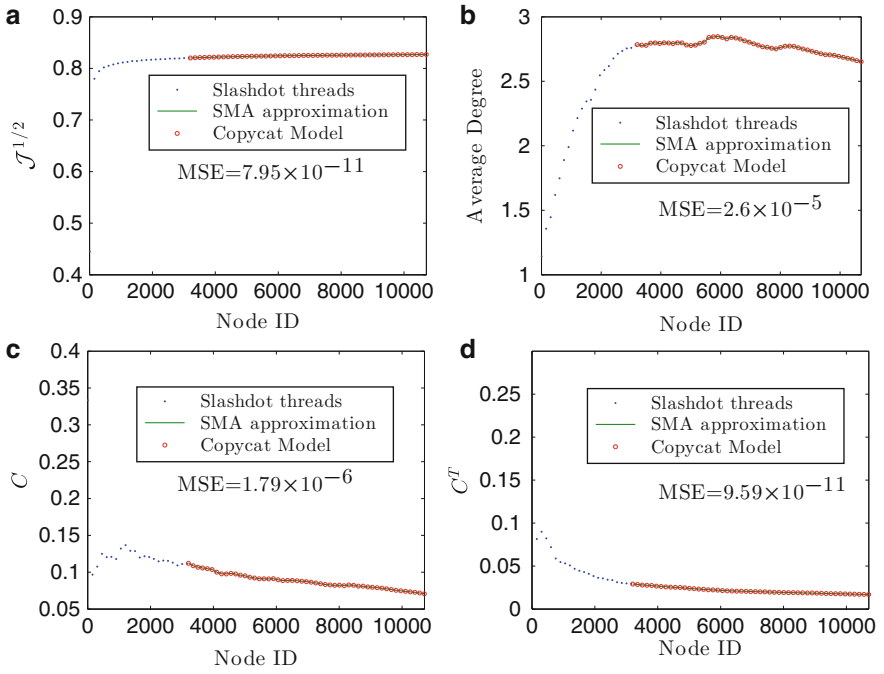

Fig. 9 Input functions of the CP model for the Slashdot Threads network: the mean squared error was performed for $n \geq 3{,}207$. (**a**) Evolution of the square root of the Jensen–Shannon divergence values; (**b**) evolution of the average degree; (**c**) evolution of the clustering coefficient; (**d**) evolution of the transitive clustering coefficient

The framework shows good results for all the cases analyzed. As expected the results are better for the prediction of scale-free networks. They possess almost constant values of clustering coefficient, mean degree, and Jensen–Shannon divergence values during their evolution, making easier its prediction. The good performance of the methodology depends on the quality of the prediction procedures as well as on the step of the prediction (n_{step}), characteristics that need to take into account, the time series profile, and the application analyzed. The use of this methodology allows the user to predict the network's growth as well as to conjecture about different scenarios.

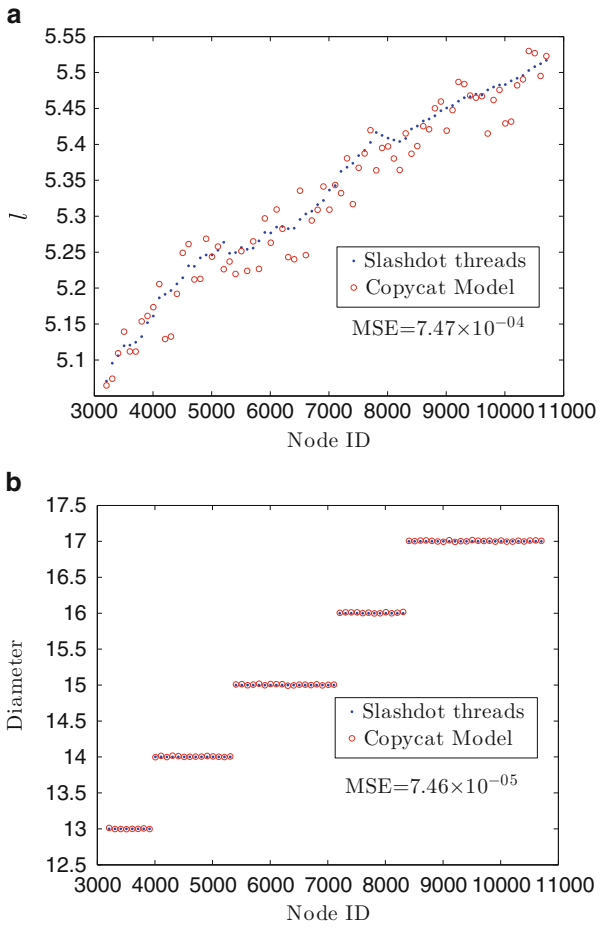

Fig. 10 Distance measures for the Slashdot Threads network: the mean squared error was performed for $n \geq 3{,}207$. (**a**) Average path length; (**b**) diameter

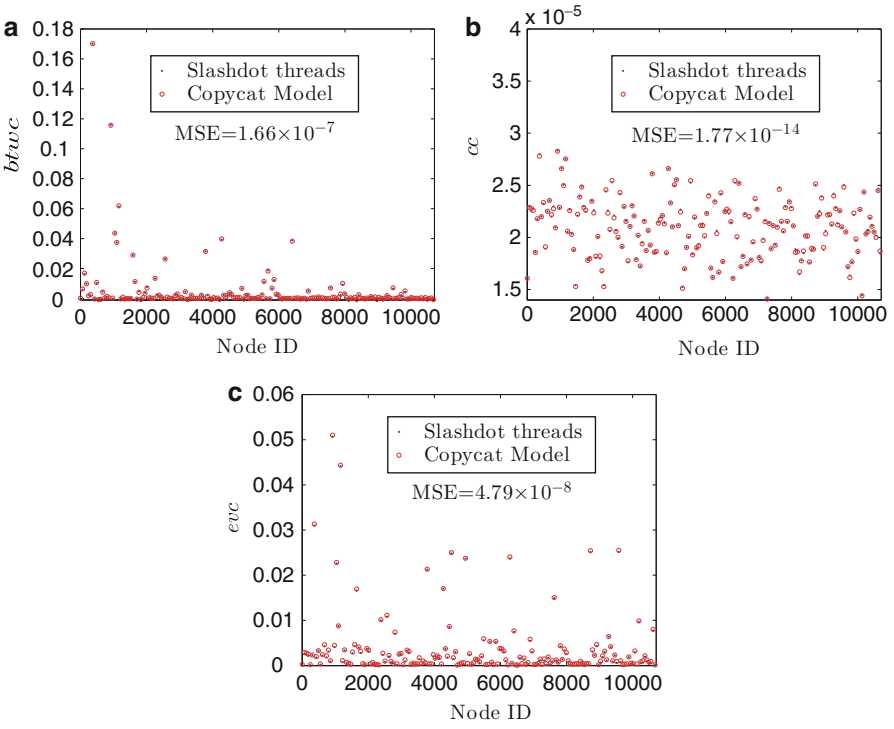

Fig. 11 Centrality measures for Slashdot Threads network. (**a**) Betweenness; (**b**) closeness; and (**c**) eigenvector

References

1. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, New York (2003)
2. Erdös , P., Rényi, A.: On random graphs, i. Publicationes Mathematicae (Debrecen) **6**, 290–297 (1959)
3. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. Science **286**(5439) 509–512 (1999). DOI:10.1126/science.286.5439.509
4. Krapivsky, P.L., Redner, S., Leyvraz, F.: Connectivity of growing random networks. Phys. Rev. Lett. **85**, 4629–4632 (2000)
5. Dorogovtsev, S.N., Mendes, J.F.F., Samukhin, A.N.: Structure of growing networks with preferential linking. Phys. Rev. Lett. **85**, 4633–4636 (2000)
6. Dorogovtsev, S.N., Mendes, J.F.: Evolution of networks with aging of sites. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **62**(2 Pt A), 1842–1845 (2000)
7. Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A **311**, 590–614 (2002)
8. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
9. Ghoshal, G., Newman, M.E.J.: Growing distributed networks with arbitrary degree distributions. Eur. Phys. J. B Condensed Matter Complex Syst. **58**(2), 175–184 (2007)

10. Diestel, R.: Graph Theory (Graduate Texts in Mathematics). Springer, Heidelberg (2005)
11. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data **1**(1), (2007)
12. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, pp. 462–470, ACM, New York (2008)
13. Barthélemy, M.: Spatial networks. Phys. Reports **499**(1-3), 1–101 (2011)
14. Herrera, C., Zufiria, P.J.: Generating scale-free networks with adjustable clustering coefficient via random walks. In: Proceedings of the 2011 IEEE Network Science Workshop, NSW '11, pp. 167–172, IEEE Computer Society, Washington, DC, USA (2011)
15. Papadopoulos, F., Kitsak, M., Angeles Serrano, M., Boguna, M., Krioukov, D.: Popularity versus similarity in growing networks. Nature **489**(7417), 537–540 (2012)
16. Barabasi, A.-L.: Network science: luck or reason. Nature **489**(7417), 507–508 (2012)
17. Alves Schieber, T., Gómez Ravetti, M.: Simulating the dynamics of scale-free networks via optimization. PLoS ONE **8**(12), e80783, 12 (2013)
18. Rodrigues, L., Travieso, G., Villas Boas, P.R.: Characterization of complex networks: a survey of measurements. Adv. Phys. **56**(1), 167–242 (2006)
19. Österreicher, F., Vajda, I.: A new class of metric divergences on probability spaces and its applicability in statistics. Ann. Inst. Stat. Math. **55**(3), 639–653 (2003)
20. Endres, D.M., Schindelin, J.E.: A new metric for probability distributions. IEEE Trans. Inf. Theory **49**(7), 1858–1860 (2003)
21. Feo, T.A., Resende, M.G.C.: A probabilistic heuristic for a computationally difficult set covering problem. Oper. Res. Lett. **8**(2), 67–71 (1989)
22. Manufacturing Emails Network Dataset - KONECT. http://konect.unikoblenz.de/networks/radoslaw_email. Accessed Aug 2014
23. Michalski, R., Palus, S., Kazienko, P.: Matching organizational structure and social network extracted from email communication. In: Lecture Notes in Business Information Processing, vol. 87, pp. 197–206. Springer, Berlin (2011)
24. Kunegis, J.: KONECT: the Koblenz network collection. In: Proceedings of an International Web Observatory Workshop, Rio de Janeiro, Brazil, pp. 1343–1350 (2013)
25. Slashdot Threads Network Dataset - KONECT. http://konect.unikoblenz.de/networks/slashdot-threads (2014). Accessed Aug 2014
26. Gómez, V., Kaltenbrunner, A., López, V.: Statistical analysis of the social network and discussion threads in slashdot. In: Proceedings of the 17th International Conference on World Wide Web, Beijing, China, pp. 645–654 (2008)

Optimal Control Formulations for the Unit Commitment Problem

Dalila B.M.M. Fontes, Fernando A.C.C. Fontes, and Luís A.C. Roque

Abstract The unit commitment (UC) problem is a well-known combinatorial optimization problem arising in operations planning of power systems. It involves deciding both the scheduling of power units, when each unit should be turned on or off, and the economic dispatch problem, how much power each of the on units should produce, in order to meet power demand at minimum cost while satisfying a set of operational and technological constraints. This problem is typically formulated as nonlinear mixed-integer programming problem and has been solved in the literature by a huge variety of optimization methods, ranging from exact methods (such as dynamic programming and branch-and-bound) to heuristic methods (genetic algorithms, simulated annealing, and particle swarm). Here, we discuss how the UC problem can be formulated with an optimal control model, describe previous discrete-time optimal control models, and propose a continuous-time optimal control model. The continuous-time optimal control formulation proposed has the advantage of involving only real-valued decision variables (controls) and enables extra degrees of freedom as well as more accuracy, since it allows to consider sets of demand data that are not sampled hourly.

Keywords Unit commitment problem • Power systems planning • Nonlinear optimization • Mixed-integer programming • Optimal control

D.B.M.M. Fontes (✉)
LIAAD - INESC TEC and Faculdade de Economia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
e-mail: fontes@fep.up.pt

F.A.C.C. Fontes
Instituto de Sistemas e Robótica do Porto and Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
e-mail: faf@fe.up.pt

L.A.C. Roque
Departamento de Matemática, Instituto Superior de Engenharia do Porto,
R. Dr. Ant. Bernardino de Almeida, 431, 4200-072 Porto, Portugal
e-mail: lar@isep.ipp.pt

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_6

1 Introduction

In this work, we address the unit commitment (UC) problem using optimal control methodologies. Despite being a highly researched problem with dynamical and multi-period characteristics, it appears that it has not been addressed by optimal control methods before, except in [12].

A problem that must be solved frequently by a power utility is to economically determine a schedule of which units are to be used and how much each unit should produce in order to meet the forecasted demand while satisfying operational and technological constraints, over a short time horizon [28, 29]. Good solutions are of most importance since they not only may provide substantial savings (tens to hundreds of millions of dollars) in operational and fuel costs but also maintain system reliability by keeping a proper spinning reserve [42]. Due to its combinatorial nature, multi-period characteristics, and nonlinearities, this problem is highly computationally demanding and, thus, solving the UC problem for real-sized systems is a hard optimization task: it is an NP-hard problem. The UC problem has been extensively studied in the literature. Several numerical optimization techniques, based both on exact and on approximate algorithms, have been reported.

Several approaches based on exact methods have been used, such as dynamic programming, mixed-integer programming, benders decomposition, lagrangian relaxation, and branch-and-bound methods; see, e.g., [2, 7, 20, 36]. The main drawbacks of these traditional techniques are the large computational time and memory requirements for large complexity and dimensionality problems. Dynamic programming [20, 26] is a powerful and flexible methodology; however it suffers from the dimensionality problem, not only in computational time but also in storage requirements. Recently a stochastic dynamic programming approach to schedule power plants was proposed [30]. In [2], a solution using lagrangian relaxation is proposed. However, the problem becomes too complex as the number of units increases and there are some difficulties in obtaining feasible solutions. Takriti [36] addresses the unit commitment problem by using mixed-integer programming which is a very hard task when the number of units increases since it requires large memory and leads to large computational time requirements. Other authors have proposed the use of mixed-integer linear programming to solve the linearized versions of the problem; see, e.g., [13,39]. The branch-and-bound method proposed in [7] uses a linear function to represent the fuel consumption and a time-dependent start-up cost, but has an exponential growth in the computational time with problem dimension.

More recently, several metaheuristic methods such as evolutionary algorithms and their hybrids have been proposed; see, e.g., [1, 6, 9, 34, 38]. These approaches have, in general, better performances than the traditional heuristics. The most commonly used metaheuristic methods are simulated annealing [25, 34], evolutionary programming [18, 27], memetic algorithms [38], particle swarm optimization [41], tabu search [24,40], and genetic algorithms [8,19,31,35]. For further discussion and

comparison of these methodologies, with special focus on metaheuristic methods, and other issues related to the unit commitment problem, see the very recent review by Saravanan et al. [32].

Although the UC problem is a highly researched problem with dynamical and multi-period characteristics, it appears that it has not been addressed before by optimal control methods, except in [12] as mentioned previously. In this work, the authors have formulated the UC problem as a discrete mixed-integer optimal control problem (OCP), which has then been converted into one with only real-valued controls. Here, we discuss formulations of the UC problem as an optimal control (OC) model and propose a new optimal control modeling approach. The model derived is a continuous one and only involves real-valued decision variables (controls).

The main contributions of the proposed modeling approach are twofold. Firstly, since it allows decisions to be taken at any time moment, and not only at specific points in time (usually, hourly), it may render better solutions. It should be noticed that the proposed approach allows for decisions about unit commitment/decommitment and about power production variation at any moment in time. Secondly, it no longer forces utilities to treat demand variations as instantaneous, i.e., time steps. In addition, if one chooses to use the approximated hourly data, as usual in the literature, the solution strategies (both regarding unit commitment/decommitment and power production) of the proposed model will approximate the discrete-time solutions since actions are only required to be taken hourly.

The remaining of this article is organized as follows. In Sect. 2, the UC problem is described and its mathematical programming formulation is given. The mixed-integer optimal control formulation and the variable time transformation that allows for rewriting it with only real-valued controls are given in Sect. 3. In Sect. 4, we provide a detailed description of the continuous-time optimal control model (OCM) including only real-valued controls, which is proposed here. Finally, in Sect. 5 we draw some conclusions and discuss future work.

2 The Unit Commitment Problem

The unit commitment problem involves both the scheduling of power units (i.e., the decision when each unit is turned on or turned off along a predefined time horizon) and the economic dispatch problem (the problem of deciding how much each unit that is on should produce). The scheduling of the units is an integer programming problem and the economic dispatch problem is a nonlinear (real-valued) programming problem. The UC problem is then a nonlinear, nonconvex, and mixed-integer optimization problem [8]. The objective of the UC problem is the minimization of the total operating costs over the scheduling horizon

while satisfying the system demand, the spinning reserve requirements, and other generation constraints such as capacity limits, ramp rate limits, and minimum uptime/downtimes.

The objective function is expressed as the sum of the fuel, start-up, and shutdown costs.

2.1 Mixed-Integer Mathematical Programming Model

The model has two types of decision variables: the binary variables and the real-valued variables. The binary decision variables $u_j(t)$ are either set to 1, meaning that unit j is committed at time t, or otherwise are set to zero. The real-valued variables $y_j(t)$ indicate the amount of power produced by unit j at time t. For the sake of simplicity, we also define the auxiliary variables $T_j^{\text{on/off}}(t)$, which represent the number of time periods for which unit j has been continuously online/off-line until time t.

Objective Function

The objective function has three cost components: generation costs, start-up costs, and shutdown costs. The generation costs, also known as the fuel costs, are conventionally given by the following quadratic cost function:

$$F_j(y_j(t)) = a_j \cdot (y_j(t))^2 + b_j \cdot y_j(t) + c_j, \tag{1}$$

where a_j, b_j, c_j are the cost coefficients of unit j.

The start-up costs, that depend on the number of time periods during which the unit has been off, are given by

$$S_j(t) = \begin{cases} S_{H,j}, & \texttt{if } T_{\min,j}^{\text{off}} \le T_j^{\text{off}}(t) \le T_{\min,j}^{\text{off}} + T_{c,j}, \\ S_{C,j}, & \texttt{if } T_j^{\text{off}}(t) > T_{\min,j}^{\text{off}} + T_{c,j}, \end{cases} \tag{2}$$

where $S_{H,j}$ and $S_{C,j}$ are, respectively, the hot and cold start-up costs of unit j and $T_{\min,j}^{\text{on/off}}$ is the minimum uptime/downtime of unit j. The shutdown costs S_{dj} for each unit, whenever considered in the literature, are not time dependent.

Therefore, the cost incurred with an optimal scheduling is given by the minimization of the total costs for the whole planning period.

Minimize

$$\sum_{t=1}^{T} \left(\sum_{j=1}^{N} \left\{ F_j(y_j(t)) \cdot u_j(t) + S_j(t) \cdot (1 - u_j(t-1)) \cdot u_j(t) \right\} + S_{dj} \cdot (1 - u_j(t)) \cdot u_j(t-1) \right). \tag{3}$$

Constraints

As said before, there are two types of constraints: the operational constraints and the technological constraints. The first set of constraints can be further divided into unit output range limit [Eq. (4)], maximum output variation, i.e., ramp rate constraints [Eq. (5)], and minimum number of time periods that a unit must be continuous in each status (online or off-line) [Eqs. (6) and (7)], while the second set of constraints can be divided into load requirements [Eq. (8)] and spinning reserve requirements [Eq. (9)].

$$Y\min_j \cdot u_j(t) \leq y_j(t) \leq Y\max_j \cdot u_j(t), \text{ for } t \in \{1,\ldots,T\} \text{ and } j \in \{1,\ldots,N\}. \tag{4}$$

$$-\Delta_j^{dn} \leq y_j(t) - y_j(t-1) \leq \Delta_j^{up}, \text{ for } t \in \{1,\ldots,T\} \text{ and } j \in \{1,\ldots,N\}. \tag{5}$$

$$T_j^{\text{on}}(t) \geq T_{\text{min},j}^{\text{on}}, \text{ for each time } t \text{ in which unit } j \text{ is turned off and } j \in \{1,\ldots,N\}. \tag{6}$$

$$T_j^{\text{off}}(t) \geq T_{\text{min},j}^{\text{off}}, \text{ for each time } t \text{ in which unit } j \text{ is turned on and } j \in \{1,\ldots,N\}. \tag{7}$$

$$\sum_{j=1}^{N} y_j(t) \cdot u_j(t) \geq D(t), t \in \{1,\ldots,T\}. \tag{8}$$

$$\sum_{j=1}^{N} Y\max_j \cdot u_j(t) \geq R(t) + D(t), t \in \{1,\ldots,T\}. \tag{9}$$

The parameters used in the above equations are defined as follows:

T: Number of time periods (hours) of the scheduling time horizon
N: Number of generation units
$R(t)$: System spinning reserve requirements at time t, in [MW]
$D(t)$: Load demand at time t, in [MW]
$Y\min_j$: Minimum generation limit of unit j, in [MW]
$Y\max_j$: Maximum generation limit of unit j, in [MW]
$T_{c,j}$: Cold start time of unit j, in [hours]
$T_{\text{min},j}^{\text{on/off}}$: Minimum uptime/downtime of unit j, in [hours]
$T_{j,0}^{\text{on}}$: Initial state of unit j at time 0, time since the last status switch off/on, in [hours]
$T_{j,0}^{\text{off}}$: Initial state of unit j at time 0, time since the last status switch on/off, in [hours]
Δ^{dn/up_j}: Maximum allowed output level decrease/increase in consecutive periods for unit j, in [MW]

3 Discrete-Time Optimal Control Approach

In this section, we describe the work in [12], where a mixed-integer OCM is proposed to the UC problem. Although it is possible to address OCPs with discrete control sets (see, e.g., [10, 14]), it is computationally demanding. Thus, it was proposed to convert this model into another OCM with only real-valued controls. The conversion process requires the use of a novel variable time transformation that was able to address adequately several discrete-valued control variables arising in the original problem formulation. Finally, the transformed real OCM was transcribed into a nonlinear programming problem (NLP) to be solved by a nonlinear optimization solver.

3.1 Discrete-Time Mixed-Integer Optimal Control Model

The mixed-integer OCM has two types of decision/control variables: on the one hand, binary control variables $u_j(t)$, which are either set to 1, meaning that unit j is committed at time t, or otherwise set to zero and on the other hand, real-valued variables $\Delta_j(t)$, which enable to control, by increasing or decreasing, the power produced by unit j at time t. We consider two types of state variables: variables $y_j(t)$, which represent the power generated by unit j at time t, and variables $T_j^{\text{on/off}}(t)$, which represent the number of time periods for which unit j has been continuously online/off-line until time t. For convenience, let us also define the index sets: $\mathbb{T} := \{1, \ldots, T\}$ and $\mathbb{J} := \{1, 2, \ldots, N\}$. The parameters related to the problem data are as defined in the previous section. The UC problem can now be formulated as a mixed-integer OCM.

Objective Function
Minimize

$$\sum_{t=1}^{T}\left(\sum_{j=1}^{N}\left\{F_j(y_j(t))u_j(t) + S_j(t)(1-u_j(t-1))u_j(t) + S_{dj}\cdot(1-u_j(t))\cdot u_j(t-1)\right\}\right), \tag{10}$$

where the costs are as before.

The State Dynamics
The state dynamics in this model are as follows:

The production of each unit, at time t, depends on the amount produced in the previous time period and is limited by the maximum allowed decrease and increase of the output that can occur during one time period:

$$y_j(t) = \left[y_j(t-1) + \Delta_j(t)\right].u_j(t), \text{ for } t \in \mathbb{T} \text{ and } j \in \mathbb{J}. \tag{11}$$

The number of time periods for which unit j has been continuously online until time t is given by

$$T_j^{\text{on}}(t) = \left[T_j^{\text{on}}(t-1)+1\right].u_j(t), \text{ for } t \in \mathbb{T} \text{ and } j \in \mathbb{J}. \tag{12}$$

The number of time periods for which unit j has been continuously off-line until time t is given by

$$T_j^{\text{off}}(t) = \left[T_j^{\text{off}}(t-1)+1\right].\left(1-u_j(t)\right), \text{ for } t \in \mathbb{T} \text{ and } j \in \mathbb{J}. \tag{13}$$

Pathwise Constraints
The constraints are as before, except for the ramp rate constraints, and thus they are given by Eq. (4) and Eqs. (6)–(9). The ramp rate constraints, which were given by Eq. (5), are now handled by the control constraints:

$$\Delta_j(t) \in \left[-\Delta_j^{dn}, \Delta_j^{up}\right], \text{ for } t \in \mathbb{T} \text{ and } j \in \mathbb{J}. \tag{14}$$

3.2 The Variable Time Transformation Method

The idea here is to develop a variable time transformation in order to convert the mixed-integer OCM into an OCM with only real-valued controls. The transformation of a mixed-integer OCP into a problem with only real-valued controls is not new nor is the general idea of a variable time transformation method. See the classical reference [17] and also [21–23,33,37]. See also the recent work [15] for a discussion on several variable time transformation methods.

Consider, for each unit j, a non-decreasing real-valued function $t \mapsto \tau_j(t)$. Consider also a set of values $\bar{\tau}_1, \bar{\tau}_2, \ldots$ such that when $\tau_j(t) = \bar{\tau}_k$ for odd k we have a transition from off to on for unit j and when $\tau_j(t) = \bar{\tau}_k$ for even k we have a transition from on to off. So, we consider that unit j is

- on if $\tau_j(t) \in [\bar{\tau}_1, \bar{\tau}_2) \cup [\bar{\tau}_3, \bar{\tau}_4) \cup \ldots \cup [\bar{\tau}_{2k-1}, \bar{\tau}_{2k})$;
- off if $\tau_j(t) \in [0, \bar{\tau}_1) \cup [\bar{\tau}_2, \bar{\tau}_3) \cup \ldots \cup [\bar{\tau}_{2k}, \bar{\tau}_{2k+1})$.

It might help to interpret τ_j to be a transformed time scale and that the values $\bar{\tau}_1, \bar{\tau}_2, \ldots$ are switching "times" in the transformed time scale. We can consider, without loss of generality, that the values $\bar{\tau}_k$ are equidistant. Nevertheless, in real time t, the distance between the two events $\bar{\tau}_k$ and $\bar{\tau}_{k+1}$ can be stretched or shrunk to any nonnegative value, including zero, depending on the shape of the function $t \mapsto \tau_j(t)$.

To simplify the exposition, and without loss of generality, let us consider that $\bar{\tau}_k - \bar{\tau}_{k-1}$ is constant and equal to 1, for all $k = 1, 2, \ldots$. In such case, unit j is

- on if $\tau_j(t) \in [1, 2) \cup [3, 4) \cup \ldots \cup [2k-1, 2k)$;

- off if $\tau_j(t) \in [0,1) \cup [2,3) \cup \ldots \cup [2k, 2k+1)$.

Now, consider the controls

$$w(t) \in [0,1], \quad t = 0, 1, \ldots, T-1,$$

that represent the increment from $\tau(t)$ to $\tau(t+1)$ such that

$$\tau(t) = \tau_0 + \sum_{k=0}^{t-1} w(k)$$

or

$$w(t) = \tau(t+1) - \tau(t), \quad \text{with } \tau(0) = \tau_0.$$

3.3 The Optimal Control Model with Real-Valued Controls

We recall the index set $\mathbb{J}$ and redefine $\mathbb{T}$ to be more consistent with usual discrete-time control formulations.

$\mathbb{T} := \{0, \ldots, T-1\}$ and $\mathbb{J} := \{1, 2, \ldots, N\}$.

In the same spirit, we redefine the control $\Delta_j(t)$ for $t \in \{0, \ldots, T-1\}$ to be the amount of power generation incremented or decremented for the next time period (rather than comparatively to the previous period).

Note that the controls are all real-valued and comprise

$\Delta_j(t) \in \left[-\Delta_j^{dn}, \Delta_j^{up}\right]$,
$w_j(t) \in [0,1]$.

Define the sets of time periods:

$$I_j^{\text{on}} := \{t \in \mathbb{T} : \tau_j(t) \in [2k-1, 2k), k \geq 1\},$$

$$I_j^{\text{off}} := \mathbb{T} \setminus I_j^{\text{on}},$$

$$I_j^{\text{off}>\text{on}} := \{t \in \mathbb{T} : \tau_j(t) \geq 2k+1, \tau_j(t-1) < 2k+1, k \geq 0\},$$

$$I_j^{\text{on}>\text{off}} := \{t \in \mathbb{T} : \tau_j(t) \geq 2k, \tau_j(t-1) < 2k, k \geq 1\}\}.$$

Finally, the unit commitment problem can be formulated as an OCP, as follows:
Minimize

$$\sum_{j=1}^{N} \left(\sum_{t \in I_j^{\text{on}}} F_j(y_j(t)) + \sum_{t \in I_j^{\text{off}>\text{on}}} S_j(t) + \sum_{t \in I_j^{\text{on}>\text{off}}} S_{dj}(t) \right), \tag{15}$$

subject to the dynamic constraints

$$\tau_j(t+1) = \tau_j(t) + w_j(t) \quad j \in \mathbb{J},\ t \in \mathbb{T}, \tag{16}$$

$$T_j^{\text{on}}(t+1) = \begin{cases} T_j^{\text{on}}(t) + 1 & j \in \mathbb{J},\ t \in I_j^{\text{on}}, \\ 0 & j \in \mathbb{J},\ t \in I_j^{\text{off}}, \end{cases} \tag{17}$$

$$T_j^{\text{off}}(t+1) = \begin{cases} T_j^{\text{off}}(t) + 1 & j \in \mathbb{J},\ t \in I_j^{\text{off}}, \\ 0 & j \in \mathbb{J},\ t \in I_j^{\text{on}}, \end{cases} \tag{18}$$

$$y_j(t+1) = \begin{cases} y_j(t) + \Delta_j(t) & j \in \mathbb{J},\ t \in I_j^{\text{on}}, \\ 0 & j \in \mathbb{J},\ t \in I_j^{\text{off}}, \end{cases} \tag{19}$$

the initial state constraints

$$T_j^{\text{on}}(0) = T_{j,0}^{\text{on}} \quad \text{(given)}, \tag{20}$$

$$T_j^{\text{off}}(0) = T_{j,0}^{\text{off}} \quad \text{(given)}, \tag{21}$$

$$\tau_j(0) = \begin{cases} 0 \text{ if } T_{j,0}^{\text{on}} = 0 \\ 1 \text{ if } T_{j,0}^{\text{on}} > 0, \end{cases} \tag{22}$$

$$y_j(0) = \begin{cases} 0 & \text{if } T_{j,0}^{\text{on}} = 0 \\ y_{j,0} \in [Y\min_j, Y\max_j] & \text{if } T_{j,0}^{\text{on}} > 0, \end{cases} \tag{23}$$

the control constraints

$$\Delta_j(t) \in \left[-\Delta_j^{dn}, \Delta_j^{up}\right], \tag{24}$$

$$w_j(t) \in [0,1], \tag{25}$$

and the pathwise state constraints

$$y_j(t) \in [Y\min_j, Y\max_j] \quad j \in \mathbb{J},\ t \in I_j^{\text{on}}, \tag{26}$$

$$\sum_{j \in \mathbb{J}} y_j(t) \geq D(t) \quad t = 1, 2, \ldots, T, \tag{27}$$

$$\sum_{j \in \mathbb{J}} Y\max_j(t) \geq R(t) + D(t) \quad t = 1, 2, \ldots, T, \tag{28}$$

where $Y\max_j(t) = Y\max_j$ if $t \in I_j^{\text{on}}$, $Y\max_j(t) = 0$ otherwise

$$y_j(t) \in [Y_{\min_j}, \max\{Y_{\min_j}, \Delta_j^{up}\}] \quad j \in \mathbb{J}, t \in I_j^{\text{off}>\text{on}}, \tag{29}$$

$$T_j^{\text{on}}(t-1) \geq T_{\min,j}^{\text{on}} \quad j \in \mathbb{J},\ t \in I_j^{\text{on}>\text{off}}, \tag{30}$$

$$T_j^{\text{off}}(t-1) \geq T_{\min,j}^{\text{off}} \quad j \in \mathbb{J},\ t \in I_j^{\text{off}>\text{on}}. \tag{31}$$

3.4 Conversion into a Nonlinear Programming Problem

To construct the NLP, we start by defining the optimization variable x containing both the control and state variables. That is

$$x = [\Delta, w, \tau, T^{\text{on}}, T^{\text{off}}, y]$$

with dimension $(6T + 1) \times N$).

(We could have considered just the controls Δ, w together with the free initial state $y(0)$. An option which, despite having the advantage of a lower dimensional decision variable, is known to frequently have robustness problems, specially in OCPs with pathwise state constraints such as ours. For further discussion, see, e.g., Betts [3].)

The objective function should be rewritten in terms of x: Minimize $J(x)$ over x.

To facilitate the optimization algorithm, we separate the constraints that are simple variable bounds, linear equalities, linear inequalities, and the remaining:

- upper/lower bounds: Eqs. (24)–(26);
- linear equalities: Eq. (16);
- linear inequalities: Eq. (27);
- nonlinear equalities: Eqs. (17)–(19); and
- nonlinear inequalities: Eqs. (28)–(31).

Note that Eqs. (20)–(23) are not implemented as constraints since the initial values of these state variables are considered as parameters and not variables.

With these considerations the problem is formulated as the following NLP:

$$\begin{aligned}
&\text{Minimize}_{x \in \mathbb{R}^{(6T+1)\times N}}\, J(x)\\
&\text{subject to}\\
&\quad LB \le x \le UB\\
&\quad A_{\text{eq}} x = b_{\text{eq}}\\
&\quad A_{\text{ineq}} x \le b_{\text{ineq}}\\
&\quad g(x) = 0\\
&\quad h(x) \le 0.
\end{aligned}$$

More specifically

Minimize over x

$$J(x) = \sum_{j=1}^{N} \left(\sum_{t \in I_j^{\text{on}}} F_j(y_j(t)) + \sum_{t \in I_j^{\text{off>on}}} S_j(t) + \sum_{t \in I_j^{\text{on>off}}} S_{dj}(t) \right),$$

Subject to

- lower bounds:

$$\begin{aligned}
&\Delta_j(t) \geq -\Delta_j^{dn}, \text{ for } t \in \mathbb{T} \text{ and } j \in \mathbb{J},\\
&w_j(t) \geq 0, \quad j \in \mathbb{J},\ t \in \mathbb{T};\\
&\tau_j(t) \geq 0, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&T_j^{\text{on}}(t) \geq 0, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&T_j^{\text{off}}(t) \geq 0, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&y_j(t) \geq 0, \quad j \in \mathbb{J},\ t \in \mathbb{T};
\end{aligned}$$

- upper bounds:

$$\begin{aligned}
&\Delta_j(t) \leq \Delta_j^{up}, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&w_j(t) \leq 1, \quad j \in \mathbb{J},\ t \in \mathbb{T};\\
&\tau_j(t) \leq T, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&T_j^{\text{on}}(t) \leq 2T, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&T_j^{\text{off}}(t) \leq 2T, \quad j \in \mathbb{J},\ t \in \mathbb{T},\\
&y_j(t) \leq Y\max_j, \quad j \in \mathbb{J},\ t \in \mathbb{T};
\end{aligned}$$

- linear equalities:

$$\tau_j(t+1) - \tau_j(t) - w_j(t) = 0 \quad j \in \mathbb{J},\ t \in \mathbb{T};$$

- linear inequalities:

$$\textstyle\sum_{j\in\mathbb{J}} y_j(t) - D(t) \geq 0 \quad t \in \mathbb{T};$$

- nonlinear equalities:

$$T_j^{\text{on}}(t+1) = \begin{cases} T_j^{\text{on}}(t) + 1 & \text{if } j \in \mathbb{J},\ t \in I_j^{\text{on}},\\ 0 & \text{if } j \in \mathbb{J},\ t \in I_j^{\text{off}}, \end{cases}$$

$$T_j^{\text{off}}(t+1) = \begin{cases} T_j^{\text{off}}(t) + 1 & \text{if } j \in \mathbb{J},\ t \in I_j^{\text{off}},\\ 0 & \text{if } j \in \mathbb{J},\ t \in I_j^{\text{on}}, \end{cases}$$

$$y_j(t+1) = \begin{cases} y_j(t) + \Delta_j(t) & \text{if } j \in \mathbb{J},\ t \in I_j^{\text{on}},\\ 0 & \text{if } j \in \mathbb{J},\ t \in I_j^{\text{off}}; \end{cases}$$

- nonlinear inequalities:

$$\begin{aligned}
&y_j(t) \geq Y\min_j \quad j \in \mathbb{J},\ t \in I_j^{\text{on}},\\
&\textstyle\sum_{j\in\mathbb{J}} Y\max_j(t) - R(t) - D(t) \geq 0 \quad t \in \mathbb{T},\\
&y_j(t) - Y_{\min_j} \geq 0 \quad j \in \mathbb{J}, t \in I_j^{\text{off>on}},\\
&y_j(t) - \max\{Y_{\min_j}, \Delta_j^{up}\} \leq 0 \quad j \in \mathbb{J}, t \in I_j^{\text{off>on}},\\
&T_j^{\text{on}}(t-1) - T_{\min,j}^{\text{on}} \geq 0 \quad j \in \mathbb{J},\ t \in I_j^{\text{on>off}},\\
&T_j^{\text{off}}(t-1) - T_{\min,j}^{\text{off}} \geq 0 \quad j \in \mathbb{J},\ t \in I_j^{\text{off>on}}.
\end{aligned}$$

Of course, since this (real-valued) NLP is a problem that originally was a MI-NLP, it is still a very hard problem. Namely, it is a nonconvex problem and standard

NLP solvers will find just a local, not necessarily global, optimum. Nevertheless, this is very useful since it can be embedded, as a local search optimizer, into a global search heuristic method.

4 Continuous-Time Optimal Control Approach

In this section, we develop a continuous-time optimal control formulation for the unit commitment problem that uses only real-valued decision variables.

To introduce the ideas and concepts used in this formulation let us start by analyzing a specific and simple situation.

Consider a generation unit for which the minimum time it must be consecutively on is 2 h ($T^{\text{on}}_{\text{min}} = 2$) and the minimum time it must be consecutively off is 3 h ($T^{\text{off}}_{\text{min}} = 3$). Furthermore, consider also the unit to be initially off-line. Let the unit be turned off and turned on as soon as the elapsed time reaches $T^{\text{on}}_{\text{min}}$ and $T^{\text{off}}_{\text{min}}$, respectively. Such a strategy corresponds to the unit having the maximum number of status switches. Thus, for a 24 h period, we would obtain a profile as given in Fig. 1.

For the example just described, the times at which status switching occurs are given by

$$t_{i+1} = \begin{cases} t_i + T^{\text{on}}_{\text{min},j}, & \text{if } i \text{ is odd,} \\ t_i + T^{\text{off}}_{\text{min},j}, & \text{if } i \text{ is even.} \end{cases}$$

All other feasible status switching strategies can be obtained from the one just described by stretching any number of time intervals $[t_i, t_{i+1})$ with $i = 1, \ldots, S$, where S—the maximum number of status switches that can occur within the 24-h scheduling period—is given by

$$S = 1 + 2 * \left(24 \text{ DIV } \left(T^{\text{on}}_{\text{min},j} + T^{\text{off}}_{\text{min},j}\right)\right),$$

where DIV denotes integer division.

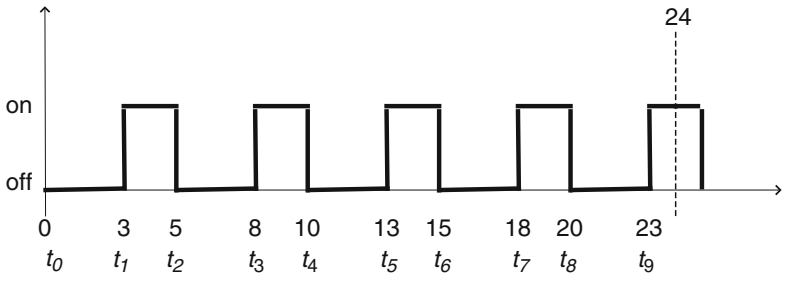

Fig. 1 Unit status, when the status switching strategy is as often as possible

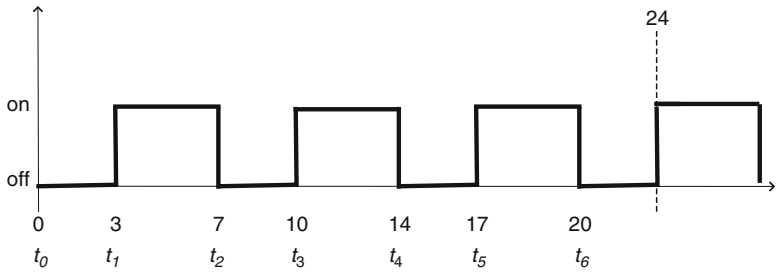

Fig. 2 Status of unit obtained with $\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_9] = [1, 2, 1, 2, 1, 2, 1, 1, 1, 1]$

The stretching magnitude α_i in the time interval $[t_i, t_{i+1})$ is bounded from below by 1, since the interval is initially defined as small as possible, and from above by $[1, (24-t_i)/T_{\min}]$, where $T_{\min}$ is set to $T^{\text{on}}_{\min,j}$ or $T^{\text{off}}_{\min,j}$ depending on whether i is odd or even, respectively, which allows for reaching the end of the scheduling period. It should be noticed that all switches occur at times $t_i \leq 24 - T_{\min}$ with $T_{\min}$ as defined.

Using a convenient selection of the α_i's we can generate any admissible switching profile. For example, choosing $\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_9] = [1, 2, 1, 2, 1, 2, 1, 1, 1, 1]$ leads to the profile given in Fig. 2.

Therefore, in any situation, the computation of the switching times is given by

$$t_{i+1} = \begin{cases} t_i + \alpha_i T^{\text{on}}_{\min,j} & \text{if } i \text{ is odd,} \\ t_i + \alpha_i T^{\text{off}}_{\min,j} & \text{if } i \text{ is even.} \end{cases}$$

4.1 Formulation

Let us define some parameters before introducing the formulation. When considering several units, the maximum number of switches is not the same for all units since they may have different limits on the number of periods that must elapse before a switch is possible. The same is true for the maximum magnitude of the stretch. Therefore and in order to have one single value for these parameters, we compute upper bounds rather than their true value. By defining

$$T^{\text{on+off}}_{\min} = \min_j \{T^{\text{on}}_{\min,j} + T^{\text{off}}_{\min,j}\}$$

we obtain a limit for the maximum number of switches as

$$S = 1 + 2 * 24 \text{ DIV } \left(T^{\text{on+off}}_{\min}\right)$$

and for the maximum magnitude of the stretch of an interval as

$$s_{\max} = 24/\min_j\{T^{\text{on}}_{\min,j}, T^{\text{off}}_{\min,j}\}.$$

For convenience, let us also define the index sets:

$$\mathbb{I} := \{0, 1, \ldots, S\}\text{—switching times indexes,}$$
$$\mathbb{J} := \{1, 2, \ldots, N\}\text{—generation unit indexes,}$$

and the time horizon

$$\mathbb{T} := [0, 24]\text{—time horizon interval.}$$

Decision/Control Variables
The model has two types of control variables, since two types of decisions are taken. On the one hand, one has to decide for how much time each unit is in each status, that is the magnitude of stretch applied to each time interval for each unit, $\alpha_{i,j}$. On the other hand, one also must decide on the amount of power production for each unit at each time instant. In our case, we do this by deciding on the variation of the production at each time instant $\delta_j(t)$.

$\alpha_{i,j}$: stretch magnitude applied to the time interval $[t_i, t_{i+1})$ for unit j. These are real-valued variables in the range $[1, s_{\max}]$.
$\delta_j(t)$: rate of change (increase or decrease) for the production of unit j at instant t. These variables are also real-valued and must be within $[-\Delta_j^{dn}, \Delta_j^{up}]$.

State Variables
The state variables characterize the system and are as follows:

$t_{i,j}$: ith switching time of unit j;
$u_{i,j}$: Status of unit j in the interval $[t_i, t_{i+1})$, (1 if the unit is on; 0 otherwise);
$u_j(t)$: Status of unit j at instant t, (1 if the unit is on; 0 otherwise);
$y_j(t)$: Power generation of unit j at instant t, in [MW].

Objective Function
The objective of the UC problem is the minimization of the total costs for the whole planning period, in which the total costs are expressed as the sum of fuel costs and start-up and shutdown costs of the generating units. Therefore, the objective function is as follows:

Minimize

$$\sum_{j\in\mathbb{J}}\int_0^T \Big(F_j\left(y_j(t)\right)u_j(t) + S_j(t)\left(1-u_j(t-1)\right)u_j(t)$$
$$+S_{dj}\cdot(1-u_j(t))\cdot u_j(t-1)\Big)\,dt.$$

Dynamic Constraints

We must define the unit status during each time interval. Unit j must have its status switched at the beginning of each interval $[t_i, t_{i+1})$. Thus if in the interval $[t_i, t_{i+1})$ the unit was 1 (on), then in the interval $[t_{i+1}, t_{i+2})$ it becomes 0 (off) and vice versa.

$$u_{i+1,j} = |u_{i,j} - 1|, \quad j \in \mathbb{J}, i \in \mathbb{I}.$$

The ending time instant of a time interval, which is the beginning of the next one, is obtained by adding up the starting time instant with the length of the interval.

$$t_{i+1,j} = t_{i,j} + \alpha_{i,j}[T^{\text{on}}_{\text{min},j} u_{i,j} + T^{\text{off}}_{\text{min},j}(1 - u_{i,j})], \quad j \in \mathbb{J}, i \in \mathbb{I}.$$

In addition, we also must define the power production at each time instant and, for convenience, also the unit status at each time instant.

$$u_j(t) = u_{i,j}, \quad j \in \mathbb{J}, i \in \mathbb{I}, t \in [t_i, t_{i+1}),$$

$$y_j(t) = \begin{cases} 0 & \text{if } u_j(t) = 0, \\ y_j(t_i) + \int_{t_i}^{t} \delta_j(s)ds, \text{with } i = \max\{i : t_i \le t\}, & \text{if } u_j(t) = 1, \end{cases} \quad t \in \mathbb{T}, j \in \mathbb{J}.$$

Control Constraints

Due to the mechanical characteristics and thermal stress limitations, the instantaneous output variation level of each online unit is restricted by ramp rate constraints, both up and down.

$$\delta_j(t) \in [-\Delta_j^{dn}, \Delta_j^{up}], \quad j \in \mathbb{J}, t \in \mathbb{T}.$$

The magnitude of the stretch is limited both from below and from above, since one must assure that the $T^{\text{on/off}}_{\text{min},j}$ are satisfied and that the scheduling does not go beyond the scheduling horizon.

$$\alpha_{i,j} \in [1, A], \text{ for } i \in \mathbb{I}, j \in \mathbb{J},$$

$$\text{with } A = \begin{cases} \frac{24 - t_j}{T^{\text{on}}_{\text{min},j} u_{i,j} + T^{\text{off}}_{\text{min},j}(1 - u_{i,j})} & \text{if } t_i \le 24 - T^{\text{on}}_{\text{min},j} u_{i,j} + T^{\text{off}}_{\text{min},j}(1 - u_{i,j}), \\ 1 & \text{otherwise.} \end{cases}$$

Pathwise State Constraints

Each unit has maximum and minimum output capacity limits.

$$y_j(t) \in [Y\min_j u_j(t), Y\max_j u_j(t)] \quad j \in \mathbb{J}, \ t \in \mathbb{T}.$$

The power generated at each time instant must meet the respective load demand.

$$\sum_{j \in \mathbb{J}} y_j(t) \ge D(t) \quad t \in \mathbb{T}.$$

The spinning reserve is the amount of real power available from online units net of their current production level and it must satisfy a pre-specified value, at each time instance.

$$\sum_{j \in \mathbb{J}} Y\max_j(t) u_j(t) \geq R(t) + D(t) \quad t \in \mathbb{T}.$$

Initial State Constraints
The initial status of each unit is given.

$$u_{0,j} = \text{Initial Status}_j, \quad j \in \mathbb{J}.$$

Also

$$u_j(0) = u_{0,j}, \quad j \in \mathbb{J}.$$

The first switching interval starts at the beginning of the scheduling horizon and thus

$$t_{0,j} = 0, \quad j \in \mathbb{J}.$$

Finally, the power production of each online unit has to be within its capacity limits.

$$y_j(0) \in [Y\min_j u_{j,0}, Y\max_j u_{j,0}], \quad j \in \mathbb{J}.$$

The numerical solution of continuous-time OCPs has been a well-studied subject for many decades [5] and also has been having recent developments and available solvers such as ICLOCS [11], BOCOP [4], and ACADO [16]. Although the use of one of these solvers is recommended, an alternative is always to discretize the problem, transcribe it into a NLP, and use directly an NLP solver.

The use of a continuous-time formulation for the UC problem has some advantages: (i) the possibility of accommodating any changes in the data or parameters that occur not an hourly basis, but at any time in between; (ii) in particular, the formulation proposed can deal with continuous-time varying demand (which is more realistic), resulting in an output strategy that responds with continuous-time variations; (iii) however, in case the demand and all remaining data vary only on an hourly basis, the resulting output strategy will follow very closely to the one obtained with a discrete-time model; (iv) the complexity of the optimization problem obtained is not increased, possibly being easier to find an optimal solution, since the decision variables involved are all real-valued. It is well known that real-valued NLPs are, in general, less difficult to solve than mixed-integer NLPs.

5 Conclusions

We have addressed the UC problem, a well-researched problem in the literature, which is usually formulated using a mixed-integer nonlinear programming model. Here we have explored the formulation of this problem using OCMs. Previous works on an optimal control approach to the UC problem, as far as we are aware of, are limited to the work in [12] that uses a discrete-time OCM.

We have proposed here a formulation of the UC problem using a continuous-time OCM. An interesting feature of the continuous-time formulation is the fact that, contrary to the usual mixed-integer programming models in the literature, all decision variables are real-valued, which enables the use of more efficient optimization methods for its solution.

Additional advantages of the continuous-time optimal control formulation are the possibility of dealing more accurately with data that is provided with irregular or fast-sampled time intervals, or even continuous-time varying. In particular, this formulation can deal appropriately with continuous-time varying demand data.

Acknowledgments This research has been partially supported by Marie Curie project FP7-ITN-264735-SADCO, ERDF (FEDER), and COMPETE through FCT projects PTDC/EEI-AUT/1450/2012 and FCOMP-01-0124-FEDER-037281 and by North Portugal Regional Operational Programme (ON.2 O Novo Norte), under the National Strategic Reference Framework.

References

1. Abookazemi, K., Mustafa, M.W., Ahmad, H.: Structured genetic algorithm technique for unit commitment problem. Int. J. Recent Trends Eng. **1**(3), 135–139 (2009)
2. Bakirtzis, A.G., Zoumas, C.E.: Lambda of Lagrangian relaxation solution to unit commitment problem. Proc. Inst. Elect. Eng. Gener. Trans. Distr. **147**(2), 131–136 (2000)
3. Betts, J.: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2010)
4. Bonnans, F., Giorgi, D., Grelard, V., Maindrault, S., Martinon, P.: BOCOP: User Guide. http://bocop.saclay.inria.fr/docs/UserGuide-BOCOP.pdf (2014)
5. Bryson, A.E., Ho, Y.C.: Applied Optimal Control. Wiley, New York (1975)
6. Chen, Y.M., Wang, W.S.: Fast solution technique for unit commitment by particle swarm optimisation and genetic algorithm. Int. J. Energ. Tech. Pol. **5**(4), 440–456 (2007)
7. Cohen, A.I., Yoshimura, M.: A branch-and-bound algorithm for unit commitment. IEEE Trans. Power Apparatus Syst. **102**(2), 444–451 (1983)
8. Dang, C., Li, M.: A floating-point genetic algorithm for solving the unit commitment problem. Eur. J. Oper. Res. **181**(4), 1370–1395 (2007)
9. Dudek, G.: Unit commitment by genetic algorithm with specialized search operators. Electr. Power Syst. Res. **72**(3), 299–308 (2004)
10. Khmelnitsky, E.: A combinatorial, graph-based solution method for a class of continuous-time optimal control problems. Math. Oper. Res. **27**, 312–325 (2002)
11. Falugi, P., Kerrigan E., van Wyk, E.: Imperial college london optimal control software user guide (iclocs) http://www.ee.ic.ac.uk/ICLOCS/user_guide.pdf (2010)

12. Fontes, F.A.C.C., Fontes, D.B.M.M, Roque, L.A.C.: An optimal control approach to the unit commitment problem. In: CDC: Conference on Decision and Control, pp. 7069–7074, Maui, USA (2012)
13. Frangioni, A., Gentile, C., Lacalandra, F.: Tighter approximated MILP formulations for unit commitment problems. IEEE Trans. Power Syst. **24**(1), 105–113 (2009)
14. Gerdts, M.: Solving mixed-integer optimal control problems by branch and bound: a case study from automobile test driving with gear shift. Optim. Contr. Appl. Methods **26**, 1–18 (2005)
15. Gerdts, M.: A variable time transformation method for mixed-integer optimal control problems. Optim. Control Appl. Meth. **27**, 169–182 (2006)
16. Houska, B., Ferreau, H.J., Diehl, M.: Acado toolkit: an open-source framework for automatic control and dynamic optimization. Optim. Control Appl. Meth. **32**(3), 298–312 (2011)
17. Ioffe, A.D., Tihomirov, V.M.: Theory of Extremal Problems, Studies in Mathematics and its Applications, vol. 6. North-Holland Publishing Company, Amsterdam (1979)
18. Juste, K.A., Kita, H., Tanaka, E., Hasegawa, J.: An evolutionary programming solution to the unit commitment problem. IEEE Trans. Power Syst. **14**(4), 1452–1459 (1999)
19. Kazarlis, S.A., Bakirtzis, A.G., Petridis, V.: A genetic algorithm solution to the unit commitment problem. IEEE Trans. Power Syst. **11**, 83–92 (1996)
20. Lee, F.N.: Short-term unit commitment-a new method. IEEE Trans. Power Syst. **3**(2), 421–428 (1998)
21. Lee, H.W.J., Teo, K.L., Rehbock, V., Jennings, L.S.: Control parameterization enhancing technique for time optimal control problems. Dyn. Syst. Appl. **6**, 243–262 (1997)
22. Lee, H.W.J., Teo, K.L., Cai, X.Q.: An optimal control approach to nonlinear mixed integer programming problems. Comput. Math. Appl. **36**, 87–105 (1998)
23. Lee, H.W.J., Teo, K.L., Rehbock, V., Jennings, L.S.: Control parametrization enhancing technique for optimal discrete valued control problems. Automatica **35**, 1401–1407 (1999)
24. Mantawy, A.H., Abdel-Magid, Y.L., Selim, S.Z.: Unit commitment by tabu search. IEE Proc. Gener. Trans. Distr. **145**(1), 56–64 (1998)
25. Mantawy, A.H., Abdel-Magid Youssef, L., Selim Shokri, Z.: A simulated annealing algorithm for unit commitment. IEEE Trans. Power Syst. **13**(1), 197–204 (1998)
26. Padhy, N.P.: Unit commitment using hybrid models: a comparative study for dynamic programming, expert systems, fuzzy system and genetic algorithm. Int. J. Electr. Power Energy Syst. **23**(8), 827–836 (2001)
27. Rajan, C.C.A., Mohan, M.R.: An evolutionary programming-based tabu search method for solving the unit commitment problem. IEEE Trans. Power Syst. **19**(1), 577–585 (2004)
28. Rebennack, S., Pardalos, P.M., Pereira, M.V.F., Iliadis, N.A.: Handbook of Power Systems I. Energy Systems. Springer, Berlin (2010)
29. Rebennack, S., Pardalos, P.M., Pereira, M.V.F., Iliadis, N.A.: Handbook of Power Systems II. Energy Systems. Springer, Berlin (2010)
30. Rebennack, S., Flach, B., Pereira, M.V.P., Pardalos, P.M.: Stochastic hydro-thermal scheduling under CO2 emission constraints. IEEE Trans. Power Syst. **27**(1), 58–68 (2012)
31. Roque, L., Fontes, D.B.M.M., Fontes, F.A.C.C.: A biased random key genetic algorithm approach for unit commitment problem. Lect. Notes Comput. Sci. **6630**(1), 327–339 (2011)
32. Saravanan, B., Siddharth, D.A.S., Sikri, S., Kothari, D.P.: A solution to the unit commitment problem: a review. Front. Energ. **7**(2), 223–236 (2013)
33. Siburian, A.: Numerical methods for robust, singular and discrete valued optimal control problems. Ph.D. thesis, Curtin University of Technology, Perth, Australia (2004)
34. Simopoulos, D.N., Kavatza, S.D., Vournas, C.D.: Unit commitment by an enhanced simulated annealing algorithm. IEEE Trans. Power Syst. **21**(1), 68–76 (2006)
35. Swarup, K.S., Yamashiro, S.: Unit commitment solution methodology using genetic algorithm. IEEE Trans. Power Syst. **17**, 87–91 (2002)
36. Takriti, S., Birge, J.R.: Using integer programming to refine lagrangian-based unit commitment solutions. IEEE Trans. Power Syst. **15**(1), 151–156 (2000)
37. Teo, K.L., Jennings, L.S., Rehbock, V.: The control parameterization enhancing transform for constrained optimal control problems. J. Aust. Math. Soc. **40**, 314–335 (1999)

38. Valenzuela, J., Smith, A.E.: A seeded memetic algorithm for large unit commitment problems. J. Heuristics **8**(2), 173–195 (2002)
39. Viana, A., Pedroso, J.P.: A new MILP-based approach for unit commitment in power production planning. Electr. Power Energy Syst. **44**(1), 997–1005 (2013)
40. Victoire, T.A.A., Jeyakumar, A.E.: A tabu search based hybrid optimization approach for a fuzzy modelled unit commitment problem. Electr. Power Syst. Res. **76**, 413–425 (2006)
41. Zhao, B., Guo, C.X., Bai, B.R., Cao, Y.J.: An improved particle swarm optimization algorithm for unit commitment. Int. J. Electr. Power Energy Syst. **28**(7), 482–490 (2006)
42. Zheng, Q.P., Wang, J., Pardalos, P.M., Guan, Y.: A decomposition approach to the two-stage stochastic unit commitment problem. Ann. Oper. Res. **210**(1), 387–410 (2013)

On the Far from Most String Problem, One of the Hardest String Selection Problems

Daniele Ferone, Paola Festa, and Mauricio G.C. Resende

Abstract This paper describes the *far from most string problem*, one of the computationally hardest string selection problems that has found its way into numerous practical applications, especially in computational biology and bioinformatics, where one is interested in computing distance/proximity among biological sequences, creating diagnostic probes for bacterial infections, and/or discovering potential drug targets.

With special emphasis on the optimization and operational research perspective, this paper studies the intrinsic properties of the problem and overviews the most popular solution techniques, including some recently proposed heuristic and metaheuristic approaches. Future directions are discussed in the last section.

Keywords Computational biology • Molecular structure prediction • String selection • Consensus • Combinatorial optimization • Metaheuristics

1 Introduction

The *far from most string problem* (FFMSP) is one of the *string selection* and *comparison problems*, also called *sequence consensus problems*.

Generally speaking, given a finite set of sequences, one is interested in finding their *consensus*, i.e., a new sequence that "agrees" as much as possible with all the given sequences. In other words, the objective is to determine a sequence called consensus, because it represents in some way all the given sequences.

D. Ferone • P. Festa (✉)
Department of Mathematics and Applications, University of Napoli FEDERICO II, Compl. MSA, Via Cintia, 80126 Napoli, Italy
e-mail: danieleferone@gmail.com; paola.festa@unina.it

M.G.C. Resende
AT&T Labs Research, Florham Park, NJ, USA
e-mail: mgcr@research.att.com

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_7

The concept of being *representative* of a given set of sequences strongly depends on the specific objective pursued by the project of studying the information contained in the given sequences and their specific properties under experimentation. The most common objectives are listed in the following:

(i) the consensus is a new sequence whose total distance from all given sequences is minimum (*closest string problem*);
(ii) the consensus is a new sequence whose total distance from all given sequences is maximum (*farthest string problem*);
(iii) the consensus is a new sequence far from most of the given sequences (*FFMSP*).

String selection problems find thousands of applications in several and heterogenous fields, ranging from coding theory to molecular biology. A fundamental remark made by researchers in molecular biology regards the abstraction of the real three-dimensional structure of DNA and its representation as a unidimensional sequence of characters from an alphabet of four symbols. The same type of assumption involves also the protein represented as a sequence of characters from an alphabet of 20 symbols. As a result of the linear coding of DNA and proteins, many molecular biology problems have been formulated as computational and optimization problems involving strings and sequences, such as to rebuild long DNA sequences starting from overlapping fragments (fragment assembly), to compare two or more sequences looking for their similarities (strings coding the same function), and to look for patterns that occur with a certain frequency in DNA and/or protein sequences. Useless to say that many further targets can be pursued in molecular biology applications involving sequences. For example, another possible application arises in creating diagnostic probes for bacterial infections. Given a set of DNA sequences from a group of closely related pathogenic bacteria, the task is to find a substring that occurs in each of the bacterial sequences (as close as possible) without occurring in the host's DNA. Probes are then designed to hybridize to these target sequences, so that the detection of their presence indicates that at least one bacterial species is likely to be present in the host. Another biological application related to string selection and comparison problems is related to discovering potential drug targets. Given a set of sequences of orthologous genes from a group of closely related pathogens and a host (such as human, crop, or livestock), the goal is to find a sequence fragment that is more conserved in all or most of the pathogens' sequences but not as conserved in the host. Information encoded by this fragment can then be used for novel antibiotic development or to create a drug that harms several pathogens with minimal effect on the host. All these applications reduce to the task of finding a pattern that with some error occurs in one set of strings (closest string problem) and/or does not occur in another set (farthest string problem). The FFMSP can help to identify a sequence fragment that distinguishes the pathogens from the host, so the potential exists to create a drug that harms several but not all pathogens.

The remainder of this article is organized as follows. In Sect. 2, the FFMSP is described and its properties are analyzed. The most popular solution techniques

for this problem are surveyed in Sect. 3, along with the computational results obtained and analyzed in the literature. Concluding remarks and future directions are discussed in the last section.

2 Notation and Problem Description

Throughout this paper, the following notation and definitions will be used:

- An *alphabet* $\Sigma = \{c_1, c_2, \ldots, c_k\}$ is a finite set of elements, called *characters*.
- $s^i = (s^i_1, s^i_2, \ldots, s^i_m)$ is a sequence of length m ($|s^i| = m$) on Σ ($s^i_j \in \Sigma,\ j = 1, 2, \ldots, m$).
- Given two sequences s^i and s^l on Σ such that $|s^i| = |s^l|$, $d_H(s^i, s^l)$ denotes their Hamming distance and is given by

$$d_H(s^i, s^l) = \sum_{j=1}^{|s^i|} \Phi(s^i_j, s^l_j), \tag{1}$$

  where s^i_j and s^l_j are the characters in position j in s^i and s^l, respectively, and $\Phi : \Sigma \times \Sigma \to \{0, 1\}$ is the predicate function such that

$$\Phi(a, b) = \begin{cases} 0, \text{ if } a = b; \\ 1, \text{ otherwise.} \end{cases}$$

- Given a set of sequences $\Omega = \{s^1, s^2, \ldots, s^n\}$ on Σ ($s^i \in \Sigma^m, i = 1, 2, \ldots, n$) d^{Ω}_H denotes the Hamming distance among the sequences in Ω and it is given by

$$0 \leq d^{\Omega}_H = \min_{i,l=1,\ldots,n \mid i<l} d_H(s^i, s^l) \leq m. \tag{2}$$

Given a set of sequences $\Omega = \{s^1, s^2, \ldots, s^n\}$ on Σ ($s^i \in \Sigma^m, i = 1, 2, \ldots, n$), the FFMSP consists in determining a string far from most of the strings in Ω. This can be formally stated by saying that given a threshold t, a string $s \in \Sigma^m$ must be found maximizing the variable x such that

$$d_H(s, s^i) \geq t,\ \forall\, s^i \in P \subseteq \Omega \text{ and } |P| = x, \tag{3}$$

or, equivalently

$$d^{P \cup \{s\}}_H \geq t, \text{ for } P \subseteq \Omega \text{ and } |P| = x. \tag{4}$$

For most consensus problems, Hamming distance (1) is used instead of any alternative measure (such as the editing distance) and biological reasons justifying this choice are very well described and motivated by Lanctot et al. in [24].

The FFMSP is one of the computationally hardest sequence consensus problems. The intractability of the general sequence consensus problem was proved in 1997 by Frances and Litman [16] and in 1999 by Sim and Park [31]. In 2003, Lanctot et al. [25] demonstrated that for sequences over an alphabet Σ with $|\Sigma| \geq 3$, approximating the FFMSP within a polynomial factor is NP-hard.

3 Several Alternative State-of-the-Art Algorithms for the FFMSP

Since polynomial time algorithms for approaching the FFMSP can yield only solutions with no constant guarantee of approximation, heuristic methods must be devised to find good-quality solutions in reasonable running times. This section overviews main heuristic/metaheuristic algorithms to efficiently find good suboptimal solutions for the FFMSP, starting from the first attempt done in 2005 by Meneses et al. [27] to the latter proposed techniques in 2013 by Ferone et al. [5, 6], who designed several pure and hybrid metaheuristics.

3.1 A Simple Heuristic Approach

The first attempt in trying to obtain good approximate solutions in reasonable running times has been done by Meneses et al. [27], who in 2005 proposed a simple heuristic consisting of the following two phases:

Phase I: A *construction phase* that iteratively builds a feasible solution $s \in \Sigma^m$. Initially,

- for each position $j \in \{1, \ldots, m\}$, compute the set V_j of characters appearing in that position in any of the n strings in Ω;
- for each character $c \in V_j$, compute the number of times that c appears in the input on position j.

Then, for each position $j \in \{1, \ldots, m\}$, a string s is iteratively built by choosing the character in V_j that appears in the smallest number of strings.
For each position $j > 1$, check the effect that assigning a character to this position will have on previous assignments.

Phase II: A *local search phase* that starting from s explores a suitably defined *neighborhood* of s (a set of feasible solutions "close" to s) until a local optimum is found and returned as final solution.

```
algorithm iter-impr (c, s, 𝒩)
1   NoLocal := true;
2   while (NoLocal) do
3         /* exploration of the neighborhood N(s) */
4         if (∃ s̄ ∈ N(s) : c's̄ > c's, c's̄ ≥ c'y, ∀ y ∈ N(s)) then
5             s := s̄;
6         else NoLocal := false;
7         endif;
8   endwhile;
9   return (s);
end iter-impr
```

Fig. 1 Local search procedure `iter-impr` for a maximization problem

Figure 1 shows the pseudo-code of the simplest local search procedure, known as *iterative improvement*. It takes as input a cost vector c, an initial solution s, and a predefined neighborhood function N. Starting from s, the iterative improvement procedure explores the neighborhood $N(s)$ looking for a better solution $\overline{s}$ in terms of objective function value. If such a solution exists, then the search continues from $\overline{s}$; otherwise, the procedure provides as output the current solution s which is locally optimal with respect to the defined neighborhood.

Meneses et al. [27] proposed a 2-*exchange* local search procedure, whose basic step consists in randomly selecting a position $j \in \{1, \ldots, m\}$ and changing it to another character in V_j selected at random.

3.2 A GRASP

Meneses et al.'s algorithm has been the first attempt in the design of heuristic approaches for the FFMSP. It is basically a greedy construction of a feasible solution followed by a local search procedure to generate a local optimal solution.

The second step along this research line has been to design a *multi-start* or *iterative process* as conceived in [26]. In general, in a multi-start technique, a solution is built either only once and usually at the beginning of the solution process or it is built at each iteration or "start" of the algorithm. In the first case, the unique solution built by the approach can be constructed applying any criterion, ranging from a pure greedy to a pure random strategy. Conversely, when a solution is built at each start of the algorithm, then a pure greedy strategy should not be followed since each time the pure greedy construction is invoked it would lead to the same solution or to a set of "close" solutions.

```
algorithm GRASP (f(·), 𝒩, Seed)
1  s_best:=∅; f(s_best):=−∞;
2  while stopping criterion not satisfied→
3        s:=GreedyRandomAdapBuild(Seed);
4        s:=LocalSearch(s, 𝒩, f(·));
5        if (f(s) > f(s_best)) then
6           s_best:=s;
7        endif
8  endwhile
9  return (s_best);
end GRASP
```

Fig. 2 Pseudo-code of a generic GRASP for a maximization problem

The first multi-start iterative process designed for the FFMSP appeared in 2007 in [7], where a GRASP (greedy randomized adaptive search procedure) and a genetic algorithm have been proposed.

Originally proposed in the literature by Feo and Resende [3,4], in the last 20 years GRASP has been successfully applied to several computationally intractable combinatorial problems. The reader interested in a comprehensive study of GRASP strategies and variants is referred to the survey chapter by Resende and Ribeiro [29] and the more recent articles by Festa and Resende [11–13], as well as to the annotated bibliography of Festa and Resende [8–10].

Figure 2 depicts the pseudo-code of a generic GRASP heuristic for a maximization problem. GRASP is an iterative multi-start heuristic algorithm that for a certain number of iterations realizes two phases (loop while in lines 2–8): a construction phase (line 3) and a local search phase (line 4). Similarly to the semi-greedy heuristic proposed independently by [22], the basic GRASP construction phase starts from an empty solution and iteratively adds one element at a time to the partial solution under construction, ending up with a representation of a feasible solution. At each iteration, an element is randomly selected from a *restricted candidate list* (RCL), whose elements are among the best ordered, according to some greedy function that measures the (myopic) benefit of selecting each element. Once a feasible solution is obtained, the local search procedure attempts to improve it by producing a locally optimal solution with respect to a predefined neighborhood structure. Construction and local search phases are repeatedly applied until stopping criterion is met and the best local optimal solution found over all iterations is returned as output.

A GRASP construction phase generally makes use of an adaptive greedy function for constructing the RCL and of a probabilistic selection criterion of a well-ranked element from the restricted list. In the case of the FFMSP, it is intuitive to relate the greedy function to the occurrence of each character in a given position. In fact, as in [27], for each position $j \in \{1, \ldots, m\}$, it is computed as the set V_j of characters appearing in that position in any of the strings in Ω and then for each character $c \in V_j$, $g_j(c)$ is computed as the number of times that c appears in the input on position j. Starting from an empty solution, at each construction iteration the choice of the next element to be added to the partial solution is determined by ordering all candidate characters in a candidate list C with respect to the above defined greedy function. The probabilistic component of the GRASP here proposed is characterized by *randomly* choosing one of the best candidates in the list, but not necessarily the top candidate. As in any GRASP heuristic, the construction procedure is *adaptive*, because the benefits associated with every element are updated at each iteration of the construction phase to reflect the changes brought on by the selection of the previous element.

There are several different mechanisms to build the RCL. Typically, it can be limited by the number of elements (cardinality-based criterion) or by their quality (value-based criterion). If the cardinality-based criterion is chosen, then the cardinality of RCL is a priori fixed to some p and the RCL is made up of those elements having the p best greedy function values, while in the value-based case, the cardinality of RCL depends on a threshold parameter $0 \leq \alpha \leq 1$. In the GRASP proposed in [7], at each iteration $j \in \{1, \ldots, m\}$ of the construction procedure, RCL is formed by all possible candidates y whose greedy function value $g_j(y)$ is better or equal to $\alpha \cdot g_j^*$, where g_j^* is the best greedy function value. Note that the extreme case $\alpha = 0$ corresponds to a pure greedy strategy, while the extreme case $\alpha = 1$ is equivalent to a completely random strategy.

To realize the local search phase the 2-exchange procedure is used as in [27]. The procedure takes as input the solution $s = (s_1, \ldots, s_m) \in \Sigma^m$ built at the end of the GRASP construction phase and $\{V_j\}_{j=1,\ldots,m}$, where V_j is the set of characters appearing in position j in any of the strings in Ω. Then, for each position $j = 1, \ldots, m$ and for each character $c \in V_j$, $c \neq s_j$, the 2-exchange procedure checks if the solution $\overline{s} = (s_1, \ldots, s_{j-1}, c, s_{j+1}, \ldots, s_m)$ obtained from s exchanging the character in position j is better than s in terms of objective function value. Note that $\overline{s}$ is a neighbor of solution s such that $d_H(s, \overline{s}) = 1$.

A local search may be implemented using either a *best improving* or a *first improving strategy*. The first improving strategy stops the current iteration as soon as an improving neighbor is found, while in the best improving strategy all neighbors are evaluated and the best among them is kept as new current solution from which to start the next iteration. In [7], both strategies have been implemented and, accordingly with results from the literature about these different possible strategies, the best improving produces better-quality solutions for most of the problem instances but in a higher amount of time as compared to the first improving strategy.

As any multi-start iterative heuristic, stopping criteria in a GRASP could be maximum number of iterations, maximum number of iterations without improvement of the incumbent solution, maximum running time, or solution quality at least as good as a given target value. In the GRASP for the FFMSP proposed in [7], the adopted stopping criterion has been a maximum number of iterations.

3.3 A New Solution Evaluation Function

Starting from a given solution to the problem under studying, any local search procedure explores a suitable neighborhood set of solutions "close" to the starting solution and outputs a locally optimal solution with respect to the used neighborhood structure definition. In exploring the neighborhood set, a local search needs to evaluate many candidate solutions in order to compare them. The most used function to perform this evaluation is the objective function, so that a solution s is better than a different solution $\overline{s}$ if and only if the objective function evaluated in s produces a value strictly better than the value assumed by the objective function when evaluated in $\overline{s}$.

This way of investigating the neighborhood of a solution is basically a *steepest descend/ascend* process that presents serious drawbacks when the search landscape includes many local optimal solutions. Unfortunately, this is exactly the case of the FFMSP, because the set of possible objective function values is $\{0, 1, \ldots, n\}$ and is therefore rather small. To overcome this limit, Mousavi et al. [28] in 2012 proposed to use in the GRASP local search of Festa [7] an alternative solution evaluation function that takes into account both the classical objective function and the so-called *estimated Gain-per-Cost* heuristic function that expresses the likelihood of a solution to lead to better solutions with as few changes as possible.

The authors compared the original Festa's GRASP with their proposal on both randomly generated and real-world problem instances and as result of their experiments the variant of the GRASP they proposed overcomes in terms of solutions quality the original GRASP in all cases.

3.4 Hybrid and Pure Metaheuristics

An issue from the heuristic/metaheuristic research community involves the idea of combining the main characteristics of pure metaheuristic frameworks in the attempt to take advantage of their best properties in terms of total running times and/or solution quality. Following this recent trend, Ferone et al. in [5] have designed the following pure and hybrid metaheuristics for finding good-quality solutions to the FFMSP:

- a pure GRASP, inspired by [7];
- a GRASP that uses Path-relinking for intensification;

algorithm GRASP $(m, \Sigma, f(\cdot), \{V_j(c)\}_{j\in\{1,\ldots,m\}}^{c\in\Sigma}$, Seed)
1 s_{best}:=$\emptyset$; $f(s_{best})$:=$-\infty$;
2 **for** $j = 1$ **to** $m\rightarrow$
3 $V_j^{\min}$:=$\min_{c\in\Sigma} V_j(c)$; $V_j^{\max}$:=$\max_{c\in\Sigma} V_j(c)$;
4 **endfor**
5 **while** stopping criterion not satisfied$\rightarrow$
6 $[s, \{\mathrm{RCL}_j\}_{j=1}^{m}]$:=GrRand $(m, \Sigma, \{V_j(c)\}_{j\in\{1,\ldots,m\}}^{c\in\Sigma}, V_j^{\min}, V_j^{\max}$, Seed);
7 s:=LocalSearch $(m, s, f(\cdot), \{\mathrm{RCL}_j\}_{j=1}^{m})$;
8 **if** $(f(s) > f(s_{best}))$ **then**
9 s_{best}:=s;
10 **endif**
11 **endwhile**
12 **return** (s_{best});
end GRASP

Fig. 3 Pseudo-code of a GRASP for the FFMSP

- a pure variable neighborhood search (VNS);
- a VNS that uses Path-relinking for intensification;
- a GRASP that uses VNS to implement the local search phase; and
- a GRASP that uses VNS to implement the local search phase and Path-relinking for intensification.

3.4.1 A Pure GRASP

The pure GRASP proposed in [5] has been inspired by the GRASP proposed in 2007 [7], but presents some different details in the design of its main ingredients. Figure 3 depicts its pseudo-code, where $f : \Sigma^m \mapsto \mathbb{N}$ denotes the objective function to be maximized according to (3) and (4).

As any GRASP framework and as the GRASP proposed in [7], also the GRASP proposed in [5] for the FFMSP proceeds for a certain number of iterations. At each iteration, it builds a solution sequence s, starting from which to look for a locally optimal solution with respect to a predefined neighborhood structure. In the following, the main ingredients of both construction and local search procedures are detailed.

The operations performed during the construction phase are described in Fig. 4, where a sequence $s = (s_1, \ldots, s_m) \in \Sigma^m$ is iteratively built, one character at each iteration. As described in [7], the greedy function is related to the occurrence of each character in a given position. In more detail, for each position $j \in \{1, \ldots, m\}$ and for each character $c \in \Sigma$, the number $V_j(c)$ of times c appears in position j in any of the strings in Ω is computed. Then, to build the RCL, let

```
function GrRand (m, Σ, {V_j(c)}_{j∈{1,...,m}}^{c∈Σ}, V_j^min, V_j^max, Seed)
1  for j = 1 to m→
2        RCL_j:=∅; α:=Random([0,1], Seed);
3        μ:=V_j^min + α·(V_j^max − V_j^min);
4        for all c ∈ Σ→
5            if (V_j(c) ≤ μ)) then
6                RCL_j:=RCL_j∪{c};
7            endif
8        endfor
9        s_j:=Random (RCL_j, Seed);
10 endfor
11 return (s, {RCL_j}_{j=1}^m);
end GrRand
```

Fig. 4 Pseudo-code of the GRASP construction for the FFMSP

$$V_j^{\min} = \min_{c \in \Sigma} V_j(c), \qquad V_j^{\max} = \max_{c \in \Sigma} V_j(c).$$

Denoting by $\mu = V_j^{\min} + \alpha \cdot (V_j^{\max} - V_j^{\min})$ the cutoff value (line 3), where α is a parameter such that $0 \leq \alpha \leq 1$ (line 2), the RCL is built by selecting as its members all characters whose greedy function value is not greater than μ (line 6). Once the RCL is built, character is then randomly selected from it (line 9) (Fig. 4).

The GRASP local search proposed in [5] and presented in Fig. 5 analyzes all positions $j \in \{1, \ldots, m\}$ (loop in lines 4–14) and changes the character in position j in the sequence s to another character in RCL_j. During the local search process, the current solution is replaced by the first improving neighbor (lines 8–11). The search in the neighborhood of the current solution stops after all possible moves have been evaluated and no improving neighbor of the current solution was found, returning a local optimal solution (line 16).

3.4.2 A Pure VNS

In [5], a pure VNS has been designed for the FFMSP. Contrary to other metaheuristics based on local search methods, given a solution s, VNS [21] is based on the exploration of increasingly distant neighborhoods $N_k(s)$, $k = 1, \ldots, k_{\max}$, until some stopping condition is satisfied. Figure 6 reports the pseudo-code of a VNS for the FFMSP, as proposed in [5].

At any VNS iteration a solution S is built at random (line 3). Then, in loop lines 4–13, increasingly distant neighborhoods $N_k(s)$, $k = 1, \ldots, k_{\max}$, are explored by applying the same local search strategy used within the pure GRASP

```
function LocalSearch (m, s, f(·), {RCL_j}_{j=1}^{m})
1   max:=f(s); change:=.TRUE.;
2   while (change)→
3         change:=.FALSE.;
4         for j = 1 to m→
5             temp:=sj;
6             for all c ∈RCLj→
7                 sj:=c;
8                 if ( f(s) > max) then
9                     max:=f(s);temp:=c;change:=.TRUE.;break;
10                endif
11            endfor
12            sj:=temp;
13        endfor
14  endwhile
15   return(s);
end LocalSearch
```

Fig. 5 Pseudo-code of the GRASP local search for the FFMSP

```
algorithm VNS (m, Σ , f(·), k_max, Seed)
1   s_best :=∅; f(s_best):=−∞;
2   while stopping criterion not satisfied→
3         k:=1; s:=BuildRand(m,Σ, Seed); /* pure randomly */
4         while (k ≤ k_max)→
5             s′:=Random (N_k(s), Seed);
6             s″:=locsearch (m, s′, f(·), {RCL_j}_{j=1}^{m});
7             if (f(s″) > f(s)) then
8                 s:=s″; k:=1;
9                 if (f(s″) > f(s_best)) then s_best:=s″;
10                endif
11            else k:=k+1;
12            endif
13        endwhile
14  endwhile
15  return (s_best);
end VNS
```

Fig. 6 Pseudo-code of a VNS for the FFMSP

algorithm described in Sect. 3.4.1. For the FFMSP, the kth-order neighborhood $N_k(s)$ is the set of all sequences that can be derived from the current sequence s by selecting k positions $j_1, \ldots, j_k$ and changing $s_{j_1}, \ldots, s_{j_k}$ with a character in $\mathrm{RCL}_{j_1}, \ldots, \mathrm{RCL}_{j_k}$, respectively.

3.4.3 Path-Relinking

Path-relinking is an intensification strategy exploring trajectories connecting elite solutions. It was proposed in 1996 by Glover [17] and later hybridized with tabu search and scatter search [18–20]. Generally speaking, starting from one elite solution, Path-relinking generates a path in the solution space leading towards another guiding elite solution. This path connecting the two solutions in the solution space is built by selecting moves that introduce in the initial solution attributes contained in the guiding solution. At each iteration, all moves are analyzed and the move that best improves (or least deteriorates) the initial solution is chosen. The scope of building this path is to explore it in the search for better solutions.

In [5], Path-relinking is applied to a pair of sequences $(\mathbf{s}, \hat{\mathbf{s}})$, where $\mathbf{s}$ is a given input solution and $\hat{\mathbf{s}}$ is a solution *sufficiently different* from $\mathbf{s}$ selected at random (line 1) from an elite set $\mathscr{E}$ of solutions that has a fixed size that does not exceed `MaxElite`.

$\mathbf{s}$ and $\hat{\mathbf{s}}$ have been retained sufficiently different if $|\Delta(\mathbf{s}, \hat{\mathbf{s}})| \geq \frac{m}{2}$, where $\Delta(\mathbf{s}, \hat{\mathbf{s}})$ is their the symmetric difference set and it is clearly defined as follows:

$$\Delta(\mathbf{s}, \hat{\mathbf{s}}) := \{i = 1, \ldots, m \mid \mathbf{s}_i \neq \hat{\mathbf{s}}_i\}. \tag{5}$$

Roughly speaking, $\Delta(\mathbf{s}, \hat{\mathbf{s}})$ is the set of components for which the two solutions differ. Once known and selected the sequences $\mathbf{s}$ and $\hat{\mathbf{s}}$, a path is explored in the solution space linking the worst solution $\mathbf{s}'$ between $\mathbf{s}$ and $\hat{\mathbf{s}}$ to the best one (line 3). $\mathbf{s}'$ is called the *initial solution* and $\hat{\mathbf{s}}$ the *guiding solution* (Fig. 7).

The procedure then computes (line 4) the symmetric difference $\Delta(\mathbf{s}', \hat{\mathbf{s}})$ between the two solutions as defined in Eq. (5), i.e., the set of moves needed to reach $\hat{\mathbf{s}}$ from $\mathbf{s}'$. A path of solutions $\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_{|\Delta(\mathbf{s}', \hat{\mathbf{s}})|}$ is generated linking $\mathbf{s}'$ and $\hat{\mathbf{s}}$. The best solution s^* in this path is returned by the algorithm (line 13).

The path of solutions is computed in the loop in lines 5 through 12. This is achieved by advancing one solution at a time in a greedy manner. At each iteration, all possible moves $i \in \Delta(\mathbf{s}', \hat{\mathbf{s}})$ are examined. Then, the best move i^* is made, producing solution $\mathbf{s}' \oplus i^*$ (line 8) and, if necessary, the best solution $\mathbf{s}^*$ is updated (lines 9–11). The algorithm stops as soon as $\Delta(\mathbf{s}', \hat{\mathbf{s}}) = \emptyset$.

3.4.4 Hybrid GRASP with Path-Relinking

Hybridizations of GRASP with Path-relinking have been proposed in the literature to incorporate into a general GRASP framework some sort of *memory mechanisms*.

```
algorithm Path-relinking(m, f(·),s,ℰ,Seed)
1   ŝ:=Random(ℰ,Seed);
2   f* := max{f(s), f(ŝ)}; s* := argmax{f(s), f(ŝ)};
3   s′:= argmin{f(s), f(ŝ)}; ŝ:=s*;
4   Δ(s′, ŝ):=:={i=1,...,m |s′_i≠ŝ_i};
5   while (Δ(s′, ŝ)≠∅) →
6         i* := argmax{f(s′ ⊕ i) | i ∈ Δ(s′,ŝ)};
7         Δ(s′ ⊕ i*,ŝ) := Δ(s′,ŝ) \ {i*};
8         s′ := s′ ⊕ i*;
9         if(f(s′) > f*) then
10           f* := f(s′); s* := s′;
11        endif
12  endwhile
13  return (s*);
end Path-relinking
```

Fig. 7 Pseudo-code of a path-relinking for the FFMSP

The first attempt along this direction has been done by Laguna and Martí [23] in 1999. This attempt has been successful followed by several further extensions and improvements [2, 11, 13–15].

In [5], the authors have integrated Path-relinking into the pure GRASP algorithm described in Sect. 3.4.1. In more detail, at each GRASP iteration, Path-relinking is applied to a pair $(\mathbf{s}, \hat{\mathbf{s}})$ of solutions, where $\mathbf{s}$ is the locally optimal solution obtained by GRASP local search and $\hat{\mathbf{s}}$ is randomly chosen from a pool with a limited number `MaxElite` of high-quality solutions found along the search. The pseudo-code for the proposed GRASP with Path-relinking hybrid algorithm is shown in Fig. 8.

The pool $\mathscr{E}$ of elite solutions is originally empty (line 1) and it is then populated by the first `MaxElite` GRASP iteration local optima solutions. After `MaxElite` iteration, the best solution $\overline{s}$ found along the relinking trajectory is considered as a candidate to be inserted into $\mathscr{E}$. In more detail, the procedure `AddToElite` inserts $\overline{s}$ into the pool replacing the worst elite solution if $\overline{s}$ is better than the best elite solution or if it is better than the worst elite solution and *sufficiently different* (i.e., if $|\Delta(\overline{s}, \epsilon)| \geq \frac{m}{2}$, for all $\epsilon \in \mathscr{E}$) from all elite solutions.

3.4.5 Hybrid GRASP with VNS

As underlined in Sect. 3.4.2, until a stopping criterion is met, VNS generates at each iteration a sequence s at random. In the hybrid GRASP with VNS designed in [5], VNS is applied as local search and its starting solution is the sequence s output of the GRASP construction procedure.

```
algorithm GRASP+PR (m, Σ, f(·), {V_j(c)}^{c∈Σ}_{j∈{1,...,m}}, Seed, MaxElite)
1  s_best:=∅; f(s_best):=−∞; ℰ := ∅; iter:=0;
2  for j = 1 to m→
3        V_j^min:=min_{c∈Σ} V_j(c); V_j^max:=max_{c∈Σ} V_j(c);
4  endfor
5  while stopping criterion not satisfied→
6        iter:=iter + 1;
7        [s, {RCL_j}^m_{j=1}]:=GrRand (m, Σ, {V_j(c)}^{c∈Σ}_{j∈{1,...,m}}, V_j^min, V_j^max, Seed);
8        s:=LocalSearch (m, s, f(·), {RCL_j}^m_{j=1});
9        if (iter ≤ MaxElite) then
10           ℰ := ℰ ∪ {s};
11           if (f(s) > f(s_best)) then s_best:=s;
13           endif
14       else
10           s̄:=Path-relinking (m, f(·), s, ℰ, Seed);
15           AddToElite (ℰ,s̄);
11           if (f(s̄) > f(s_best)) then s_best:=s̄;
13           endif
13       endif
11 endwhile
12 return (s_best);
end GRASP+PR
```

Fig. 8 Pseudo-code of a hybrid GRASP with path-relinking for the FFMSP

3.4.6 Hybrid VNS with Path-Relinking

VNS described in Sect. 3.4.2 has been hybridized with Path-relinking applied as an intensification procedure at each VNS iteration. In the pair of solutions $(\mathbf{s}, \hat{\mathbf{s}})$ to which Path-relinking is applied, $\mathbf{s}$ is the locally optimal solution while $\hat{\mathbf{s}}$ is randomly chosen from the `MaxElite` high-quality solutions.

3.4.7 Hybrid GRASP with VNS and Path-Relinking

The main blocks of each iteration of this hybrid heuristic are the GRASP construction procedure, followed by VNS local search phase, and Path-relinking as intensification procedure. See [5] for a detailed description of the resulting hybrid algorithm.

3.4.8 Experimental Results and Recent Trends

In [5], the above described pure and hybrid metaheuristic approaches have been experimentally evaluated to determine which algorithm seems to be more effective to solve the FFMSP.

The objectives of the computational study were to compare the running times and the solution qualities achieved by the several alternative pure and hybrid algorithms when applied to solve FFMSP instances pseudo-randomly generated and characterized by several different sizes. In fact, in the set of test instances, the sequence length m ranged from 300 to 800, the number n of sequences in Ω ranged from 100 to 200, and threshold t varied from $75\,\%m$ to $85\,\%m$. The stopping criterion for all algorithms was $\texttt{MaxIterations} = 500$ or the obtainment of an incumbent solution with objective function value $z = n$ (i.e., an optimal solution).

For each problem size, 100 random instances have been generated for each possible value of $t \in \{75\,\%m, 80\,\%m, 85\,\%m\}$. Each algorithm has been run on the 100 random instances and average solution values and the corresponding average running times (in seconds) were computed. For a detailed analysis of the experimental evaluation of the different algorithms, the reader can refer to [5]. Here, about the experiments carried by the authors, we report only the main conclusions which are the following:

- on all instances, better-quality solutions have been found by the hybrid GRASP with VNS and Path-relinking;
- at the expense of increased running times, both hybridizations of Path-relinking with all different metaheuristics and the use of VNS in the local search phase of GRASP have been beneficial in terms of solution quality.

Given the random component of each proposed algorithm and since their running times per iteration vary substantially, in [5] a further experiment has been carried, involving the empirical distributions of the random variable *time-to-target-solution-value* considering the following four random instances:

1. $n = 100, m = 300, t = 240$, and target value $\hat{z} = 0.70 \times n$ (Fig. 9a);
2. $n = 100, m = 300, t = 252$, and target value $\hat{z} = 0.12 \times n$ (Fig. 9b);
3. $n = 200, m = 300, t = 240$, and target value $\hat{z} = 0.40 \times n$ (Fig. 10a);
4. $n = 300, m = 300, t = 240$, and target value $\hat{z} = 0.28 \times n$ (Fig. 10b).

100 independent runs of each heuristic have been performed and the time taken to find a solution at least as good as the target value $\hat{z}$ has been saved. As in [1], to plot the empirical distribution, with the ith sorted running time (t_i) a probability $p_i = \frac{i-1/2}{100}$ is associated, and the points $z_i = (t_i, p_i)$, for $i = 1, \ldots, 100$, are then plotted. About these further experiments, looking at Figs. 9 and 10, the authors concluded that, in a fixed amount of running time, the hybrid GRASP with Path-relinking has a higher probability than all competitors of finding a solution whose objective function value is at least as good as the target objective function value.

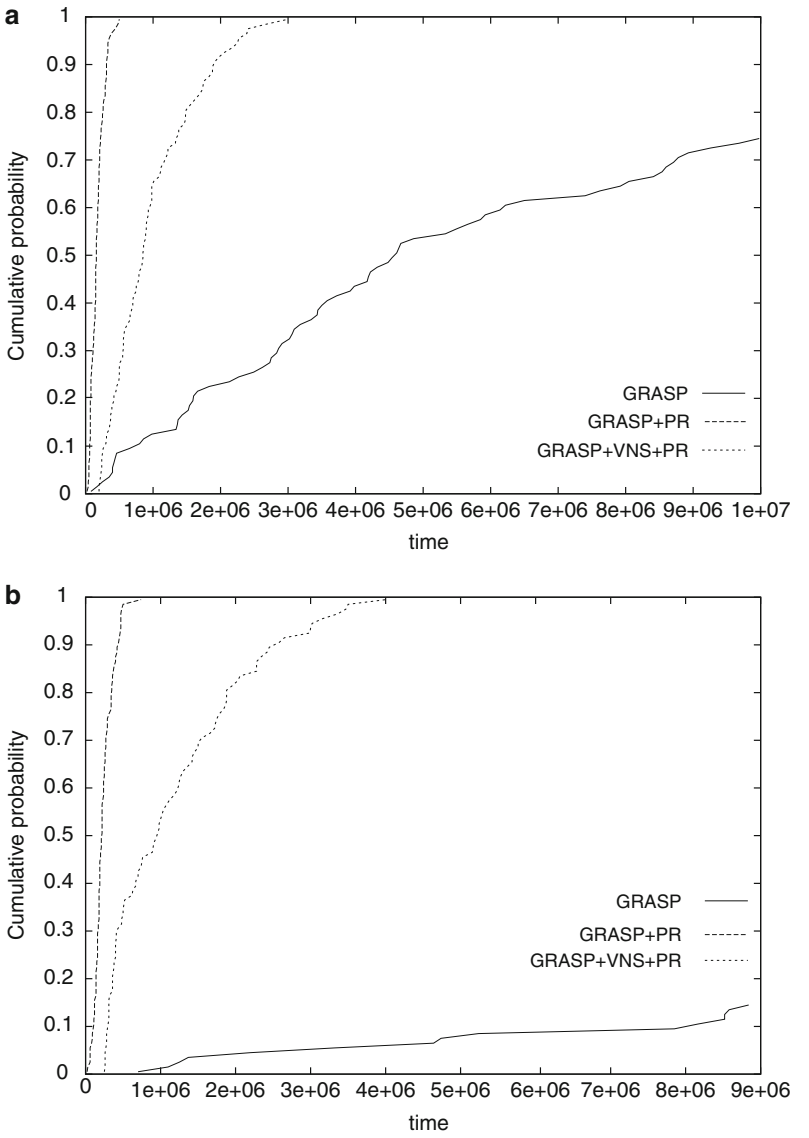

Fig. 9 Time to target distributions comparing `GRASP`, `GRASP+PR`, and `GRASP+VNS+PR` (**a**) Random instance with $n = 100$, $m = 300$, $t = 240$, and target value $\hat{z} = 0.70 \times n$ (**b**) Random instance with $n = 100$, $m = 300$, $t = 252$, and target value $\hat{z} = 0.12 \times n$

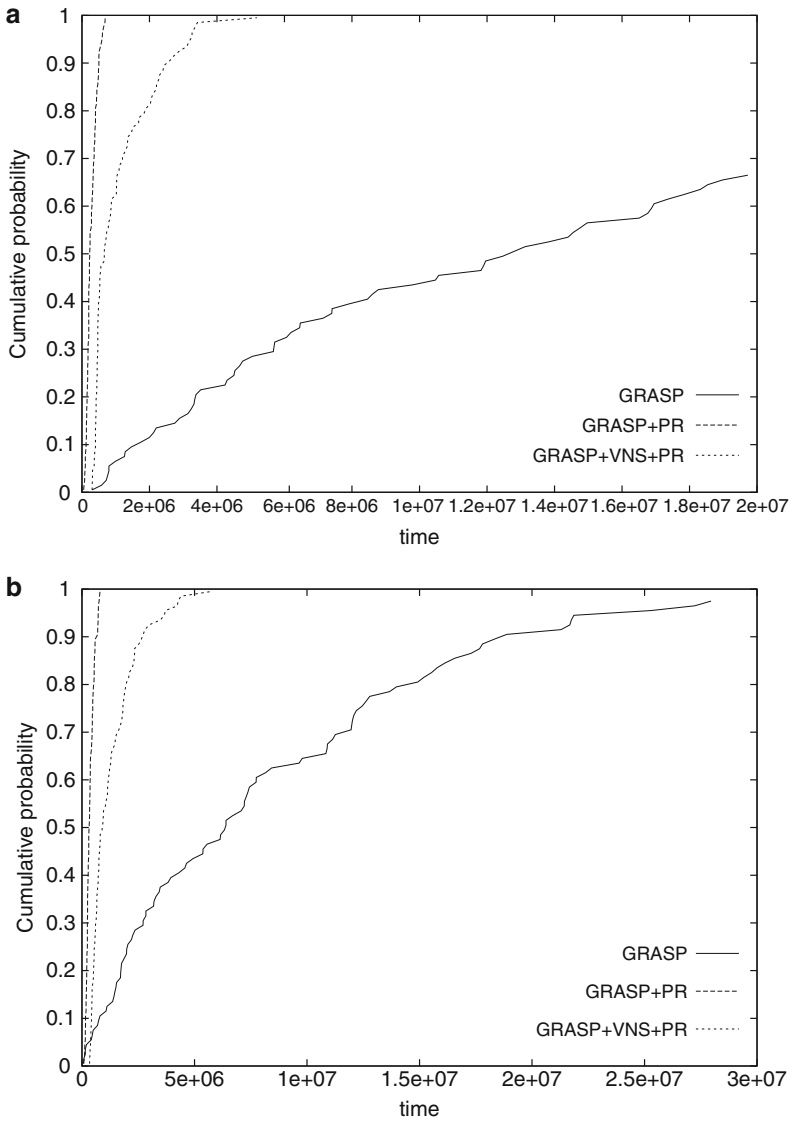

Fig. 10 Time to target distributions comparing `GRASP`, `GRASP+PR`, and `GRASP+VNS+PR` (**a**) Random instance with $n = 200$, $m = 300$, $t = 240$, and target value $\hat{z} = 0.40 \times n$ (**b**) Random instance with $n = 300$, $m = 300$, $t = 240$, and target value $\hat{z} = 0.28 \times n$

4 Conclusions and Future Directions

The *FFMSP*—as well as all consensus problems—occurs as a problem and/or as a subproblem in many real-world scenarios, especially in computational biology and bioinformatics.

Among consensus problems, the FFMSP is one of the computationally hardest. For this reason, contrary to other problems in the same family, only very recently efficient, fast, and robust solution techniques have been designed and proposed in the scientific literature. With special emphasis on the optimization and operational research perspective, this paper surveyed the most popular solution techniques, starting from the first ingenuous and naive heuristic attempt by Meneses et al. [27] in 2005 and ending with recently proposed heuristic and metaheuristic approaches [5, 28] that appeared in 2012 and 2013.

In [6], as current and immediately future work, we are performing the following steps:

1. To better understand the practical behavior of the algorithms proposed in [5], we are analyzing the numerical results by applying a recently published tool designed by Ribeiro et al. [30] for characterizing stochastic algorithms' running times under the assumption that the running times of the algorithms follow exponential (or shifted exponential) distributions, as it is the case of our hybrid heuristics.
2. We are collecting and interpreting further numerical results, by applying the heuristics proposed in [5] on a larger dataset of instances, both randomly generated and taken from real-world applications of the problem.
3. We are designing some further variants of the approaches proposed in [5]. Three natural extensions are

   - to perform a post-optimization phase, consisting in performing Path-relinking among pairs of elite solutions;
   - to implement alternative Path-relinking strategies, known as backward, mixed, and randomized Path-relinking;
   - to integrate in the local search of the algorithms Mousavi et al.'s function [28].

In addition, thinking about future research it would also be interesting to propose some further metaheuristic approaches, analyzing the possibility of designing some sophisticated nature-inspired techniques, whose main ingredients could be ad hoc tuned for this problem.

References

1. Aiex, R.M., Resende, M.G.C., Ribeiro, C.C.: Probability distribution of solution time in GRASP: an experimental investigation. J. Heuristics **8**, 343–373 (2002)
2. Canuto, S.A., Resende, M.G.C., Ribeiro, C.C.: Local search with perturbations for the prize-collecting Steiner tree problem in graphs. Networks **38**, 50–58 (2001)
3. Feo, T.A., Resende, M.G.C.: A probabilistic heuristic for a computationally difficult set covering problem. Oper. Res. Lett. **8**, 67–71 (1989)
4. Feo, T.A., Resende, M.G.C.: Greedy randomized adaptive search procedures. J. Global Optim. **6**, 109–133 (1995)
5. Ferone, D., Festa, P., Resende, M.G.C.: Hybrid metaheuristics for the far from most string problem. In: Proceedings of 8th International Workshop on Hybrid Metaheuristics, Lecture Notes in Computer Science, Springer, vol. 7919, pp. 174–188 (2013)
6. Ferone, D., Festa, P., Resende, M.G.C.: Hybrid metaheuristics for the far from most string problem. Technical Report, Department of Mathematics and Applications, University of Napoli FEDERICO II (2013)
7. Festa, P.: On some optimization problems in molecular biology. Math. Biosc. **207**(2), 219–234 (2007)
8. Festa, P., Resende, M.G.C.: GRASP: an annotated bibliography. In: Ribeiro, C.C., Hansen, P. (eds.) Essays and Surveys on Metaheuristics, pp. 325–367. Kluwer Academic Publishers, Norwell (2002)
9. Festa, P., Resende, M.G.C.: An annotated bibliography of GRASP: part I: algorithms. Int. Trans. Oper. Res. **16**(1), 1–24 (2009)
10. Festa, P., Resende, M.G.C.: An annotated bibliography of GRASP: part II: applications. Int. Trans. Oper. Res. **16**(2), 131–172 (2009)
11. Festa, P., Resende, M.G.C.: Hybrid GRASP heuristics. Stud. Comput. Intell. **203**, 75–100 (2009)
12. Festa, P., Resende, M.G.C.: GRASP: basic components and enhancements. Telecommun. Syst. **46**(3), 253–271 (2011)
13. Festa, P., Resende, M.G.C.: Hybridizations of GRASP with path-relinking. Stud. Comput. Intell. **434**, 135–155 (2013)
14. Festa, P., Pardalos, P.M., Resende, M.G.C., Ribeiro, C.C.: Randomized heuristics for the MAX-CUT problem. Optim. Meth. Software **7**, 1033–1058 (2002)
15. Festa, P., Pardalos, P.M., Pitsoulis, L.S., Resende, M.G.C.: GRASP with path-relinking for the weighted MAXSAT problem. ACM J. Exp. Algorithmics **11**, 1–16 (2006)
16. Frances, M., Litman, A.: On covering problems of codes. Theor. Comput. Syst. **30**(2), 113–119 (1997)
17. Glover, F.: Tabu search and adaptive memory programing: advances, applications and challenges. In: Barr, R.S., Helgason, R.V., Kennington, J.L. (eds.) Interfaces in Computer Science and Operations Research, pp. 1–75. Kluwer Academic Publishers, New York (1996)
18. Glover, F.: Multi-start and strategic oscillation methods: principles to exploit adaptive memory. In Laguna, M., Gonzáles-Velarde, J.L. (eds.) Computing Tools for Modeling, Optimization and Simulation: Interfaces in Computer Science and Operations Research, pp. 1–24. Kluwer Academic Publishers, Norwell (2000)
19. Glover, F., Laguna, M.: Tabu Search. Kluwer Academic Publishers, Boston (1997)
20. Glover, F., Laguna, M., Martí, R.: Fundamentals of scatter search and path relinking. Contr. Cybern. **39**, 653–684 (2000)
21. Hansen, P., Mladenović, N.: Developments of variable neighborhood search. In: Ribeiro, C.C., Hansen, P. (eds.) Essays and Surveys in Metaheuristics, pp. 415–439. Kluwer Academic Publishers, Norwell (2002)

22. Hart, J.P., Shogan, A.W.: Semi-greedy heuristics: an empirical study. Oper. Res. Lett. **6**, 107–114 (1987)
23. Laguna, M., Martí, R.: GRASP and path relinking for 2-layer straight line crossing minimization. INFORMS J. Comput. **11**, 44–52 (1999)
24. Lanctot, J., Li, M., Ma, B., Wang, S., Zhang, L.: Distinguishing string selection problems. Information and Computation, Elsevier **185**(1), 41–55 (2003)
25. Lanctot, J., Li, M., Ma, B., Wang, S., Zhang, L.: Distinguishing string selection problems. Inf. Comput. **185**(1), 41–55 (2003)
26. Lin, S., Kernighan, B.W.: An effective heuristic algorithm for the traveling-salesman problem. Oper. Res. **21**, 498–516 (1973)
27. Meneses, C.N., Oliveira, C.A.S., Pardalos, P.M.: Optimization techniques for string selection and comparison problems in genomics. IEEE Eng. Med. Biol. Mag. **24**(3), 81–87 (2005)
28. Mousavi, S.R., Babaie, M., Montazerian, M.: An improved heuristic for the far from most strings problem. J. Heuristics **18**, 239–262 (2012)
29. Resende, M.G.C., Ribeiro, C.C.: Greedy randomized adaptive search procedures. In: Glover, F., Kochenberger, G. (eds.) State-of-the-Art Handbook of Metaheuristics. Kluwer Academic Publishers, Norwell (2002)
30. Ribeiro, C.C., Rosseti, I., Vallejos, R.: Exploiting run time distributions to compare sequential and parallel stochastic local search algorithms. J. Global Optim. **54**, 405–429 (2012)
31. Sim, J.S., Park, K.: The consensus string problem for a metric is NP-complete. J. Discrete Algorithms. **1**(1), 111–117 (2003)

IGV-*plus*: A Java Software for the Analysis and Visualization of Next-Generation Sequencing Data

Antonio Agliata, Marco De Martino, Maria Brigida Ferraro, and Mario Rosario Guarracino

Abstract In this work we describe IGV-*plus*, a software for next-generation sequencing (NGS) data analysis and visualization. It integrates de facto standard tools for the discovery of genetic mutations in genomic-wide association studies. We describe the software specification that led to the development of IGV-*plus*. Finally, we show how we integrate a single-nucleotide polymorphism (SNP) calling software of the genome analysis toolkit (GATK) in the genome browser integrative genomics viewer (IGV), in order to create a centralized platform, as a possible one-stop shop for biologists dealing with NGS data.

1 Introduction

High-throughput technologies for genome sequencing, namely next-generation sequencing (NGS) platforms [1], produce large amounts of biological data. Such data need extensive preprocessing before they can be used for genome-wide association studies, in which genetic variants are detected and correlated to specific traits. The availability of integrated software tools for the analysis, comparison, view, and annotation of such data becomes a discriminating factor for the success of a biological analysis [2].

The idea of NGS technology is similar to capillary electrophoresis (CE)-based Sanger sequencing. The bases of a small fragment of DNA are sequentially identified from signals emitted as each fragment is resynthesized from a DNA template strand. NGS extends this process across millions of reactions through a massive parallelization. It provides an enormous number of reads, which permits the sequencing of entire genomes at a fraction of the costs for Sanger technology. The main steps in NGS data preprocessing consist in computing the quality of base calls,

A. Agliata (✉) • M. De Martino • M.R. Guarracino
High Performance Computing and Networking Institute, National Research Council, Naples, Italy
e-mail: agliataantonio@gmail.com

M.B. Ferraro
Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_8

aligning the short-reads to a reference genome (when available), evaluating quality of such alignment for each short-read, and recalibrating the base calls quality in the context of the aligned reads. After this phase, the discovery of genomic variants is executed, and scientists visualize and analyze results using genome browsers. This visual analysis shows the nucleotide context in which variants occur and other factors of biological interest, such as splicing sites, promoters, terminators, noncoding RNAs, and repeated sequences. Although many software are available for each of such tasks and standards are emerging, some efforts are needed to use the output of one software as the input of the next. To our knowledge, no single software integrates all those steps, thus enabling the users to control the complete process underlying their analysis.

In this work we describe IGV-*plus*, an open-source software that integrates existing solutions for data preprocessing, genetic variant calling, and genome browsing, providing a single tool for each and all the steps needed in NGS data analysis. IGV-*plus* is extensible to plug in other existing tools, given they meet some general requirements. Our analysis of existing software starts from preprocessing and variant discovery software that are noncommercial and open source, with support for input files in BAM and SAM formats [3] and output in VCF format [4]. Furthermore, we only take into consideration those supporting pooled analysis of individuals. Then we analyze existing genome browsers to upload, view, and explore the alignment of the datasets to a reference genome [5]. There exist tens of genome browsers, as a simple web search will reveal. Nevertheless, only a few are widely adopted by the scientific community, usable across different operating systems, freely available and downloadable together with their source code, easy to use and with intuitive graphical user interface (GUI), executable on standard desktop computers, and supported by a developers' community. As a result of our analysis, we decided to integrate the preprocessing and the variant calling software GATK [6] with the genome browser integrative genomics viewer (IGV) [7].

The paper is organized as follows. In Sect. 2 we report a software specification based on the analysis of user needs (Sect. 2.1). In particular, a simple use case diagram is showed and the execution of the main tasks is described. In Sects. 2.2 and 2.3 we analyze tools for single-nucleotide polymorphism (SNP) calling and the existing software solution to browse genomes. In Sect. 3 we provide the implementation details, by analyzing the changes in Java class and the additional packages required for the integration. Finally, in Sect. 4 conclusions are discussed and future work is addressed.

2 Software Specification

Analysis tools in the preprocessing phase are needed to detect positions of mutations among billions of possible ones. Using only a genome browser software would mean manually searching and viewing all data position, which is a daunting and prone-to-error task. Genome browser is useful in the next phase to view the position filtered by preprocessing and analyze them with useful annotation information.

So the goal of IGV-*plus* is creating a centralize platform for biologists dealing with NGS analysis by the integration of analysis tool into genome browser.

In this section we describe the biologists' requirements, the choice of the genome viewer, and the analysis tool to obtain an integrated software.

2.1 User Requirements

Not only do biologists require the list of candidate mutations, but they also need to visually analyze these positions. After the identification of possible causative mutations, they start the biological validation of those that might be of interest to confirm their hypothesis. This is usually done with Sanger sequencing [8] or other techniques, which are expensive in terms of time, resources, and expertise. For this reason, they require to exclude variants that are obviously unrelated with the scope of their research, and artifacts introduced in the amplification process before the sequencing step, to avoid biological validations that would be unnecessary.

From these requirements, a simple use case diagram has been drawn (Fig. 1), where the actor is represented by the software user and it is assumed to be of only one type. In addition to the standard functions performed by a genome viewer, the user can conduct a new analysis and/or evaluate the results of an existing one.

In the *Execute Analysis* task, the user goal is to perform an analysis of the data. The preconditions are that the user is running the IGV-*plus* and executing the "RunTool" task. The success is obtained when the input form is correctly filled in, and the analysis is started. An exception is raised in case input forms are incorrectly filled in. In case of success, it generates a VCF output file. In the *View Results* task, the user goal is to evaluate the results of an analysis carried out earlier or by others. The preconditions require a VCF file and to execute the "Run Tool" task. The success is obtained when the input form is properly filled in, and the input

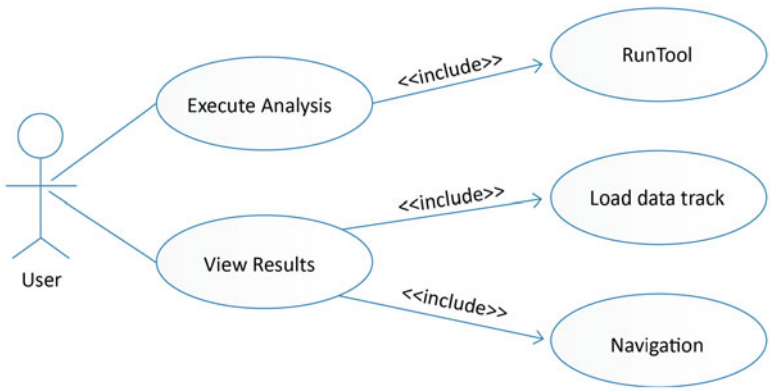

Fig. 1 Use cases diagram

Table 1 Analysis tools: a comparison

Features	GATK	FreeBayes	SNVer	CRISP
Language	Java	C++	Java	Python
Multi-platform	Yes	No	Yes	Yes
Noncommercial	Yes	Yes	Yes	Yes
File format	BAM, BED, FASTA VCF, tab-delim	BAM, FASTA VCF	BAM, BED, FASTA VCF, tab-delim	BAM, FASTA VCF
Supported	Update 2013	Update 2013	Update 2012	Update 2012

file is correctly loaded. An exception is raised in case input forms are incorrectly filled in. The use case describes the evaluation of analysis results that are available in the VCF format. In case of success, the user will be able to browse candidate mutations by including the "Load Data Track" and "Navigation" tasks.

2.2 Preprocessing and Variant Calling Software

We compare four variant calling software: GATK, FreeBayes [9], SNVer [10], and CRISP [11]. The features of these software are described in Table 1.

The choice of variant calling software to integrate falls on GATK and depends on different reasons. First, GATK provides more accurate results of sensitivity and specificity on pooled NGS data [12]. In addition, as we will detail in the following, it is not only a simple variant calling software but also a real analysis framework that provides functions also needed in the preprocessing phase of the data. Finally, it is possible to generate BAM files for other software.

GATK is a software developed at the *Broad Institute* to analyze NGS data. GATK is not only used to perform genomic variant calling, but also integrate analysis to evaluate variants (see Fig. 2). It is a structured framework designed to facilitate the development of efficient and robust analysis tools for NGS data, and it is suitable to be used in projects of any size [6]. GATK is developed in Java and follows the *MapReduce* paradigm that allows the parallelization and distribution of processing by splitting the computation into two steps: *Map* splits the initial problem into independent subproblems and *Reduce* solves the subproblems and combines the partial solutions to get the overall solution to the main problem. The use case describes the evaluation of an analysis results, available in the VCF format.

The currently available version of GATK, 2.7-2, allows us to use two different walkers for this purpose: *Unified Genotyper* and *Haplotype Caller*. Since *Haplotype Caller* does not yet support pooled data, we use the first one.

Searching SNPs in pooled data requires methods capable to distinguish mutations from sequencing errors, mainly with medium-low coverages, and amplification artifacts. With an error rate equal to 1 %, a mutation can be easily mistaken for sequencing error.

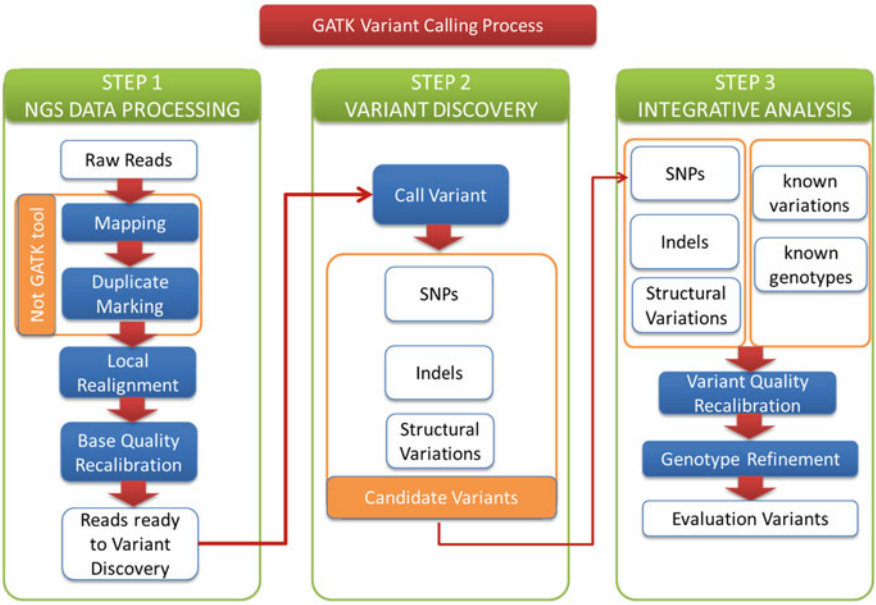

Fig. 2 Calling variants with GATK

The use of the pooling with a good sequencing quality can be very useful to make a proper distinction. Assuming that the reads of each pool are aligned to the reference genome, for each position, we consider the reads of all pools that cover that position. Since GATK it is written in Java, it is executed from the command line.

2.3 Genome Browser

With respect to the choice of the genome viewer software, it derives from the analysis summarized in Table 2.

Although, as shown in Table 2, the software under analysis have characteristics very similar to each other, we choose IGV for the integration, especially for the support it offers for remote control. This feature is very important in an area where large datasets produced by NGS are complex to manage. This means that, when started, IGV runs a web server daemon accepting remote requests. This function can be used to send commands to IGV and to receive results. This choice is also supported by an idea of future development to be implemented, that is, to realize an NGS analysis platform as a web service. The latter facilitates the management of the data, avoiding its duplication among the groups of biologists and of analysts and all those people involved in the project.

Table 2 Visualization software: a comparison

Features	IGV	Savant [13]	Artemis [14]
Open source	Yes	Yes	Yes
Language	Java	Java	Java
Platform	All	All	All
Technology	Illumina, 454, Sanger, ChIP-Seq, RNA-Seq	Illumina, 454, Sanger	Illumina, 454, Sanger
File format	BAM, SAM, GOBY, BED, GFF, GTF, PSL, CN, GCT, FASTA, . . . , . . .	SAM, BAM, FASTA, WIG, GFF, BED, tab-delimited	BAM, VCF, BCF, FASTA, tab-delimited
Remote control	HTTP, HTTPS, FTP	No	No
Supported	Yes	Yes	Yes

Other aspects for which IGV has been chosen are the high support to various file formats and, last but not least, a more complete documentation from the implementation point of view.

IGV is a high-performance viewer for genomic data, capable of handling large heterogeneous datasets while providing a simple and intuitive interface for navigation in the genome [7]. It is written in Java, so it is multi-platform. It supports dataset loading from both local and remote storage. Furthermore, it is available for free under the GNU LGPL license and the required hardware resources allow its installation even on desktop computers. IGV also enables interaction with the data through different levels of detail managed by zoom, which goes from the entire genome to the single base/nucleotide, using an approach similar to that used by *Google Maps*. It uses preprocessed images of genomic data, representing different resolutions, leaving the display at runtime with a better resolution only for the required parts. This approach is called *data tiling* (see Fig. 3).

The software architecture of IGV is divided into three conceptual levels (see Fig. 4). The *application layer* deals with the management of IGV main window and the interaction of the user with the interface elements. IGV displays data in rows that are called *tracks*. They appear in the data panel that handles the layout and the rendering, as well as events related to shared actions, such as zooming. It delegates to object tracks that may refer to object renderers, instanced at runtime, events related to navigation, loading of features, and track design. The *data layer* reads and handles different file formats and makes the data available to the application layer for displaying and on-demand indexing. This allows the optimization of computer resources at runtime. It also creates the data caching to improve efficiency, when requesting a genomic region that is already displayed in the same session. Finally, the *stream layer* manages all the protocols that IGV uses for local and remote access to files and for uploading of the reference genome.

As for all the genome viewers, IGV is mainly used to check the alignment of the analyzed data to the reference genome. Since in the alignment process artifacts

Fig. 3 Integrative genome viewer GUI

are introduced by the sequencing platform, SNPs must be detected from errors. The misalignment, in practice, is suitably highlighted, at certain zoom levels. IGV also offers the ability to display different characteristics of the aligned reads such as mapping quality, the alleles frequency, and many other features used by biologists to validate mutations. Finally, it supports about 30 different input file formats.

3 Implementation Details

In this section we detail the integration of *UnifiedGenotyper*, a specific GATK walker, with IGV. There are two possibilities: integrating the entire GATK code, which is open source, or allowing IGV to execute an instance of GATK. The adopted solution enables the genome viewer to run instances of GATK, whereas the code integration is not convenient in terms of usability and code complexity. Furthermore, this option would make the GUI strictly linked to one version of GATK.

The implementation can be divided into two phases. The first one is reengineering IGV to integrate the variant caller by means of the identification of the package and the integration of the software. The second one consists of developing a GUI that allows the use of GATK.

3.1 IGV Reengineering

The reengineering of IGV starts with its structural analysis. We have identified the packages and classes of interest related to the management of the user interface from the class diagram of IGV (Fig. 4), provided in the documentation. Hence, most of

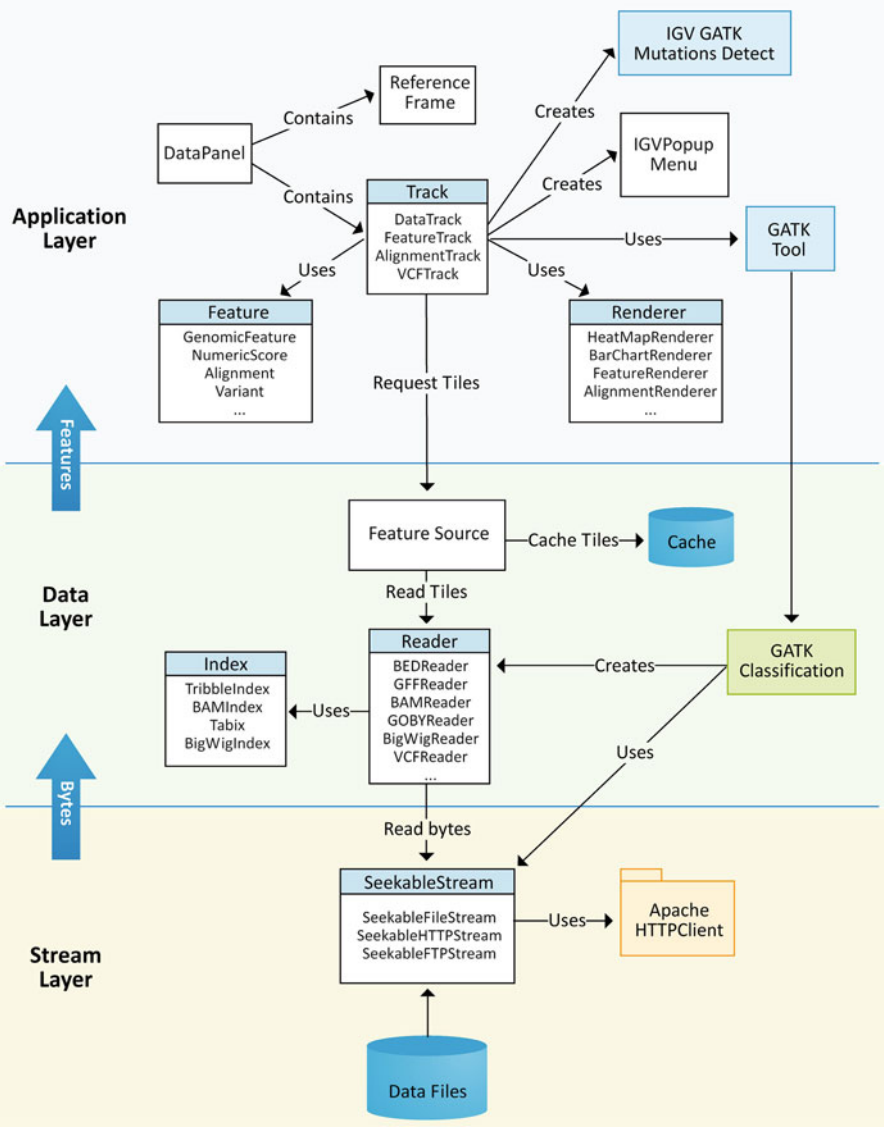

Fig. 4 IGV class diagram

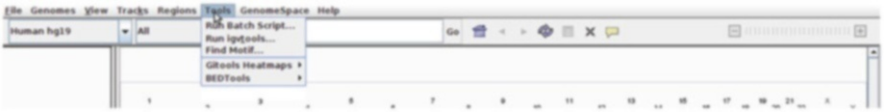

Fig. 5 Menu bar of IGV

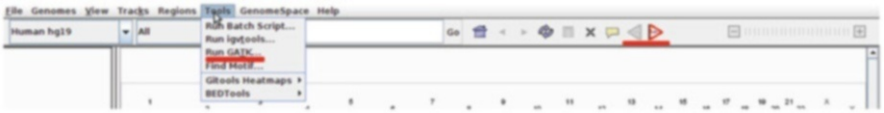

Fig. 6 New menu bar of IGV

the necessary changes of IGV are basically related to interface and, in particular, the command bar and some other menu items (Fig. 5).

The package *org.broad.igv.ui* consists of all the classes dealing with the management of the graphical interface of IGV, the display preferences, and all the involved aspects and graphical events. The goal is to integrate a menu item to run the GATK GUI and to add two buttons *forward* and *back* for browsing the candidate mutations arising from the analysis.

The *IGVMenuBar.java* class deals with the management of the GUI main frame, bringing together items of the main menu and the command bar, in which we want to add the menu item *Tools/Run GATK*. It extends the class *JComponent* that implements the interfaces for the management of mouse events and keyboard for the items in the menu. The *IGVCommandBar.java* deals with the management of the IGV command bar, handling the events of mouse and keyboard inherent to the buttons that appear on the toolbar. Such buttons allow to choose the reference genome, to click the chromosome, to zoom, and other controls to which we want to add the navigation buttons for the candidate mutations.

The edit of the above described classes leads to the addition of new features in IGV (see Fig. 6)

3.2 GATK Graphical User Interface

The capabilities of GATK are structured in two levels: *traversals* and *walkers*. The *traversals layer* is composed of the management modules for common functions. These deal with partitioning and preparation of the data for analysis to be passed to the walkers. The partition of any amount of data is a fundamental problem for the scalability, memory consumption and parallelization task. GATK splits the data into fragments called *shards* whose size is defined by GATK engine and depends on the characteristics of the input BAM file. The *walkers layer* is composed of the management modules of specific functions receiving data from traversals and it

applies the MapReduce paradigm. For example, the function *SNP calling*, used in this work, operates in a Map level as it performs independent processing for each position of the genome. Furthermore, walkers can also run on specific ranges of input files, allowing users to perform analysis only on regions of interest. Each thread performs independently one *MapReduce* call on a single instance of the walker and GATK merges the results of the step *Reduce* of each thread in sequential order, returning the overall step *Reduce*.

The walkers are the core of the functions offered by GATK, making it an essential software to conduct analysis of NGS data even in the case where different tools for the variant calling are used. In the following, we show how to implement a GUI for the *variant calling* walker. This interface provides a GATL function within IGV.

The GATK GUI requires the implementation of new packages for GUI, the management of VCF file format, and the GATK execution command, which are implemented in *org.broad.igv.gatk.ui* (*Gatk.java*), *org.broad.igv.gatk.vcf* (*ReadVcfFile.java*, *VcfBean.java*), and *org.broad.igv.gatk.run* (*RunGatk.java*), respectively.

Gatk.java deals with all the aspects related to the implementation of the GUI, with the construction of the GATK execution string command and with the management of all parameters that it is possible to set. It provides a robust control on entered inputs, properly reporting the wrong and/or missing inputs. The GUI also integrates tooltips menu for each parameter and each option that contains information on the use and meaning of each of them, taken from the manual. *ReadVcfFile.java* contains the implementation of a parser for VCF files. It reads the mutations reported in the VCF file and its properties, setting proper variables. *VcfBean.java* implements a Java bean for the input fields, by means of *get* and *set* for their management. *RunGatk.java* deals with the GATK command that receives input from the GUI, redirecting the standard output from GUI.

The GUI is organized in different areas. As shown in Fig. 7, A area contains forms to select GATK jar for the execution walker, input BAM file to analyze, input humane genome FASTA file to run GATK analysis, input VCF file of known SNPs, and output VCF file path for called variants. If properly inserted, text box is highlighted in green, otherwise in red. In B area it is possible to set GATK command options like memory usage, number of threads to use in the execution, and input interval to analyze. In C area we can set basic parameters related to the *UnifiedGenotyper* walker, to change the default values. In D area we can set advanced UnifiedGenotyper parameters, activated by Advanced checkbox. In E and F areas there are two buttons to run GATK command with selected parameters and to view results in IGV. Finally, in G area, we redirect GATK command standard output, if the execution is successful, and standard errors, otherwise.

To better explain GATK GUI we consider the following execution example. If we want to run a complete NGS analysis we can fill in input forms with GATK jar, input file BAM, human genome FASTA file, and VCF output file path. Then, we select proper parameters to run GATK jar (area B) and analysis (areas C, D). When selecting the Execute Analysis button all parameters are collected for variant calling and analysis starts if errors are not signaled. At the end of execution, we can evaluate

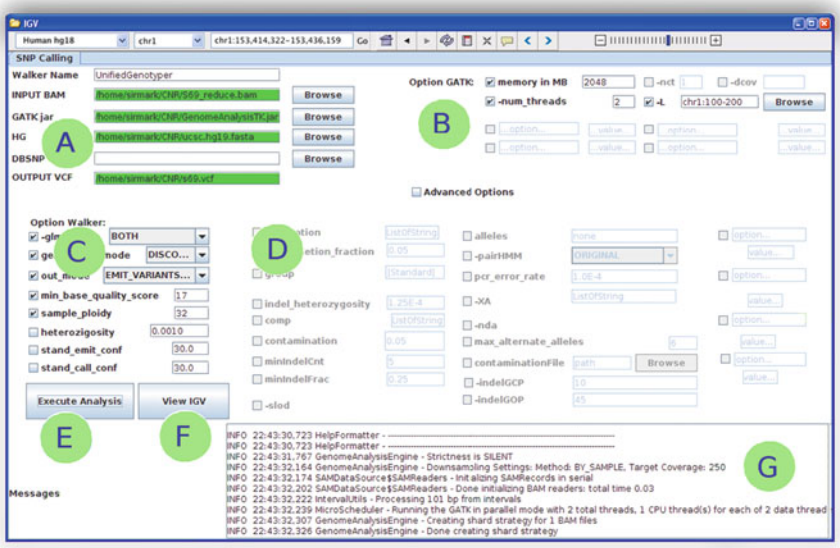

Fig. 7 GATK GUI

GATK execution (area G) and navigate in variants found by clicking the View IGV button that loads input bam file automatically. So we can use the new buttons in IGV GUI to navigate through the called variants.

4 Conclusion and Future Work

In genomic analysis it is necessary to validate the positions of candidate mutations. To reduce the cost of validation, it is useful to view these positions in a software viewer, in order to evaluate also the biological context.

In this work we propose IGV-*plus*, a software for providing biologists with a simple and centralized platform to make an independent genomic analysis.

It is of great interest for biologists to annotate the discovered mutations in order to create an appropriate report of the visual inspection. Furthermore, in the near future, we will integrate other analysis software and mutation reports, and we will provide them as web services. Finally, we will integrate a database to handle data and associate it to patients, and for storing information about them, as in a laboratory information and management systems (LIMS).

References

1. Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S.: Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. **12**, 443–451 (2011)
2. Illumina, Inc.: An introduction to next-generation sequencing technology (2013). http://www.illumina.com/
3. The SAM/BAM Format Specification Working Group: Sequence alignment/map format specification. https://github.com/samtools/hts-specs (2014)
4. The Variant Call Format (VCF) Version 4.1 Specification. https://github.com/samtools/hts-specs (2013)
5. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nat. Biotechnol. **29**, 24–26 (2011)
6. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**, 1297–303 (2010)
7. Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. **14**, 178–192 (2012)
8. Sanger, F., Coulson, A.R.: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. **94**, 441–448 (1975)
9. Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing (2012). http://arxiv.org/abs/1207.3907
10. Wei, Z., Wang, W., Hu, P., Lyon, G.J., Hakonarson, H.: SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucl. Acids Res. **39** (2011). doi:10.1093/nar/gkr599
11. Bansal, V.: A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics **26**, 318–324 (2010)
12. Ferraro, M.B., Guarracino, M.R.: Prediction of single-nucleotide polymorphisms causative of rare diseases. In: Computational Intelligence Methods for Bioinformatics and Biostatistics, pp. 213–224. Springer, Berlin (2014)
13. Fiume, M., Williams, V., Brook, A., Brudno, M.: Savant: genome browser for high-throughput sequencing data. Bioinformatics **26**, 1938–1944 (2010)
14. Carver, T., Harris, S.R., Berriman, M., Parkhill, J., McQuillan, J.A.: Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics **28**, 464–469 (2012)

Statistical Techniques for Assessing Cyberspace Security

Alla R. Kammerdiner

Abstract This chapter describes statistical approaches to cyber security. Telecommunication and computer network systems form a physical foundation of a cyberspace. Cyberspace is of critical importance to national security and economy. Because cyber attacks and cybercrime pose a considerable and increasing threat to society, security of cyber systems must be improved. In intrusion detection, three fundamental methodologies for cyber attack detection are known, namely anomaly detection, signature recognition, and attack norm separation. Based on these methodologies, various statistical approaches for detecting cyber attacks have been developed. However, cyber attacks continue to evolve. Distributed attacks are emerging that hijack and harvest the power of cloud technologies to cause disruption. System-wide approaches are needed to enable early detection of such attacks. We propose a new graph-based modeling framework and derive a network-based statistical method that could help detect compromised structures in a cyberspace. We illustrate our new approach with an example of a small cyberspace.

Keywords Graph theory • Statistical inference • System of systems • Cyber security

1 Introduction

Cyberspace is commonly thought as an abstract environment enabled by telecommunication and computer networks. Increasingly, cyberspace continues to move from an abstraction to being an integral part of our reality. Cyberspace has become "a defining feature of modern life" [6] utilized by billions of people and communities all over the world to connect, socialize, and self-organize and by countless organizations to reach out, self-promote, and engage in business.

Cyberspace is embedded into a nation's economy, infrastructures, and defense. Cyberspace is an operational basis for online banking and retail with e-commerce

A.R. Kammerdiner (✉)
New Mexico State University, P.O. Box 30001, MSC 4230, Las Cruces, NM 88012, USA
e-mail: alla@nmsu.edu

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_9

accounting for 4.1 % of gross domestic product (GDP) in the G-20 countries in 2010 [7]. The percentage of the internet economy in the GDP in the G-20 is growing rapidly and is projected to reach 5.3 % by 2016 [7]. The operation of the US critical infrastructure also depends on cyberspace. Through information technologies and industrial control systems, many parts of our critical infrastructure including communication, energy, finance, and transportation are, to some degree, connected to cyberspace [6]. This connection allows greater access to and more efficient use of these resources. From the standpoint of national defense, cyberspace is a complex and dynamic operational domain of strategic importance to the military. The US military has capability to use cyberspace for rapid communication and information sharing in support of its critical missions. The military developed cyber-enabled weapons and systems, whose operation is reliant on cyberspace.

Emergence of cyberspace as a reality is connected to the global expansion and penetration of telecommunications in the twenty-first century. In 2013, an estimated number of users of the internet in the world exceeded 2.7 billion people reaching 39 % of global population [10]. Moreover, in 2013 the number of mobile-cellular subscriptions reached 6.8 billion approaching 100 % of the entire global population (i.e., 7.1 billion people) [10]. The wide availability of mobile and internet technologies not only connects individuals, communities, and countries, but also exposes the internet and mobile users to threats and attacks.

Telecommunication and computer network systems have known and unknown vulnerabilities that can be exploited. Because cyberspace is enabled by these systems, cyberspace security is affected by vulnerabilities in the network systems. Despite continuing significant efforts to address the existing vulnerabilities in computer and telecommunications networks and technologies, security of our cyberspace is inadequate. This fact is particularly apparent in the proliferation of cybercrime, which costs the US economy between $70 billion and $140 billion each year [21]. Globally annual losses from cybercrime activities range between $80 billion and $400 billion (based on conservative estimates) and may reach up to $1 trillion (based on alternative estimates) [21]. The internet crime is on the rise with the reported 8.3 % growth in unverified losses reported to IC3 in 2012 [11].

In this chapter, we define a cyberspace as a system of subsystems connected via the internet and internal communication networks. Examples of subsystems include computer networks, corporate networks, clouds, industrial control systems, etc. In terms of belonging to a given organization, systems can be internal, external, and mixed (i.e., a combination of internal and external components). Traditionally, a focus of cyber security has been on the internal part of a network. We argue that due to wireless and cloud technologies the internal part is becoming increasingly more exposed to the outside threats. Consequently, cyber security should include not only detection of cyber attacks on internal networks but also an assessment of the exposure of internal networks to outside threats. The main contribution of this chapter is the derivation of a mathematical method for a system-wide assessment.

In this chapter some statistical approaches for assessing security of cyber systems and cyberspace are presented. The rest of the chapter is organized as follows. Section 2 introduces the reader to three fundamental methodologies for

intrusion detection and describes some of the statistical approaches for detecting and identifying cyber attacks based on these methodologies. An excellent survey of research papers on statistical methods for cyber attack detection with additional details can be found in [22]. In Sect. 3, we first present a motivation for network-based statistical methods for cyber attack detection and monitoring and then introduce a new statistical approach for detecting compromised connections in a cyberspace. We finish this section with an example illustrating application of our approach on a network of a small size. Finally, Sect. 4 summarizes the key points in this chapter.

2 Statistical Analyses of Cyber Systems Data for Security Assessment

An attack on a computer and network system results in an activity, a state change, or a performance change, which can be observed by monitoring the relevant variables, such as available memory on a computer, throughput for a network connection, and characteristics of a process. To detect and identify an attack on a given system, the data on such variables can be collected and compared to normal use conditions for the system (e.g., web browsing, text editing, etc.). Statistical methods are often utilized in detection and identification of cyber attack.

In intrusion detection, three basic methodologies are distinguished: signature recognition, anomaly detection, and attack norm separation [22]. Signature recognition is applied by first identifying signature patterns of attack data and then using these patterns to monitor data from a system and generate alarms. Anomaly detection is applied by first creating a model of the normal use, called a norm profile, which is then used to detect large deviation from norm profile, called an anomaly. Signature recognition and anomaly detection are two conventional methodologies for cyber attack identification and detection [22]. As an alternative to these conventional methodologies, a more recent attack norm separation methodology is designed to work on the data that represent mixed effects of normal use and cyber attack activities.

Cyber Attack Signature Recognition

Statistical methods for signature recognition include supervised clustering and classification algorithms. Such methods are particularly suitable because signature recognition aims at distinguishing patterns of data during cyber attack from data during normal use. In relation to application of classification techniques for recognizing signatures of cyber attacks, decision trees [13, 35] and artificial neural networks [22] have been studied in the literature. Scalable incremental classification technique for attack signature recognition has been studied in [25].

In signature recognition, supervised clustering methods [14–16] are used that aim at incremental clustering of data points so that any new data point can be added to the existing clusters. Incremental clustering allows to accommodate for successive modifications and an evolution seen in malicious code and attack signatures with time. Hence, incremental clustering is more suitable for signature recognition than density-based and hierarchical clustering methods. In application of clustering for recognizing signatures of cyber attacks, dummy-based clustering [14] and grid-based clustering [14, 15] have been studied in the literature. To overcome dependence of the conventional incremental clustering techniques on the order in which the data points are presented, a more robust clustering procedure for signature-recognition-based intrusion detection was developed in [15]. Recently, an approach for network intrusion detection based on unsupervised clustering has been proposed [3].

Association rules is an alternative method developed for recognition of cyber attack signatures [12]. This data-mining method is used to find interesting relationships among two or more variables. In application to attack signature recognition, it works by learning the frequent itemsets for various types of cyber attacks from the combined normal use and attack data. The discovered relationships are viewed as attack signatures and used for recognizing cyber attacks.

Signature recognition cannot handle data with mixed cyber attack and normal use activities [22].

Anomaly Detection for Detecting Cyber Attacks

In cyber attack detection, statistical methods focused on anomaly detection include statistical process control techniques and stochastic process models [22]. Statistical process control (SPC) is particularly suitable for anomaly detection because SPC procedures aim to detect out-of-control events, i.e., the events that are statistically rather unlikely given the model that fits the data from a normal (in-control) process. In application of SPC for cyber attack detection both univariate and multivariate approaches have been used to construct the normal use profile (i.e., a model of an in-control process) from normal use data.

For example, in [24,27,28,32] a univariate SPC approach called the exponentially weighted moving average (EWMA) control chart has been applied for statistical anomaly detection in cyber systems. This technique works well on both normally and non-normally distributed data. It is shown [22] that variables describing activities, performance, and states of a cyber system do not necessarily fit a normal distribution. Instead they often have skewed or multimodal distributions. Because the EWMA control chart is robust to non-normality, it is better suited than the SPC procedures requiring normality of the data. Recently, the SPC approach called the failure quality control (FQC) has been proposed for detection of anomalous packets in the network for real-time intrusion detection [17].

Many conventional multivariate SPC methods require computing large covariance matrices and taking inverse of covariance. Hence, these methods are poorly

suited for analyzing large-scale, complex cyber networks. Similarly to univariate SPC, some multivariate SPC techniques assume normally distributed data. To avoid these pitfalls a scalable multivariate SPC approach known as the chi-square distance monitoring (CSDM) has been developed in [8, 23, 26, 29, 31, 33]. The approach is capable of dealing with large datasets containing both normally and non-normally distributed variables.

An alternative to SPC, stochastic process models for detecting anomalies in cyber systems take into account the sequential order of the events, which is not utilized by the SPC methods. Stationary Markov chains have been used in [30, 34] to build the norm profile of event transitions for anomaly detection. First the Markov chain model is estimated from the normal use data. Then anomalies are detected when the joint probability of a sequence of most recent events (which is calculated based on the estimated model) is small. Both first-order and higher-order Markov chains can be used for cyber attack detection; however there is a natural trade-off between the computational cost and modeling accuracy. Anomaly detection, even using more powerful stochastic models instead of SPC, cannot handle mixture data of cyber attack and normal use activities [22].

Attack Norm Separation

Attacks typically occur during the normal use activities. This greatly affects the detection accuracy of pattern recognition and anomaly detection approaches that view cyber attacks and normal use as activities occurring at separate times. Attack norm separation has been developed to be able to handle data with mixed activities of attack and normal use. Attack norm separation applies signal-noise separation method to filter out the normal use activities co-occurring with attack from the mixed activities data.

In attack norm separation, first the attack and norm models are defined, second the norm model is used to cancel out the effect of the normal use activities in the mixture data and obtain the residual data, and third the attack model is applied to detect an attack in the residual data. In the first step, the data characteristics of attack and norm activities can be discovered based on the data features of mean, probability distribution, autocorrelation, and various wavelet transforms. In the second and third steps, the normal use is considered as noise and the cyber attack is treated as the signal. Hence, the characteristics of norm activities need to be filtered out to get the residual data. The cancellation depends on the modeling assumption about how the norm and attack are mixed together. For instance, an α-stable model for norm-attack data is used in [20], and an additive mixture model for norm-attack data is presented in [22]. Many signal processing techniques can be used for noise cancellation and signal detection. In [1], the SCP method based on the cumulative score (cuscore) statistics is used to detect changes in the residual data signaling an attack. Attack norm separation approach has much better accuracy and timeliness than anomaly detection and signature recognition when detecting cyber attacks from the mixture of normal use and attack data [1, 22].

3 Statistical Detection of Compromised Graph Structure Under Variable Routing

In Sect. 3 we propose a new statistical technique for assessing cyberspace security. First, we present a motivation behind the development of new approach. Second, we describe the graph-based modeling framework and then use it to develop the statistical procedure for detection of compromised graph structure. Third, we a present a simple example to demonstrate our approach.

3.1 Motivation for New Statistical Approaches to Network Security

In the last quarter of 2012, major US banks experienced a series of directed cyber attacks. These attacks were so powerful because they were directed through hijacked data centers (i.e., the cloud) around the world instead of individual computers, according to the report by New York Times [19]. Clearly, these cyber attacks from hijacked cloud nodes as other organized distributed denial-of-service (DDoS) attacks represent an emergent type of threat, which *utilizes various paths to the target* sites or nodes to cause significant disruptions. Most existing techniques do not adequately address the distributed nature of DDoS attacks.

To address the distributed nature and variability in the attack path, new system-wide-based approaches for attack detection and identification are needed. A cyber system can naturally be represented by a network model. Although surveillance of entire internet is computationally infeasible, locally monitoring a network graph within some radius may be possible. This type of monitoring could provide additional and earlier information for detection of distributed cyber attack based on traffic, resource utilization, and other variables.

In response to increasing threat of directed massive cyber attacks such as DDoS to interconnected network and computer systems, recent research in cyber security is concerned with the development of *federated networks protection system* [5]. Military networks, critical infrastructure, and public administration computer systems are often connected into a federation of systems (FoS), together forming a complex system of subsystems. In [5], it is argued that the synergy effect for security can be gained by adopting the idea of federation networks. Specifically, the federated network and systems can share and exchange information about events in the network, detected attacks, and proposed countermeasures [5]. The federated network approach to cyber security may replace inefficient approach of "closed security" [4], which was develop from the perspective of a single network. This new approach has already received significant attention not only in computer security but also in the context of military networks and critical systems, including NATO network-enabled capability (NNEC) [2, 9, 18].

Motivated by the security of FoS and other cyber-physical systems, we define a cyberspace as a system of subsystems connected via the internet and internal communication networks. The individual subsystems can be computer networks, corporate networks, clouds, and industrial control systems. The components of each subsystem interact with various components of other subsystems. Three types of subsystems can be distinguished in terms of inclusion into a given organization or federation, namely internal, external, and mixed (i.e., comprised of internal and external components). Traditionally, a focus of cyber security has been on the internal part of a network. Most statistical approaches for cyber security described in Sect. 2 have been developed for detection of intrusions into an internal network. This "closed security" [4] approach is now considered largely ineffective in dealing with directed large-scale cyber attacks [2, 4, 5, 9, 18]. Moreover, wireless and cloud technologies increasingly expose an internal network to the outside threats. To address these new threats, cyber security should also incorporate an assessment of the exposure of internal networks to outside threats. In the remainder of this chapter, we present the derivation of a mathematical method for such a system-wide assessment.

3.2 Detecting Structure Under Variability in Graph Dynamics

In this Sect. we derive a procedure for detecting hidden structure of the network from data. We formulate the procedure for the networks where a hidden connection to a compromised network resource may exist. We assume the variability in routing through the networks and utilize this uncertainty to construct a random graph model, which is used in the inference procedure.

Modeling Framework

Let $V = \{1, \ldots, |V|\}$ be a vertex set representing network resources and $E = \{(v, u) : v \in V_{out} \subseteq V, u \in V_{in} \subseteq V\}$ be an edge set denoting (directed) cyber connections between pairs of resources. Then a connected directed graph (digraph) $G = (V, E)$ represents cyberspace. Some vertices $v \in V$ in this cyberspace are compromised (for instance, resources with known vulnerabilities or unsafe location).

Furthermore, suppose that some information about E is unknown (e.g., hidden or obscured). Unknown connections from or to compromised resources represent hidden threats. Because identifying these threats is important for assessing cyberspace security, the following question arises:
Can some of this unknown information about E be inferred through monitoring the network data?

Suppose that we can monitor the transmission of data on the network G using a pair of trustworthy resources $s, t \in V$. Suppose that we send the data from s to t.

Hence, vertex s is a source, whereas t is a sink. We can assume without loss of generality that $(s,t) \notin E$. Otherwise the edge (s,t) through two trustworthy vertices s and t can be removed from graph G.

We suppose that each message from s to t is sent through one path to t, although different path may be utilized at different times. In other words, we assume that the routing of data through G exhibit variability and model the uncertainty in routing as follows. If we know that a connection exists (i.e., a known connection $(i,j) \in E$), then the corresponding edge of G can route our message with probability p, $0 < p < 1$. Let E_1 denote a set of unknown or hidden edges (connections). If we do not know whether a potential connection exists (i.e., an unknown or hidden connection $(i,j) \in E_1, (i,j) \notin E$ with $i,j \in V$), then the corresponding edge can route our message with probability r, $0 \leq r < p < 1$.

Notice that through introduction of these probabilities, we have defined a probability space (Ω, P) on the set of known and hidden edges of G. Where $\Omega = \{\omega_{ij} : i,j \in V\}$ and the random outcomes $\omega_{ij} \in \{0,1\}$ are such that

- $P(\omega_{ij} = 1) = p$ and $P(\omega_{ij} = 0) = 1 - p$ for known edges $(i,j) \in E$,
- $P(\omega_{ij} = 1) = r$ and $P(\omega_{ij} = 0) = 1 - r$ for hidden edges $(i,j) \in E_1$, and
- $P(\omega_{ij} = 0) = 1$ and $P(\omega_{ij} = 1) = 0$ for proven nonexistent edges $(i,j) \notin E \cup E_1$.

In what follows we consider the situation when cyberspace G contains a single compromised resource denoted by vertex $c \in V$ and a single *unknown* possible two-way connection represented by a pair of opposite edges $(c,u), (u,c) \notin E$ linking c with $u \in V$. Additionally, assume that vertices c and u are such that there exist paths $\gamma_c = (s,\ldots,c,\ldots,t)$ and $\gamma_u = (s,\ldots,u,\ldots,t)$ from s to t in G passing through c and u, respectively.

We assume that when a message is routed from s to t through a compromised resource, the message is always either altered or delayed in such a noticeable way that one can detect the change or delay on t. By sending messages from s to t and observing the outcomes, we sample the integrity of transmitted data. Let $\overline{X} = (x_1,\ldots,x_N)$ denote the observed data sample of size N, where for $i = 1,\ldots,N$ the observation

$$x_i = \begin{cases} 1 \text{ if the } i\text{th message reaches } t \text{ without being altered or delayed,} \\ 0 \text{ if the } i\text{th message is altered or delayed.} \end{cases} \tag{1}$$

The data sample could be used to infer whether the hidden bidirectional connection via (u,c) and (c,u) exists. Then the problem of detecting the hidden graph structure from the data can be formulated as testing the hypothesis H_0 versus alternative H_1, where

Hypothesis H_0: $(c,u) \notin E$ and $(u,c) \notin E$, i.e., unknown edges (c,u) and (u,c) do *not* exist in cyberspace G.

Alternative H_1: $(c,u) \in E$ and $(u,c) \in E$, i.e., unknown edges (c,u) and (u,c) exist in cyberspace G.

Derivation of the Edge Detection Procedure

Let a random variable or vector γ denote a route passed by a message sent from s to t on G. Clearly, a routing path should not contain any directed loops. Let L denote the maximum length of a path without directed loops in G. Let the length of such a path be denoted by l, then $2 \leq l \leq L$ in G.

For each specific path $(s, v_1, \ldots, v_{l-1}, t) := \beta$ from s to t of length l in G, we can compute the probability of $\gamma = \gamma_0$ under the hypothesis H_0 and under the alternative H_1. If $(s, v_1), (v_i, v_{i+1}), (v_{l-1}, t) \in E$ for $i = \overline{1, l-2}$ then under H_0 we have

$$P(\gamma = (s, v_1, \ldots, v_{l-1}, t)|H_0) = p^l (1-p)^{|E|-l}. \tag{2}$$

Similarly, if $(s, v_1), (v_i, v_{i+1}), (v_{l-1}, t) \in E \cup E_1$ for $i = \overline{1, l-2}$, then under H_1 we have two cases.

1. When path does *not* contain compromised vertex c, we get

$$P(\gamma = (s, v_1, \ldots, v_{l-1}, t), v_i \neq c \ \forall i = \overline{1, l-1}|H_1) = p^l (1-r)(1-p)^{|E|-l}. \tag{3}$$

2. When path contains c, we obtain

$$P(\gamma = (s, v_1, \ldots, v_{l-1}, t), \exists i = \overline{1, l-1} : v_i = c|H_1) = p^{l-1} r (1-p)^{|E|-l}. \tag{4}$$

Let $\mathscr{A}$ denote all paths β from s to t in G without directed loops, i.e.,

$$\begin{aligned} \mathscr{A} = \Big\{ \beta : \beta = (s, v_1, \ldots, v_{l-1}, t), \\ \text{with } (s, v_1), (v_i, v_{i+1}), (v_{l-1}, t) \in E \cup E_1, \\ \text{and } v_i \in V, i = \overline{1, l-1}, \text{ for some } l = \overline{2, L} \Big\}. \end{aligned} \tag{5}$$

All possible paths from s to t in G can be divided into two disjoint classes, those containing c and those *not* containing it. Let $\mathscr{U}$ denote the collection of all paths β from s to t in G that do not pass via c, i.e.,

$$\begin{aligned} \mathscr{U} = \Big\{ \beta : \beta = (s, v_1, \ldots, v_{l-1}, t), \text{ with } (s, v_1), (v_i, v_{i+1}), (v_{l-1}, t) \in E \\ v_i \in V \text{ and } v_i \neq c \ \forall i = \overline{1, l-1}, \text{ for some } l = \overline{2, L} \Big\}. \end{aligned} \tag{6}$$

Let $\mathscr{C}$ denote all paths β from s to t in G that pass via c, i.e.,

$$\begin{aligned} \mathscr{C} = \Big\{ \beta : \beta = (s, v_1, \ldots, v_{l-1}, t), \text{ with } (s, v_1), (v_i, v_{i+1}), (v_{l-1}, t) \in E \cup E_1, \\ v_i \in V \text{ and } \exists i = \overline{1, l-1} : v_i = c, \text{ for some } l = \overline{2, L} \Big\}. \end{aligned} \tag{7}$$

We say that $\mathscr{U}$ are the safe paths in G and $\mathscr{C}$ are the compromised paths. For any $2 \le l \le L$, we introduce the following three notations for numbers of path of length l. Let

- $a(l)$ denote the number of all possible paths of length l (specifically, a total number of distinct paths from s to t in G having length l without directed loops);
- $c(l)$ be the number of compromised paths of length l (specifically, a total number of distinct paths of length l in $\mathscr{C}$ without directed loops);
- $u(l)$ denote the number of safe paths of length l (specifically, a total number of distinct paths of length l in $\mathscr{U}$ without directed loops).

Then

$$a(l) = c(l) + u(l), \text{ or } u(l) = a(l) - c(l), \text{ for } l = \overline{2, L}. \tag{8}$$

The event that a random route γ starts at s and ends at t induces a new probability space. Let us consider a graph Γ on V with edges given by a random route γ. We denote $\Gamma = \Gamma(\gamma)$. Then Γ is a random graph with the induced probability distribution. The induced probabilities P_Γ on Γ can be computed for the hypothesis H_0 and for the alternative H_1.

We compute the induced probabilities for H_0 as follows. When H_0 is true, we have

$$P(\gamma \in \mathscr{A} | H_0) = P(\gamma \text{ is a path from } s \text{ to } t | H_0) = \sum_{l=2}^{L} a(l) p^l (1-p)^{|E|-l}, \tag{9}$$

and

$$\begin{aligned} P(\gamma \in \mathscr{U} | H_0) &= P(\gamma \text{ is a safe path from } s \text{ to } t | H_0) \\ &= \sum_{l=2}^{L} [a(l) - c(l)] p^l (1-p)^{|E|-l}. \end{aligned} \tag{10}$$

Note that because we consider distinct paths of without loops, the resulting events are disjoint leading to the sums in (9) and (10). Hence,

$$P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_0) = \frac{\sum_{l=2}^{L} [a(l) - c(l)] p^l (1-p)^{|E|-l}}{\sum_{l=2}^{L} a(l) p^l (1-p)^{|E|-l}}. \tag{11}$$

Additionally, let us introduce four new notations for the numbers of incoming and outgoing paths for vertices c and u. Let

- $c_{in}(k)$ denote the number of paths from s to c in G of length k without directed loops;

- $c_{out}(k)$ denote the number of paths from c to t in G of length k without directed loops;
- $u_{in}(k)$ denote the number of such paths $(s, v_1, \ldots, v_{k-1}, u)$ in G of length k that have $v_i \neq c$ for all $i = \overline{1, k-1}$ and have *no* directed loops;
- $u_{out}(k)$ denote the number of paths $(u, v_1, \ldots, v_{k-1}, t)$ in G of length k that have $v_i \neq c$ for all $i = \overline{1, k-1}$ and have *no* directed loops.

Note that under H_0 we have

$$c(l) = \sum_{k=1}^{l-1} c_{in}(k) \cdot c_{out}(l-k), \text{ for any } l = \overline{2, L}. \tag{12}$$

Under H_1, $G = (V, E)$ is replaced with $G_1 = (V, E \cup E_1)$. Because $E_1 = \{(c, u), (u, c)\}$, there are some additional compromised paths. In particular, we have $\sum_{l=2}^{L} \sum_{k=1}^{l-1} c_{in}(k) \cdot u_{out}(l-k)$ paths passing through edge (c, u) and $\sum_{l=2}^{L} \sum_{k=1}^{l-1} u_{in}(k) \cdot c_{out}(l-k)$ paths passing through edge (u, c). Because these paths include an extra edge, they measure anywhere between 3 and $L+1$ in lengths. Therefore, the total number of paths in G_1 of length l is

$$a_1(2) = a(2), \tag{13}$$

$$a_1(l) = a(l) + \sum_{k=1}^{l-2} [c_{in}(k) \cdot u_{out}(l-1-k) + u_{in}(k) \cdot c_{out}(l-1-k)], \tag{14}$$

$$\text{for } 3 \leq l \leq L,$$

$$a_1(L+1) = \sum_{k=1}^{L-1} [c_{in}(k) \cdot u_{out}(L-k) + u_{in}(k) \cdot c_{out}(L-k)]. \tag{15}$$

Similarly, the total number of compromised paths in G_1 of length l is

$$c_1(2) = c(2), \tag{16}$$

$$c_1(l) = c(l) + \sum_{k=1}^{l-2} [c_{in}(k) \cdot u_{out}(l-1-k) + u_{in}(k) \cdot c_{out}(l-1-k)], \tag{17}$$

$$\text{for } 3 \leq l \leq L,$$

$$c_1(L+1) = \sum_{k=1}^{L-1} [c_{in}(k) \cdot u_{out}(L-k) + u_{in}(k) \cdot c_{out}(L-k)]. \tag{18}$$

Next, we compute the induced probabilities for H_1 as follows. When H_1 is true and we have

$$P(\gamma \in \mathscr{A}|H_1) = P(\gamma \text{ is a path from } s \text{ to } t|H_1) \tag{19}$$
$$= \sum_{l=2}^{L} \left(a(l)p^l(1-r)(1-p)^{|E|-l} + p^l r(1-p)^{|E|-l} \times \right.$$
$$\left. \times \sum_{k=1}^{l-1} [c_{in}(k)u_{out}(l-k) + u_{in}(k)c_{out}(l-k)] \right)$$
$$= \sum_{l=2}^{L} \left\{ a(l)(1-r) + r \sum_{k=1}^{l-1} [c_{in}(k)u_{out}(l-k) + u_{in}(k)c_{out}(l-k)] \right\}$$
$$\times p^l(1-p)^{|E|-l},$$

and

$$P(\gamma \in \mathscr{U}|H_1) = P(\gamma \text{ is a safe path from } s \text{ to } t|H_1) \tag{20}$$
$$= \sum_{l=2}^{L} [a(l) - c(l)]p^l(1-p)^{|E|-l}(1-r).$$

Hence,

$$P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U}|H_1) = \frac{P(\gamma \in \mathscr{U}|H_1)}{P(\gamma \in \mathscr{A}|H_1)}. \tag{21}$$

When the induced distributions in (11) and (21) are the same, then it is not possible to distinguish between H_0 and H_1 and so we cannot distinguish the topologies of known cyberspace G and alternative cyberspace G_1. But, when the induced distributions are different under H_0 and H_1, i.e.,

$$P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U}|H_1) \neq P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U}|H_0), \tag{22}$$

we could use this difference to construct the statistical procedures for detecting hidden edges.

To construct the procedure, let a significance level be set to some $\alpha = \alpha_0$. By passing N messages from node s to node t, we collect the observations $\overline{X} = (x_1, \ldots, x_N)$ on the integrity of transmitted data (e.g., whether any of the original messages were preserved or altered during the routing). Recall from (1) that individual values x_i's in the data sample $\overline{X}$ are realizations of independent identically distributed (i.i.d.) Bernoulli trials with the following success probability

$$\pi = P(\text{ if the ith message reaches t without being altered or delayed}). \tag{23}$$

The ith message reaches t without being altered or delayed if and only if the message passes via some safe route $\beta \in \mathscr{U}$, which is equivalent to having $\Gamma = \Gamma(\beta)$ for some $\beta \in \mathscr{U}$. Therefore, the probability in (23) is also equal to

$$\pi = P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U}). \tag{24}$$

Let random vector $X = (X_1, \ldots, X_N)$ with $X_i \sim$ Bernoulli (π), $i = \overline{1, N}$ represent the outcomes of our trials before observations were collected. One can easily construct a sufficient statistic for a sample of i.i.d. Bernoulli variables. Let $T = T(X)$ be a random variable representing the number of unchanged messages in N trials. Since our trials are Bernoulli with success probability π, then T is a Binomial random variable with parameters N and π, denoted $T \sim \text{Bin}(N, \pi)$. Thus, as long as (22) holds, the problem of the hidden edges detection can be reduced to testing a hypothesis for a binomial parameter π. Specifically, we test

$$H_0 : \pi = \pi_0 \tag{25}$$

against

$$H_1 : \pi = \pi_1 \tag{26}$$

where

$$\pi_i := P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_i), i = 0, 1. \tag{27}$$

We can distinguish three important cases based on the values of the induced probabilities in the left-hand side of (11) and (21):

1. $P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_1) < P_\Gamma(\Gamma : \Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_0)$.
   In this case, the induced distributions are different under H_0 and H_1. Hence, we could use this difference to construct the one-sided test $H_0 : \pi = \pi_0$ versus $H_1 : \pi < \pi_0$ based on statistics T at a significance level α_0.
2. $P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_1) > P_\Gamma(\Gamma : \Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_0)$.
   In this case, the induced distributions are different under H_0 and H_1. Hence, we could use this difference to construct the one-sided test $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$ based on statistics T at a significance level α_0.
3. $P_\Gamma(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_1) = P_\Gamma(\Gamma : \Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U} | H_0)$.
   In this case, the induced distributions are identical under H_0 and H_1. Hence, we could not distinguish G and G_1.

Observe that

$$P(\gamma \in \mathscr{U} | H_0)(1 - r) = P(\gamma \in \mathscr{U} | H_1). \tag{28}$$

Using (28) in (11) and (21), we can restate the above three cases as follows:

1. If $P(\gamma \in \mathscr{A}|H_1) > P(\gamma \in \mathscr{A}|H_0)(1-r)$ in (19) and (9), respectively, then conduct the one-sided test $H_0 : \pi = \pi_0$ versus $H_1 : \pi < \pi_0$ based on statistics T at a significance level α_0.
2. If $P(\gamma \in \mathscr{A}|H_1) < P(\gamma \in \mathscr{A}|H_0)(1-r)$ in (19) and (9), respectively, then conduct the one-sided test $H_0 : \pi = \pi_0$ versus $H_1 : \pi > \pi_0$ based on statistics T at a significance level α_0.
3. If $P(\gamma \in \mathscr{A}|H_1) = P(\gamma \in \mathscr{A}|H_0)(1-r)$ in (19) and (9), respectively, then G and G_1 cannot be distinguished.

3.3 *Example: Construction of a Test to Infer Compromised Routing in a Simple Cyberspace*

Consider a simple cyberspace G with 4 resources $V = \{s, 1, 2, t\}$, known connections $E = \{(s,1),(s,2),(1,t),(2,t)\}$, and hidden edges $E_1 = \{(1,2),(2,1)\}$, where exposed resource is $u = 1$ and compromised resource is $c = 2$. Also notice that as required we have $(s,t) \notin E$.

In this example the messages are sent from s to t in G, but they cannot be transmitted directly through (s,t). For known connections, it is estimated that links $(s,1)$, $(s,2)$, $(1,t)$, and $(2,t)$ can route a message with some probability $p > 0$. Whereas, for hidden connections provided they exist, the links $(1,2)$ and $(2,1)$ can route a message with probability $r > 0$, $r < p$. The compromised resource $c = 2$ changes messages. To assess the degree of exposure of resource $u = 1$, we seek to determine *whether the compromised resource* $c = 2$ *can route messages from and to the exposed resource* $u = 1$?

We apply our method to infer whether hidden connection $(1,2)$ and $(2,1)$ exist given sampled data $\overline{X}$ on node t. Specifically, we test hypothesis $H_0 : (1,2),(2,1) \notin E$ against alternative $H_1 : (1,2),(2,1) \in E$. It is easy to see that $|E| = 4$, $L = 2$, $a(2) = 2$, $u(2) = 1$, $c(2) = 1$. Moreover, $L+1 = 3$, $u_{in}(1) = u_{out}(1) = c_{in}(1) = c_{out}(1) = 1$, and $u_{in}(2) = u_{out}(2) = c_{in}(2) = c_{out}(2) = 0$. From (13) we have $a_1(2) = a(2) = 2$ and $a_1(3) = 1\cdot 1 + 1\cdot 1 + 0\cdot 0 + 0\cdot 0 = 2$. Using (16), we obtain $c_1(2) = 1$ and $c_1(3) = 1 + 1 = 2$.

Hence, for H_0, using (9), (10), and (11), we have

$$P(\gamma \in \mathscr{A}|H_0) = P(\gamma \text{ is a path from } s \text{ to } t|H_0) = 2p^2(1-p)^2, \tag{29}$$

$$\begin{aligned} P(\gamma \in \mathscr{U}|H_0) &= P(\gamma \text{ is a safe path from } s \text{ to } t|H_0) \\ &= p^2(1-p)^2, \end{aligned} \tag{30}$$

and

$$P_{\Gamma}(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U}|H_0) = \frac{1}{2}. \tag{31}$$

Similarly for H_1, applying (19), (20), and (21), we have

$$P(\gamma \in \mathscr{A}|H_1) = P(\gamma \text{ is a path from } s \text{ to } t|H_1) = 2p^2(1-p)^2, \tag{32}$$

$$P(\gamma \in \mathscr{U}|H_1) = P(\gamma \text{ is a safe path from } s \text{ to } t|H_1) = p^2(1-p)^2(1-r), \tag{33}$$

and

$$P_{\Gamma}(\Gamma = \Gamma(\beta) \text{ for some } \beta \in \mathscr{U}|H_1) = \frac{1-r}{2}. \tag{34}$$

Therefore, $\pi_0 = \frac{1}{2}$ and $\pi_1 = \frac{1-r}{2}$. Since $r > 0$ then $\pi_0 > \pi_1$.

Observe that if we had $r = 0$ then the induced distributions would have been the same under hypothesis and the alternative. But for sufficiently large r, we may be able to distinguish H_0 from H_1.

For instance, a one-sided test of the hypothesis $H_0 : \pi = \frac{1}{2}$ versus the alternative $H_1 : \pi < \frac{1}{2}$ can be performed at a significance level α using the p-value method:

1. Compute $p-\text{value} = P(t \leq T(x))$, where $T(x)$ is the value of test statistic for data sample x, and $t \sim \text{Bin}\left(n, \frac{1}{2}\right)$.
2. Reject H_0 if $p-\text{value} \leq \alpha$.

Notice that given a probability of type I error α, the *power of the test* β can be computed from the probability of type II error, which is

$$1 - \beta = P(t > C|t \sim \text{Bin}(n, \frac{1-p_e}{2})), \tag{35}$$

where

$$C = \max\left\{c : \sum_{k=0}^{c} \binom{n}{k} < 2^n\alpha\right\}. \tag{36}$$

4 Summary

In this chapter, we described both existing and new statistical approaches for cyber security. We highlighted the emergence of cyberspace and its importance for national security and economy. Telecommunication and computer network systems form a physical foundation of a cyberspace. Increasingly cyber attacks and

cybercrime pose a considerable threat to society. Therefore, security of cyber systems must be addressed. We presented three fundamental methodologies for cyber attack detection known in the intrusion literature. For each methodology, anomaly detection, signature recognition, and attack norm separation, we reviewed research publications that describe the representative statistical approaches for detecting cyber attacks. Cyber attacks continue to evolve. Distributed attacks are emerging that hijack and harvest the power of cloud technologies to cause disruption. System-wide approaches are needed to enable early detection of such attacks. We proposed a new graph-based modeling framework and derived a network-based statistical method that could help detect compromised structures in a cyberspace. Finally, we applied our new method on a small cyberspace represented by a simple graph on four nodes.

References

1. Ayutyanont, N.: Statistical Characteristics and Models of Cyber Attack and Norm Data for Cyber Attack Detection. Ph.D. Dissertation. Arizona State University, Tempe, AZ, USA (2007)
2. Calo, S., Wood, D., Zerfos, P., Vyvyan, D., Dantressangle, P., Bent, G.: Technologies for Federation and Interoperation of Coalition Networks. In: Proceedings of 12th International Conference on Information Fusion, Seattle (2009)
3. Casas, P., Mazel, J., and Owezarski, P.: Unsupervised network intrusion detection systems: Detecting the unknown without knowledge. Comput. Comm. **35**(7), 772–783 (2012)
4. Chora, M., DAntonio, S., Kozik, R., Holubowicz, W.: INTERSECTION Approach to Vulnerability Handling. In: Proceedings of 6th International Conference on Web Information Systems and Technologies, WEBIST 2010, vol. 1, pp. 171–174. INSTICC Press, Valencia (2010)
5. Choraś, M., Kozik, R. Network Event Correlation and Semantic Reasoning for Federated Networks Protection System. In Computer Information Systems Analysis and Technologies. Springer, Berlin Heidelberg, 48–54 (2011)
6. Department of Defense Strategy for Operating in Cyberspace. U.S. Department of Defense. July 2011. http://www.defense.gov/home/features/2011/0411_cyberstrategy/docs/DoD_Strategy_for_Operating_in_Cyberspace_July_2011.pdf. Cited 15 Apr 2013.
7. Dean, D. et al. The Internet Economy in the G-20. The Boston Consulting Group. March 2012. http://www.bcg.com/documents/file100409.pdf. Cited 15 Apr 2013
8. Emran, S.M., Ye, N.: Robustness of chi-squared and Canberra techniques in detecting intrusions into information systems. Qual. Reliab. Eng. Int.**18**(1), 19–28 (2002)
9. El-Damhougy, H., Yousefizadeh, H., Lofquist, H., Sackman, D., Crowley, R.: Hierarchical and federated network management for tactical environments. In: Proceedings of IEEE Military Communications Conference MILCOM, vol. 4, pp. 2062–2067 (2005)
10. ICT facts and figures: The World in 2013. ICT Data and Statistics Division. International Telecommunication Union. Geneva, Switzerland. February 2013. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf. Cited 3 Jan 2014
11. IC3 2012 Internet Crime Report Released: More than 280,000 Complaints of Online Criminal Activity. National Press Releases. FBI National Press Office. The Federal Bureau of Investigation. May 2013. http://www.fbi.gov/news/pressrel/press-releases/ic3-2012-internet-crime-report-released. Cited 30 Jun 2013.
12. Lee, W., Stolfo, S.J., Mok, K.: A data mining framework for building intrusion detection models. In: Proceedings of the 199 IEE Symposium on Security and Privacy. Anaheim, CA: IEEE Computer Society Press, pp. 120–132 (1999)

13. Li, X., Ye, N.: Decision tree classifiers for computer intrusion detection. J. Paralel and Distributed Comput. Practices, **4**(2), 179–190 (2001)
14. Li, X., Ye, N.: Grid- and dummy-cluster-based learning of normal and intrusive clusters for computer intrusion detection. Qual. Reliab. Eng. Int. **18**(3), 231–242 (2002)
15. Li, X., Ye, N.: A supervised clustering algorithm for mining normal and intusive activity patterns in computer intrusion detection. Knowledge Inform. Syst. **8**(4), 498–509 (2005)
16. Li, X., Ye, N.: A supervised clustering and classification algorithm for mining data with mixed variables. IEEE Trans. Syst. Man, Cybernet. **36**(2), 396–406 (2006)
17. Mujtaba, M., Nanda, P., and He, X.: Border gateway protocol anomaly detection using failure quality control method. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on (pp. 1239–1244). IEEE (June 2012).
18. NATO Network Enabled Feasibility Study Volume II: Detailed Report Covering a Strategy and Roadmap for Realizing an NNEC Networking and Information Infrastructure (NII), version 2.0
19. Perlroth, N., Hardy, Q.: Bank Hacking Was the Work of Iranians, Officials Say. The New York Times (2013). http://www.nytimes.com/2013/01/09/technology/online-banking-attacks-were-work-of-iran-us-officials-say.html?pagewanted=1&_r=1&. Cited 26 Feb 2013.
20. Simmross-Wattenberg, F., Asensio-Perez, J. I., Casaseca-de-la-Higuera, P., Martin-Fernandez, M., Dimitriadis, I. A., and Alberola-Lpez, C.: Anomaly detection in network traffic based on statistical inference and alpha-Stable modeling. IEEE Trans. Dependable and Secure Comput. **8**(4), 494–509 (2011)
21. The economic impact of cybercrime and cyber espionage. Center for Strategic and International Studies. McAfee, An Intel Company. July 2013. http://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime.pdf. Cited 20 Aug 2013.
22. Ye, N.: Secure Computer and Network Systems. Wiley, Chichester (2008)
23. Ye, N., Chen, Q.: An anomaly detection technique based on a chi-square statistics for detecting intrusions into information systems. Qual. Reliab. Eng. Int. **17**(2), 105–112 (2001)
24. Ye, N., Chen, Q.: Computer intrusion detection through EWMA for auto-correlated and uncorrelated data. IEEE Trans. Reliab. **52**(1), 73–82 (2003)
25. Ye, N., Li, X.: A scalable, incremental learning algorithm for classification problems. Comput Indust. Eng. **43**(4), 677–692 (2002)
26. Ye, N., Borror, C., Parmar, D.: Scalable chi square distance versus conventional statistical distance for process monitoring with uncorrelated data variables. Qual. Reliab. Eng. Int. **19**(6), 505–515 (2003)
27. Ye, N., Borror, C., Zhang, Y.: EWMA techniques for computer intrusion detection through anomalous changes in event density. Qual. Reliab. Eng. Int. **18**(6), 443–451 (2002)
28. Ye, N., Chen, Q., Borror, C.: EWMA forecast of normal system activity for computer intrusion detection. IEEE Trans. Reliab. **53**(4), 557–566 (2004)
29. Ye, N., Chen, Q., Emran, S.M., Xu, M.: Probabilistic techniques for intrusion detection based on computer audit data. IEEE Trans. Syst. Man, Cybernet. **31**(4), 266–274 (2001)
30. Ye, N., Ehiabor, T., Zhang, Y.: First-order versus high-order stochastic models for computer intrusion detection. Qual. Reliab. Eng. Int. **18**(3), 243–250 (2002)
31. Ye, N., Emran, S.M., Chen, Q., Vilbert, S.: Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE Trans. Comput. **51**(7), 810–820 (2002)
32. Ye, N., Giordano, J., Feldman, J.: A process control approach to cyber attack detection. Comm. ACM. **4**(8), 76–82 (2001)
33. Ye, N., Parmar, D., Borror, C.M.: A hybrid SPC method with the Chi-square distance monitoring procedure for large-scale, complex process data. Qual. Reliab. Eng. Int. **22**(4), 393–402 (2006)
34. Ye, N., Zhang, Y., Borror, C.M.: Robustness of the Markov-chain model for cyber attack detection. IEEE Trans. Reliab. **53**(1), 116–123 (2004)
35. Ye, N., Li, X., Chen, Q., Emran, S.M., Xu, M.: Probabilistic techniques for intrusion detection based on computer audit data. IEEE Trans. Syst. Man, Cybernet. **31**(4), 266–274 (2001)

System Safety Analysis via Accident Precursors Selection

Ljubisa Papic, Milorad Pantelic, and Joseph Aronov

Abstract Starting from a systems-based statistical safety analysis, it has been noted that safety analysis in the period of operations is performed mainly in situations when accidents have not happened. Using the FMECA method, initial events screening (exclusion from the list of initial events whose effects are not significant) was performed. Benefits of the FMECA of excavator unit enable initial events list reduction for performing the ETA. The ETA methodology allows the determination of accident scenarios on the basis of initial events, for the excavator unit. One of the initial events represents the excavator unit's disturbance, the so-called accident precursor. Safety control chart was created from which the worst initial event was determined. A new safety assessment adoption in system operation stage based on the disturbance sequence (accident scenario) modeling is proposed.

Keywords Statistical safety analysis • Excavator unit • Initial events screening • Accident scenarios • Safety accident precursor • Safety control chart

L. Papic (✉)
Faculty of Technical Sciences, Svetog Save Str. 65, 32000 Cacak, Serbia

DQM Research Center, P.O. Box 132, 32102 Cacak, Serbia
e-mail: dqmcenter@open.telekom.rs

M. Pantelic
Kolubara Metal Company, Dise Djurdjevica Str. 32, 11560 Vreoci, Serbia

Faculty of Technical Sciences, Svetog Save Str. 65, 32000 Cacak, Serbia
e-mail: milorad.pantelic@kolubarametal.com

J. Aronov
Federal Scientific Research Institute of Certification (VNIIS), Elektricheskiy pereulok 3/10, 123557 Moscow, Russia
e-mail: aiz@gost.ru

C. Vogiatzis et al. (eds.), *Dynamics of Information Systems*, Springer Proceedings in Mathematics & Statistics 105, DOI 10.1007/978-3-319-10046-3_10

1 Introduction

Energy is necessary for development. However, methods used for manufacturing and energy usage are characterized by serious financial and ecological limitations. For the countries of the Northern Earthly hemisphere, today's question consists of providing a quality development. For the countries of the Southern Earthly hemisphere the problem consists of development itself, which is in connection with the risk of impossibility to expand the progress model onto the entire planet which has been accepted by the minority [1].

The damage (loss) that is inflicted to the environment and human health by manufacturing and energy usage could, usually, be classified into several categories such as:

- accidents,
- water pollution,
- radioactive radiation,
- radioactive waste,
- air pollution,
- the greenhouse effect,
- forest massif destruction,
- alienation (expropriation) of the land,
- transportation.

The bases of contemporary industrial manufacturing are natural resources. Approximately 70 % of natural resources are mineral raw materials. Over 1,000 billion dollars are spent annually in the world today for mineral raw materials (metal ores, nonmetal ores, coal, clay, stone, sand, and gravel) which for many of the countries represent the main import and export items (products) [2].

As an economic branch, mining in many countries represents the basis of development and has a great impact on its entire economy. According to [2], the following data are the best illustration about how much the mining economy has influence on a country's economy:

- the cost of coal and ore has 80–90 % share in the cost of iron,
- the cost of ore and electric power has 90 % share in the cost of light metals,
- the coal has 60 % share in the cost of electric power,
- maintenance of mining equipment on open-pit mine has 35–40 % share in the cost of coal.

Planning, designing, building, and exploitation, along with mining equipment maintenance on open-pit mines, carry a large number of occurrences which could reduce accident as well as damage and could endanger the health and life as of people directly engaged in equipment as well as the wider environment. In short, there is a high risk level of unwanted event's occurrence and their consequences. The failure of any part of open-pit mines' mining equipment, as complex technical systems that have great mutual dependence of its subsystems and elements, could

automatically mean outage of the entire system, or the operation with reduced power, which could also consequently increase cost of equipment operation, thermal and other overloads as well as other larger damages. For these reasons it is necessary for this complex system to be both reliable and safe in operation.

Risk causes of mining equipment on open-pit mines are various and numerous, and in the phase of designing the consequences of critical failure modes have to be minimized through the prediction of protective devices during the system operation. Risks of this equipment include: impacts, vibrations, corrosion, environment, fire, mishandling, etc. Their occurrence is in connection with all the equipments as well as with all the processes and decisions made during the equipment's life cycle.

The risk, as a combination of frequency or probability occurrence and a consequence of certain "danger" event, has two aspects: quantitative (calculated on the basis of known probability of event occurrence and consequences) and qualitative (in connection with human perception, i.e., depends on a person's emotional state). The use of management rules and certain procedures in order to identify, perform analysis, assess or estimate the risk, as well as its monitoring and reporting, represents the management of risk.

The safety of mining equipment on open-pit mines could be considered from two aspects. The first and the most important aspect is protecting the operator (human) from injuries during equipment operation. The second aspect is equipment protection from damages caused by the action of external causes. A need for cost reduction of mining equipment exploitation of open-pit mines, along with realization of the required safety level, also requires continual development of reliability and safety analysis discipline [3].

2 Overview of the State-of-the-Art Within the Chosen Research Area

A production increase on open-pit mining has required the mining equipment manufacturer to: increase the capacity, increase the excavation altitude, decrease the working bulks, better adjust to the mining-geological, hydrogeological, and climate conditions, increase the reliability and safety, improve the working comfort for personnel that operate and maintain the equipment, etc.

Intensive development of excavator types and models took place between 1960s and 1970s. In that period of time a great number of excavators have been delivered with expressed tendency towards optimization of the main technical characteristics such as: capacity, grasp altitude, range, surface pressure on the ground, increase of cutting force, etc. Then, i.e., in 1978, the Orenstein and Koppel factory (O&K) has manufactured the largest rotary excavator for the demand of Raynbrown mine (lignite open-pit mine of Hanbach), Germany. Today, seven of these excavators are in use. The main parameters of rotary excavator RB 289, which are illustrated in the Fig. 1, are:

Fig. 1 Rotary excavator RB 289, by manufacturer Orenstein and Koppel (O&K)

- theoretical capacity of 19.120 m^3/h,
- rotary boom length of 70.5 m,
- rotor diameter of 21.5 m,
- bucket number of 18 pieces,
- bucket bulk volume of 6.34 m^3,
- operating wheel drive engine power of 3,360 kW,
- operating bulk of the excavator of 13,265 t.

Demands of surface coal exploitation are more and more directed towards excavators of relatively large capacities, very mobile, easy for assemblage and dismantling, easy for maintenance, with short delivery deadlines, and lower life cycle cost. Today, such mining equipment for open-pit mines is strong, complex, and on a high technological level and demands a high level of operating and maintenance capability as well as high level of reliability and safety.

According to [4], the following data and numbers directly or indirectly speak upon the safety of open-pit mining equipment in the USA and Great Britain:

- during the period of time between the years 1991 and 1999, in average, there have been 21,351 cases of miners' injuries annually,
- during the period of time between the years 1983 and 1992, in stone pits of Great Britain there have been a total of 81 great accidents with human casualties.

The maintenance cost is in direct relation with the choosing of maintenance concept for a certain excavator. The largest part of working cycle of an excavator is hidden inside: low reliability, lost capacity (approximately 50 %), endangered life environment, state of accident (SAC) of the excavator (safety risks), decrease of working results, power loss, uneconomic supplying of consumable material, outsourcing service and maintenance, etc.

During the period of 2 years, between 2010 and 2012, on the open-pit mine Kolubara, Lazarevac, Serbia, there have been eight excavator accidents, which detailed description is given in Table 1. These data from internal maintenance database [5] show that in total 456 days have been lost within observing period. Converted into money, there has been a loss of 70,238,600.00 Euros in total. This amount of money could purchase two new rotary excavators class Srs 2000 by TAKRAF, Germany, manufacturer.

During the years, in Serbia multiple investigations have been executed towards determination of causes of open-pit mines mining equipment accidents. One such investigation was executed by DQM Research Center, Prijevor, Serbia, and Kolubara Metal Vreoci Company, Serbia [6]. The report has shown that there are seven main causes of accidents on excavators as shown in Table 2.

Detailed analysis of these accidents shows that human errors are significant factors in areas of operation, maintenance, and safety of mining equipment on open-pit mines. Human errors (by operator, mechanic) have an important role in accident occurrence, for which the data are shown in Table 2. These data show that human errors make more than 30 % of initial events of excavators accident causes. Here error implies "failure mode" of human, personnel, which is not in connection with work stoppage or sabotage.

More detailed on human reliability analysis has been given in literature [7]. Data on human errors are in special databases which were formed according to the certain system type of exploitation results. Considering human factor, risk analysis by implementing the event tree method, gives significant benefits for statistical safety analysis.

3 Problem Formulation

Open-pit mines dispose of various and numerous technological equipments in which units have large overall dimensions and masses, which, during the exploitation, and because of series connection in reliability block diagram, require very well maintenance. Because of structural complexity and their specific maintenance tasks, the excavators and disposers, as parts of technological equipment on open-pit mines, are usually called by a common name: excavator units [8]. What maintenance task of excavator units on open-pit mines makes even more complex is necessity use of many other machines beside the basic mechanization. Those are the machines of auxiliary mechanization: bulldozers, pipe layer, means of transportation, lifting devices (cranes), and many others whose technical nature and thereby also their

Table 1 Accidents on open-pit mine Kolubara, Lazarevac, Serbia, between years 2010 and 2012

No	Accident description	Year	Stoppage duration (days)	Direct cost (EUR)	Indirect cost (EUR)	Total cost (EUR)
1	Collision of two dragline excavators ES 6/45, by manufacturer NKMZ, Ukraine	2010	30	30,000.00	420,000.00	450,000.00
2	Accident of toothed ring on rotary excavator's SchRs 1760 No IX rotary motion mechanism by Krupp, Germany, manufacturer	2011	25	18,000.00	3,360,000.00	3,378,000.00
3	Accident of band 3 on rotary excavator Srs 1200 No V by TAKRAF, Germany, manufacturer	2011	20	27,000.00	2,688,000.00	2,715,000.00
4	Accident of frame 1and rotor boom connection on rotary excavator SRs 1200 No III by TAKRAF, Germany, manufacturer	2011	12	10,000.00	16,128,000.00	16,138,000.00
5	Accident of rotor boom on rotary excavator Srs 1200 No III by TAKRAF, Germany, manufacturer	2011	335	1,000,000.00	45,024,000.00	46,024,000.00
6	Accident of pulley for rotor boom lifting on rotary excavator SRs 2000 No 2 by TAKRAF, Germany, manufacturer	2012	6	12,000.00	806,400.00	818,400.00
7	Accident of excavating reduction unit carrier on rotary excavator Srs 2000 by TAKRAF, Germany, manufacturer	2012	3	8,000.00	403,200.00	411,200.00
8	Accident of mast on dragline excavator ES 6/45 No 103 by NKMZ, Ukraine, manufacturer	2012	25	24,000.00	280,000.00	304,000.00

Table 2 Causes of excavators accidents

Cause of accident	Accident share (%)
Difficult exploitation conditions	27
Error in manufacture and assembly	22
Operator's error	18
Mechanic's error	13
Fatigue of materials, wear of equipment, and corrosion processes	8
Inadequacy design	7
Other miscellaneous factors	5

maintenance tasks are significantly different from excavator units. Therefore it could not be spoken upon some common concept of open-pit mines' machines maintenance, but every group of machines, from the aspect of maintenance concept, has to be approached differently. One approach is valid for excavator units and other for auxiliary excavators, bulldozers, belt conveyor transporters, and all the machines which many open-pit mines own.

Solving problems of maintenance concept on open-pit mines significantly complexes the economical and technical need for certain maintenance tasks to be entrusted, through outsourcing, to aside companies. This particularly applies to manufacturing of spare parts, but also for many other works.

The failure modes of excavator units on open-pit mines could be various, their causes, even more versatile. Disturbances in the functioning of excavator units could manifest through exceeding increase or total intermission of coal supply to the client. Consequences of excavator units' failure modes depend on the place and role of failed elements within the structure of the excavator unit.

Causes of excavator units' failures represent stochastic occurrences because they depend on the sequence of certain but random factors which influence could mainly not be entirely perceived. Preventive measures in a way could plan activities of repression and possible reactions on such a group of factors. Possibility of excavator units operation without a failure in stationary and no stationary operation mode, economic and technical maintenance availability as of elements so of the system as a whole, restrictions that follow excavator units' exploitation (power use, environment protection, exploitation cost etc.), possible use of corresponding type solutions on the basis of analogy with similar facilities, normative for state control and failure diagnostics–these are all the problems which do not have detailed calculated and experimentally reasoned basis which relates to reliability and safety of excavator units.

On the other hand, by increasing complexity of excavator units, a problem of their optimal functionality occurs as a concomitant problem, especially if it is known that such systems could often cause large economical losses. The practice shows that 1 h of excavator unit exploitation stoppage of these systems costs:

- 10,000.00–12,000.00 Euros/Hour for coal excavation,
- 8,000.00–10,000.00 Euros/Hour for tailings excavation.

Basic problems carrying out statistical safety analysis involve selecting the most hazardous scenarios which have the biggest impact on excavator unit risk assessment. Statistical safety analysis is carried out by a team of experts which, as a rule, consists of: designers, process engineers, mechanics, and experts in engineering statistics and statistical safety analysis. Their role is working out a certain calculation which enables establishing possible scenarios for the development of accidents, as well as assessment of the excavator units' accident effects. Any kind of accident condition results in damaging personnel health and life and great economic losses resulting from the cost of revitalizing and restarting up of the excavator units. Well carried out statistical safety analysis and risk assessments of excavator units, which are considered in this paper, are preconditions for choosing a suitable maintenance concept.

4 The Role of Statistical Safety Analysis During the System Operation Phase

Events related to safety are divided to as follows: disturbances, difficulties, and accidents. In practice, during the system operation, the most common are disturbances, rarely difficulties and very rarely accidents. This setting illustrates the data that relate to the rotary excavators. On the basis of reliability testing for the rotary excavator SRS 1200×24/4×0 (400 kW) + VR, results have been found out in terms of failure modes frequency (disturbances and difficulties) at the rotary excavator items, which is shown in Fig. 2. The reliability testing results [6] show that during the observed period of one calendar year for all items of the rotary excavator, there have been 373 occurrences of disturbances and difficulties and but not one accident. Also, the reliability investigation shows that *mechanism for hoist of rotor's arrow (MHA)* is one of the most reliable items.

Analysis of this result means that the number of disturbances during the system operation significantly exceeds the number of accidents. Disturbances carry a lot of information about the system safety, and because of that, their analysis, including statistical analysis, enables effective detection of accident factors. As outlined in the guideline of International Civil Aviation Organization (ICAO) for the prevention of aircraft accidents the most important characteristics of information about the disturbances are:

- disturbances and accidents similarity, which allows detection of the same accident factors, as well as during accidents, excluding the severe consequences,
- disturbances occur much more frequently than accidents (according to some estimates, 10–100 times more) and because they represent a comprehensive data sources about accident factors.

Thus, in the operation phase, the most detailed is an information channel that includes data on disturbances of the operating conditions, whose number can be

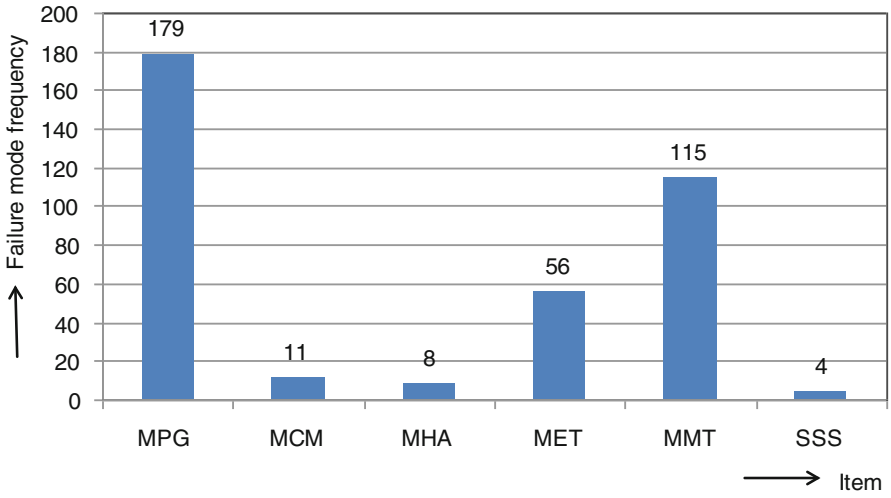

Fig. 2 Failure mode frequency per item of rotary excavator. SRS 1200×24/4×0 (400 kW) + VR. *MPG* mechanism for pawingthe the ground, *MCM* mechanism for circular motion, *MHA* mechanism for hoist of rotor's arrow, *MET* mechanism for excavator transport, *MMT* mechanism for material transport, *SSS* supporting steely structure

more than the number of serious difficulties and accidents. Thus for the operational management of safety, it is useful to direct attention towards the analysis and the results of the disturbance data. Besides, the analysis should consider that in the conditions of limited time and financial resources the need and the urgency of corrective actions depends on the degree of the hazard disturbance. This causes a justification for the determination of indicators, which characterize the severity of the disturbance, and appropriate assessment methods for the introduced indicators according to the operational data.

It should be emphasized that the ordinary reliability characteristics, which are used for safety analysis in the system operation phase, are not sufficient for a complete description of safety, because they do not define the severity of the system in operation. However, the possible proximity (similarity) level of disturbance and accident represents an important characteristic. Therefore, when it is performed, safety assessment for the system in operation is done, for example, based on the middle breakdown specific number n, from the fact that the number of disturbances n_A for a system A is smaller than the number of disturbances n_B for the system B ($n_A < n_B$), it does not necessarily mean that system A is safer than system B, since it may be that the level of proximity (dislocation) disturbance of the system A to accident is much higher than at the system B.

Accordingly, the traditionally used reliability characteristics are obviously insufficient for a complete and correct description of a system safety in operation. This supports the introduction of new indicators for the safety system evaluation in the phase of operation.

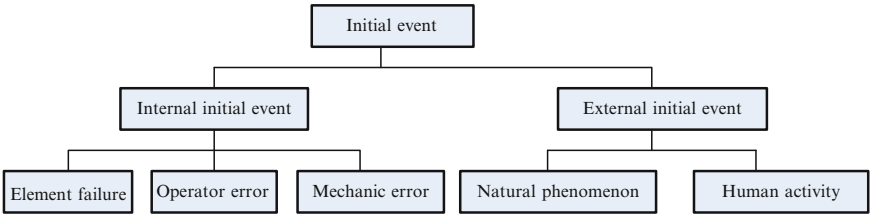

Fig. 3 Classification of initial events

5 Initial Events Analysis

At this stage, first an initial list of possible events (potentially hazardous from the aspect of damage occurrence), which exceeds the allowed level, is made, and from this list, a selection is made of the initial events group which is later used for modeling by means of event tree construction. Carrying out of this stage is necessary in order to reduce the selected scenarios of possible accidents.

When making the list of initial events, internal, and external initial events, the internal and external initial events should be separated. Internal initial events are caused by system items failures, operator's incorrect activities, or mechanic's errors, while external events are caused by influences connected with natural phenomena or human activities in the territory (region) where the system is located (earthquakes, winds, floods, terrorist attacks). The classification of the initial events is shown in Fig. 3.

As the starting data for carrying out this stage, accident analysis of similar systems is used. The importance of work in this stage is conditioned by the need for safety assurance not only in the period of normal exploitation, but when the initial event occurs [8].

6 Initial Events Screening

Safety analysis in the system operation phase is often done in situations when no accident has occurred. To a system an accident leads to an unwanted event (mode of operator error, mode of mechanic error, system failure mode), with disastrous consequences, and when the accident occurs, it is usually too late for an analysis. Therefore, system safety analysis is implemented with prevention. For the safety analysis, data on the history of events in the system are not required, which is the case in the reliability analysis being the main purpose. Therefore, the safety analysis provides prediction of unwanted events with catastrophic consequences, i.e., the analysis is used for prognosis of system accident occurrence.

Using the FMECA method [9] for the rotary excavator SRs 1200×24/4×0 (400 kW) + VR during the operation of overburden on the open-pit coal mine

of Field D, Kolubara, Serbia, was performed screening, i.e., exclusion of initial events (mode of operator error, mode of mechanic error, system failure mode) whose effects are not significant, was performed. In this way, the results from the application of the FMECA method application have helped reduce the number of initial events list for the event tree analysis from 373 to 8. The FMECA procedure in [6] has showed that the highest level of criticality (highest value of the risk priority number) is for the MHA at the rotary excavator SRs 1200×24/4×0 (400 kW) + VR, because the greatest degree of criticality is in the mode of operator error, mode of mechanic error, and system failure mode with corresponding causes. On this basis, it is concluded that the safety criterion of the rotary excavator SRs 1200×24/4×0 (400 kW) + VR is the accident of MHA, because in that case there is usually an accident (falling excavator to contra-weight) of the entire excavator.

7 Risk Calculation

If for a particular initial event I_0, we can select n scenarios of accident occurrence which are marked as: E_1, E_2, …, E_n, in this case accident may occur before the realization of n non simultaneous (random) scenarios of accident occurrence. Thus, accident is an event (in statistical sense) which represents a collection of non simultaneous (random) events E_1, E_2, …, E_n. So, the conditional accident probability is given by

$$Q(I_0) = \sum_{i=1}^{n} Q_i(E_i/I_0), \quad i = 1, 2, \ldots, n, \tag{1}$$

where $Q_i(E_i/I_0)$ is the probability of realization of the ith accident occurrence scenario for a particular initial event.

For calculating the total probability $R(I_0)$ of accident occurrence (unconditional), it is necessary to take into account the probability $P(I_0)$ of the initial event occurrence. In that case, according to the total probability formula, the accident probability $R(I_0)$ can be calculated when the initial event I_0 occurs:

$$R(I_0) = P(I_0) \cdot \sum_{i=1}^{n} Q_i(E_i/I_0) = \sum_{i=1}^{n} P(I_0) \cdot Q_i(E_i/I_0), \tag{2}$$

where $P(I_0)$ is the probability of the occurrence of the initial event I_0 for a certain period of time T, e.g., for 1 year. This probability is determined by using results of the initial events analysis.

The last expression presents the total probability formula which characterizes the unconditional (full) accident occurrence probability, i.e., accident risk R [10].

In practice, as initial events are very rare, for probability distribution of their occurrence for the time T, a Poisson's distribution can be taken:

$$P(v = m) = \lambda^m \cdot e^{-\lambda}/m!, \quad m = 0, 1, 2, \ldots, \lambda, \quad \lambda > 0 \tag{3}$$

which characterizes the occurrence probability of exactly m initial events in a time unit. Here λ is the intensity of the initial event occurrence which is measured by their average number in a unit time.

Supposing that $m = 1$, a $\lambda \cdot T \approx 0$ (which is justified for high reliable potential dangerous systems) it is obvious that:

$$P(v = 1) = P(I_0) \approx \lambda.$$

Thus, in formula (2) for calculating risk instead of the initial event occurrence probability, it is useful to change the rate (frequency) of its occurrence:

$$R(I_0) = \lambda \sum_{i=1}^{n} Q_i (E_i / I_0). \tag{4}$$

This substitution is connected with simpler risk defined as accident frequency in a time unit. Majority of quantitative safety analysis includes risk assessment exactly in this form. Apart from this, very often analysis of initial events relies on the information about frequency and not on probability of their occurrence.

On the other hand, values $Q_i (E_i/I_0)$, $i = 1, 2, \ldots, n$ are calculated according to the formula of simultaneous occurrence of independent events probability (in a set) which form a particular scenario of accident occurrence E_i. In other words, if E_i is a scenario of accident occurrence caused by k_i independent, in a set of events (items failures, personnel errors, items operation without failures) whose probabilities are equal to π_{ij}, then

$$Q_i (E_i / I_0) = \prod_{j=1}^{k_i} \pi_{ij}, \tag{5}$$

where $j = 1, 2, \ldots, k_i$ and $\pi_{ij} = p_{ij}$ is the probability of operation without failures or $\pi_{ij} = q_{ij}$ is failure probability.

It should be emphasized that assumption of independence within a group of events, which enter in accident occurrence scenario, is rather disputable. However, taking into account the dependence of events can make the calculation of probability $Q_i (E_i/I_0)$ much more difficult, that is why it is not considered here.

Calculated values $Q_i (E_i/I_0)$ are entered in the fifth column of Fig. 4. Apart from that, sometimes, it is useful to enter values of all events scenarios realization probability in this column. As an example, in Fig. 4 probability values of all possible scenarios previously classified in appropriate groups are given.

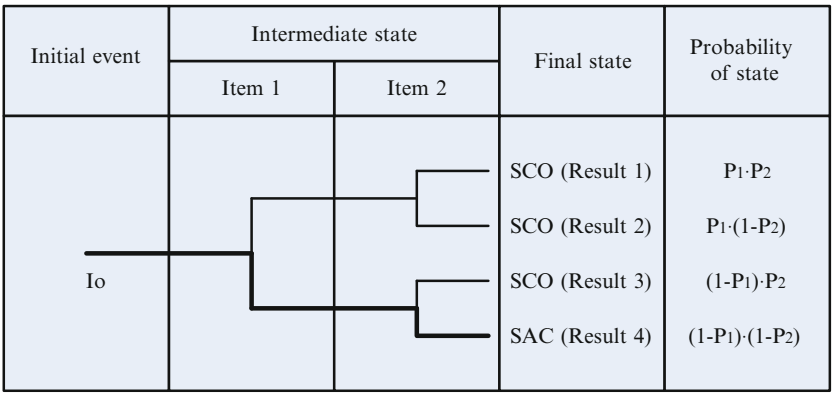

Fig. 4 Example of event tree with presentation final states probabilities. *SCO* state of capability to operate, *SAC* state of accident

The analysis of the fourth column in Fig. 4 shows that the number of accident scenarios equals unit ($i = 1$). In that case:

$$Q\,(I_0) = Q_1\,(E_1/I_0)\,. \tag{6}$$

On the other hand, conditional probability $Q_1\ (E_1/I_0)$ of accident scenario realization (failure probability of both items) is determined as:

$$Q_1\,(E_1/I_0) = (1 - P_1) \cdot (1 - P_2)\,. \tag{7}$$

Here, when calculating value Q, the factor of time is not taken into account (determined operation time) which has an important role when calculating probability operation without failure. It is obvious that if a set of final states matches with the full set of elementary events (within the limits of elementary probability theory), in that case, the sum of all final states probability equals to unity.

The risk accident value is calculated according to the formula (7) taking into account conditions (6):

$$R\,(I_0) = P\,(I_0) \cdot Q\,(E_1/I_0) = P\,(I_0) \cdot (1 - P_1) \cdot (1 - P_2)\,. \tag{8}$$

In complex cases, the event tree can be extended; thus, the analysis of risk calculation results [11] becomes complicated accordingly.

8 Event Tree Analysis

As the initial events in the event tree analysis, besides functional failure modes in the MHA, observed modes are of personnel errors (operators and mechanics). Based on data from the rotary excavator SRs 1200×24/4×0 (400 kW) + VR failure map [6] founds out the list of modes of operator error, modes of mechanic error, and failure modes of MHA:

1. List of modes of operator errors, $n = 1, 2$:
   (a) The operator often turns on mechanism for the hoist of rotor's arrow.
   (b) The operator often turns on mechanism for the hoist of rotor's arrow when the excavator is on ground level.
2. List of modes of mechanic error, $m = 1, 2, 3$:
   (a) Mechanic has not properly performed an assembly of the coupling at the small group generator.
   (b) Mechanic has not made centering of electric motors precisely.
   (c) Mechanic has not adjusted arrester for car interlocking.
3. List of failure modes of mechanism for the hoist of rotor's arrow, $k = 1, 2, 3$:
   (a) Breaking at the back gearbox shaft (front-end) for the hoist of rotor's arrow
   (b) Outage of electric-hydraulic lifter (releaser) at operating brake
   (c) Mechanical defect of ropes for the hoist of rotor's arrow

Event tree for the initial event—failure mode of mechanism for the hoist of rotor's arrow, $k = 1$: *Breaking at the back gearbox shaft (front-end) for the hoist of rotor's arrow*, is shown in Fig. 5.

The probability of occurrence of a SAC scenario realization is:

$$P\,(E_1/I_0) = (1-P_1)\cdot(1-P_2)\cdot(1-P_3)\cdot(1-P_4)\cdot(1-P_5)\cdot(1-P_6)\cdot(1-P_6)$$
$$= 0.175\cdot 0.145\cdot 0.155\cdot 0.135\cdot 0.125\cdot 0.165\cdot 0.175 = 0.1916\cdot 10^{-5}.$$

This result analysis presents the final stage of statistical safety analysis. Its content depends (to a great extent) on overall aims of statistical safety analysis. For example, risk calculation results enable solving problems:

- comparison of several system variants (according to safety criterion),
- showing of principal realization of required safety,
- choice of effective maintenance concept.

For solving the first problem it is necessary to compare risk values $R(I_0)$, calculated for several system variants, and choose the one where the risk value is minimal. Solving of the second problem is connected with comparison of calculated risk value $R(I_0)$ with criterion risk value. For solving the third problem it is special importance that with inadequate maintenance operation will not causing excavator

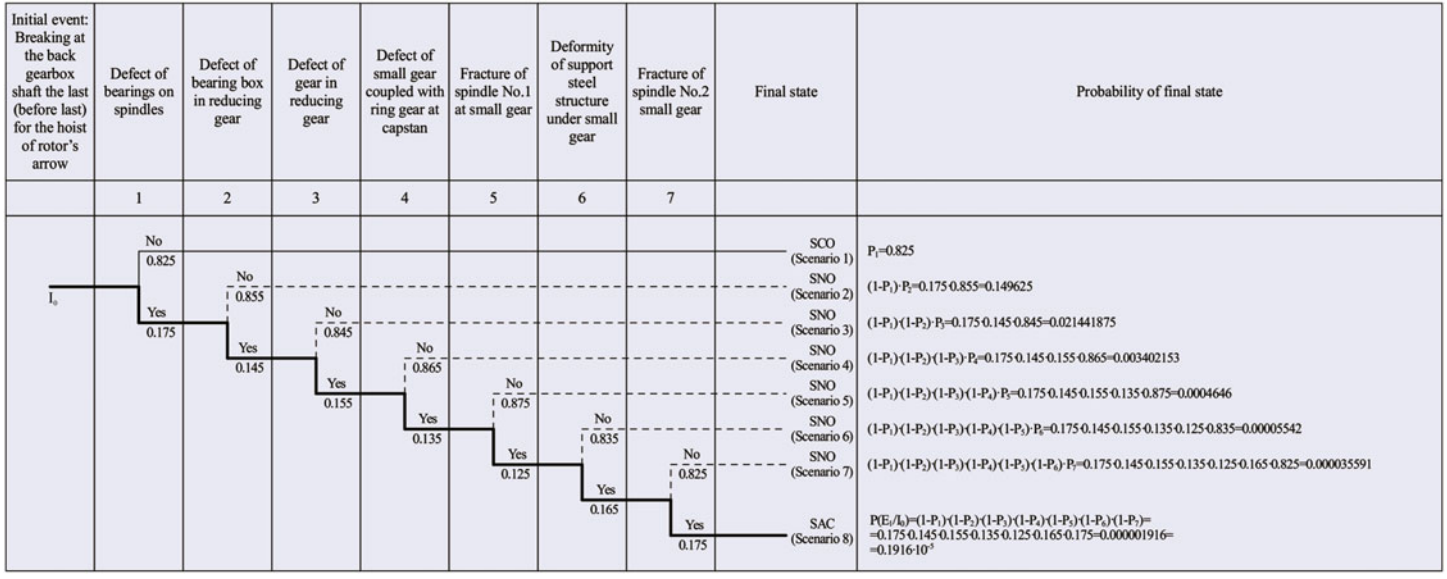

Fig. 5 Event tree for initial event breaking at the back gearbox shaft the last (before last) for the hoist of rotor's arrow. *No* without unwanted event, *Yes* unwanted event happened, *SCO* state of capability to operate, *SNO* state of noncapability to operate, *SAC* state of accident

unit's state of accident. Any SAC results in endangered health and life of personnel and great economic losses expressed through cost of reengineering and repeated starting of excavator units.

9 Systems Analysis in Operation on the Basis Safety Accident Precursors

One of the important aspects of complex safety systems analysis during operation is in connection with argumentation (explanation, demonstration) of the most serious (the most dangerous, hardest, the most important) disturbances for further production of effective corrective actions. This is the main problem within systems safety operational management in conditions of resource limitations. In that case, the body (department, office) of safety operational management or person which makes decision, first of all, should invest means in the removal of disturbances root causes, the most important ones for safety.

Disturbances of operation systems with high-value rating, i.e., with high-value probability of disturbance passing into accident, for some as certained period of operation, are called *accident precursors* [12]. Accordingly, previously established problem of extracting for systems safety important disturbances could be restated from the aspect of accident precursor determination. Introducing in practice safety disturbance analysis—accident precursor solves important problem of early accident prevention, i.e., safety prognosis.

Lack of limiting acceptable (criteria) value for disturbance rating does not allow introducing simple rule for making decisions about extracting accident precursor, based on comparison of disturbance rating values (e.g., punctuated appraisal of total disturbance rating) with criteria value. In that case it is useful to replace unknown criteria value S with some value S_{lim}, for which the numerical value is conditioned by the heaviness of recognized disturbances. Here, and further, the term "rating" describes any indicator marked by $S(t_i)$, whereby t_i is the moment (the trice) of ith operation disturbance [12].

Obviously, operation of any system, even if it has items of high reliability, beside the existence of impeccable instructions for operation and qualified personnel (operators and mechanics), is followed by random disturbances. Every disturbance could be described by certain rating values which are changeable within some limits. Such oscillations (fluctuation) of rating degree have random character. To avoid disturbances in systems operation with random character of oscillations is practically impossible. Upper limitation of such oscillations can be named *natural limitation* for rating.

In another words, it is assumed that all of the system samples, which affect disturbance rating value, stay unchanged in the condition which is being controlled and that rating fluctuation reflects only the influence of factors that are considered, which follow the process of alternate disturbance occurrence with corresponding rating values $S(t_1)$, $S(t_2)$, $\ldots$, $S(t_r)$, registered in time moments t_1, t_2, $\ldots$, t_r.

It is further assumed that $S(t_1)$, $S(t_2)$, …, $S(t_r)$ are forming the sample of rating value from the infinite basic set. Assuming that rating S is submissive to normal distribution, through the characteristics $\overline{S}$ *i* D_s (where $\overline{S}$ is mean value assessment, and D_s is rating dispersion assessment), calculated according to samples $S(t_1)$, $S(t_2)$, …, $S(t_r)$, such limit S_{lim} could be found at which, along with the confident level (confidential probability) $\gamma > 0.5$, the great part ($P > 0.5$) of rating value set would be guaranteed to come into (fall into) the interval $[0, S_{\text{lim}}]$.

For disturbance rating value $S(t_j)$ registered in time t_j ($j > r$) which does not fall into this interval, the justified condition is:

$$S\left(t_j\right) > S_{\text{lim}}. \tag{9}$$

Accordingly, with high probability γ, this disturbance registered in time t_j could be attributed to accident precursors, because the rating value of this disturbance significantly differs from the other rating values, in which the great part ($P > 0.5$) lies (is situated) under S_{lim}. In mathematical statistics the limit S_{lim} is called *upper acceptable (tolerant) limit* and for normal distribution is calculated by the following formula:

$$S_{\text{lim}} = \overline{S} + k \cdot \sqrt{D_s}, \tag{10}$$

where k is called *tolerant multiplier* and is determined by formula [2]:

$$k = U_p \cdot \left[1 + \frac{U_\gamma}{\sqrt{2r}} + \frac{\left(5U_\gamma^2 + 10\right)}{12r}\right], \tag{11}$$

where U_p is the standardized random variable of normal distribution for the probability P and U_γ is the standardized random variable of normal distribution for the probability γ.

Values of standardized random variables of normal distribution are determined from special statistical tables, whose entries are represented by probability values P or γ. For example, for $P = 0.9$ value, $U_p = 1.281$, and for $P = 0.99$ value, $U_p = 2.326$.

Values $\overline{S}$ and D_s are being calculated by the following formulas:

$$\overline{S} = \sum_{j=1}^{r} S\left(t_j\right)/r, \tag{12}$$

$$D_s = \sum_{j=1}^{r} \left[S\left(t_j\right) - S^2\right]/r. \tag{13}$$

Values of tolerant multiplier k for $P = \gamma = 0.8$; 0.9; 0.95; 0.99 and $r = 10$; 15; 20; 30 are given in Table 3. The analysis of this table shows that with the increase

Table 3 Values of tolerant multiplier

γ	P	$r = 10$	$r = 15$	$r = 20$	$r = 30$
0.80	0.80	1.096	1.035	1.002	0.965
	0.90	1.665	1.573	1.523	1.467
	0.95	2.140	2.022	1.957	1.886
	0.99	3.024	2.845	2.766	2.668
0.90	0.80	1.210	1.124	1.076	1.024
	0.90	1.842	1.710	1.637	1.558
	0.95	2.365	2.196	2.102	2.000
	0.99	3.345	3.105	2.973	2.828
0.95	0.80	1.316	1.205	1.143	1.075
	0.90	2.003	1.833	1.739	1.636
	0.95	2.572	2.353	2.234	2.100
	0.99	3.638	3.328	3.159	2.970
0.99	0.80	1.539	1.373	1.282	1.181
	0.90	2.342	2.089	1.950	1.797
	0.95	3.007	2.683	2.504	2.308
	0.99	4.252	3.794	3.540	3.263

of the number of inspections r at determined values γ *i* P, value S_{lim}—decreases. Accordingly, inequality value increases and according to that the number of possible accident precursors implemented into safety analysis increases. At increase of P and γ for determined r, the value S_{lim} increases. Accordingly, the number of possible accident precursors decreases.

The choice of values P and γ is conditioned by problems of safety analysis. For securing the system safety (analysis is performed with safety "reserve") values P and γ are useful to choose on the level 0.9 and higher.

The assumption of normal distribution of rating values is achieved approximately—under the condition that the variation coefficient, i.e., ratio of rating value dispersion and its expected value (mean value), is significantly smaller than 0.3.

In practice, the safety control chart [13], means showing the sequence of disturbances in the form of time series of values in a coordinate system disturbance rating. Theory of control charts proposed by Shewhart, in the year 1924, for statistical analysis of the technological process, is based on the separation of two kinds of causes of process variability.

The first cause of variability is random, due to the large number of random causes, which are always present and generally quite difficult to identify (diagnose). Each of these causes represents minor component of total variability, and none of them significantly contribute to the overall process variability.

Second cause of variability is related to certain events, phenomena, and processes, and as such can be identified (diagnosed) and removed without significant cost, as compared to the cost of identifying and removing random causes. These identified causes are considered particular. Control charts are used for the detection of these special causes.

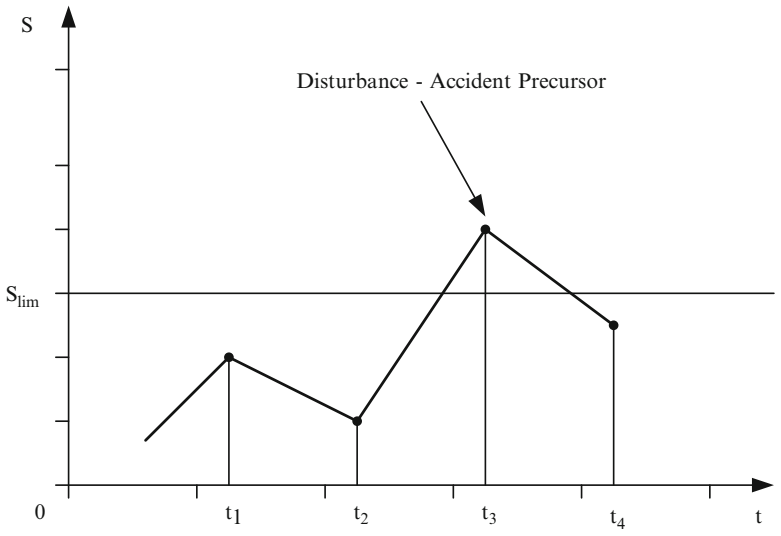

Fig. 6 Example of safety control chart

Regarding safety control chart, it could be told that disturbance rating, caused by natural random reasons (causes), with high probability ($P \gg 0.5$), should be less than the control limit of S_{lim}, while disturbance ratings in connection with some of the special root causes should be above S_{lim} value. Since these disturbances represent accident precursors, in that case it could be told that disturbances—accident precursors on safety control chart—will be illustrated by overflows of S_{lim} level.

In that way, the role of safety control chart consists of gathering data about disturbances and synoptic (visual, clear) disturbance extraction—accident precursor. Example of system safety control chart is shown in Fig. 6, on which the disturbance was extracted—accident precursor.

The order of introducing the control chart is as follows:

- for every system a safety control chart form is made which is memorized in corresponding data base,
- safety control chart form is shown in the shape of coordinate system which horizontal axis is inflicted by moments of normal operation disturbance onset, and vertical axis is inflicted by corresponding disturbance rating values,
- after expiration of certain time period T, the S_{lim} value is calculated which is supposed to be drawn into control chart as a line,
- control chart is being filled out according to disturbance emergence and according to the calculation of corresponding rating values,
- disturbances, which ratings exceed S_{lim} value with high probability ($P \gg 0.5$), are assorted as accident precursors,
- disturbance descriptions—accident precursors are memorized in the corresponding data base.

From practical point of view the important is the question of inspection (control, observation) period choice T (or number of disturbances, which is the same), necessary for further calculation of S_{lim}. Because of that a special engineering analysis of disturbance in period T should be performed, the period in which disturbances of system operation could be considered insignificant. The period T, necessary for the calculation of S_{lim}, is called founded. When the process of system exploitation is performed in conditions similar to the ones during assessment S_{lim}, in that case later values $S(t_j)$ should be under the limit S_{lim}.

Disturbance causes—accident precursor should be identified and, first of all, eliminated. Because of that the organ function of operational safety management should prepare corrective measures and to confirm its effectiveness.

The implementation procedure of safety control chart is shown in Fig. 7.

Forming the data base (list) of accident precursors enables the explanation (argumentation) of corrective actions for system safety increase, appraisal of these measures' effectiveness, and conduction of continuous monitoring of system safety. Since disturbance of normal operation does not bring to considerable damage, in that case, identifying disturbance—accident precursors in due time and removal of its causes allows prevention of accident emergence. Therein lies is the significance of safety control chart.

Besides, safety control chart is useful as obvious (visual) mean of system "safety history" for representing the information to the management and to the bodies (departments, offices) of monitoring.

In this way determined control limit S_{lim} could be used for the analysis of later periods of exploitation. While values S_j are within controlled limits, it could be highly likely considered that the process of system exploitation (from safety point of view) is in conditions which are being controlled (controlled conditions).

So, what is disturbance of normal exploitation—accident precursor, not from statistical, but from practical point of view? No doubt, for every kind of system disturbances, accident precursors have specific features (characteristics). However, some important universal features also could be emphasized, characteristic for all (or at least many) kinds of disturbances—accident precursors. First of all, they are often in connection with the interruption of exploitation rules (regulations), which points at lack of "safety culture" and to human errors [7].

Besides that, disturbances—accident precursors often in their development assume system safely protection activation. And finally, event trees of such disturbances contain within scenarios of accidents occurrence, in which realization probability is high compared to other scenarios of accidents occurrence.

Detailed analysis of disturbance—accident precursor enables the correction of exploitation instructions for considering its possible symptoms, which will in the future provide increase of exploitation quality.

The conclusion points out that variations (fluctuation) of disturbance rating values in limits $[0, S_{\text{lim}}]$ does not show any tendency (no trend), so it is not logical to speak of, for example, decrease of safety if three consequent rating disturbance values do not increase being situated in interval $[0, S_{\text{lim}}]$.

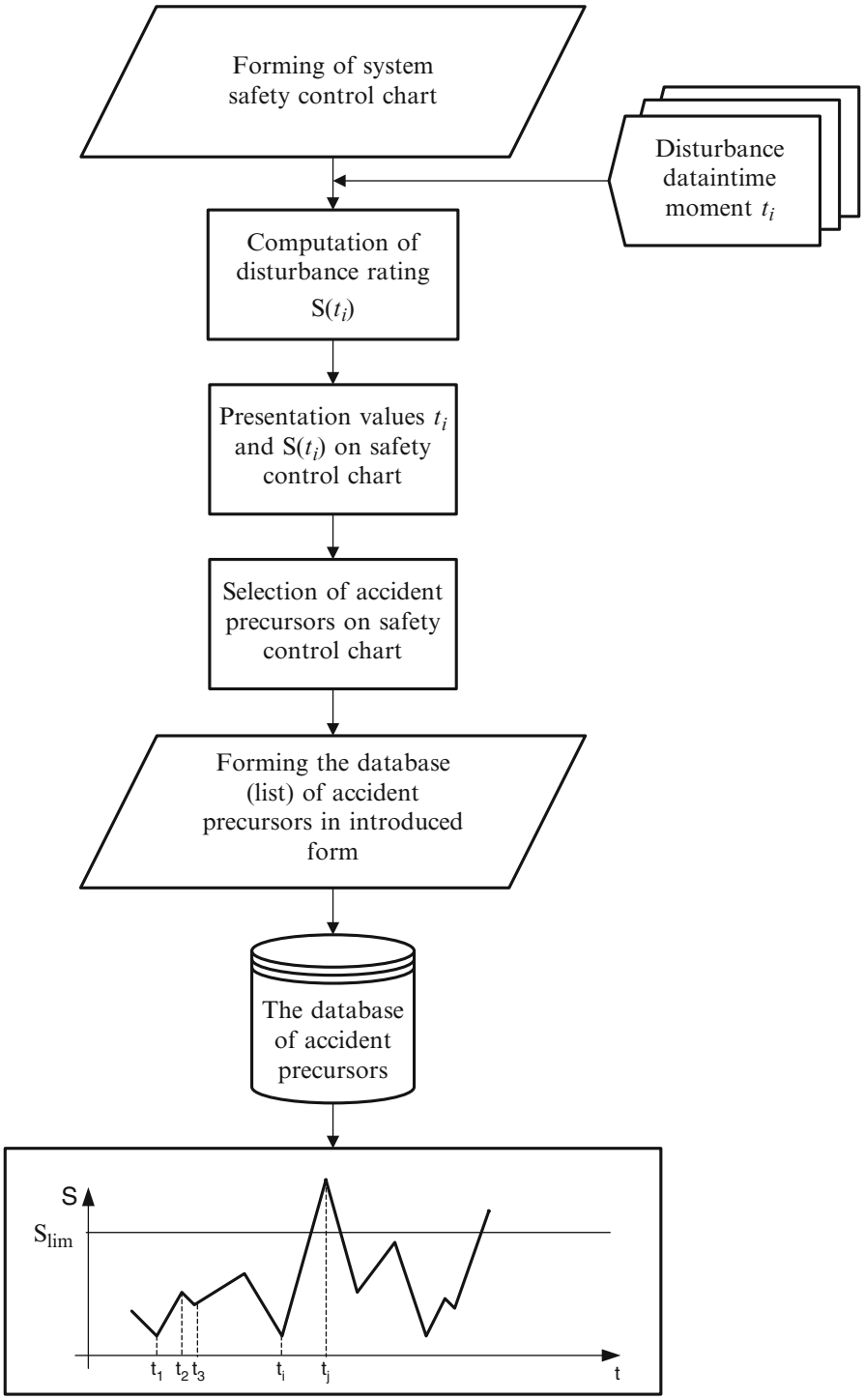

Fig. 7 Procedure of safety control chart implementation and of operational safety management on its basis

10 Safety Control Chart for Rotary Excavator SRs 1200×24/4×0 (400 kW) + VR

On the basis of obtained scenarios from a total of eight event trees ($n + m + k$), at determined (certain) conditions, some of the scenarios could get to the SAC of rotary excavators and so:

- from $n = 2$ event trees (initial event—modes of operator errors): 1 scenario,
- from $m = 3$ event trees (initial event—modes of mechanic errors): 2 scenarios,
- from $k = 3$ event trees (initial event—failure modes of mechanism for the hoist of rotor's arrow—(MDS)): 2 scenarios.

In that case it can be written the existence of

$r = 1 + 2 + 2$ initial events. Certain conditions could lead to the SAC of the rotary excavator.

For every initial event, probabilities of transition in accident in certain moment of time are:

t_1: 27 April 2006, breaking at the back gearbox shaft (front-end) for the hoist of rotor's arrow, $P_1 = 0.1916 \cdot 10^{-5}$,
t_2: 6 May 2006, operator often turns on mechanism for the hoist of rotor's arrow, $P_2 = 1.1 \cdot 10^{-15}$,
t_3: 29 July 2006, mechanic has not made centering of electric motors precisely, $P_3 = 1.65 \cdot 10^{-10}$,
t_4: 1 November 2006, mechanical defect of ropes for the hoist of rotor's arrow, $P_4 = 0.566 \cdot 10^{-6}$,
t_5: 23 November 2006, mechanic has not adjusted arrester for car interlocking, $P_5 = 1.1 \cdot 10^{-19}$.

Probabilities of initial event transition in SAC are shown in coordinate probability system—time, in Fig. 8. Safety control chart is obtained after calculation and drawing of control limit as (upper) limit value, P_{lim}, into coordinate probability system—time:

$$P_{\text{lim}} = \overline{P} + k \cdot \sigma_P,$$

where $k = 1.65$ is the standardized random variable for normal distribution and for confident level $\gamma = 0.95$ (Fig. 9), $\overline{P} = \frac{\sum_{i=1}^{5} P_i}{5} = \frac{P_1 + P_2 + P_3 + P_4 + P_5}{5} = 0,496 \cdot 10^{-6}$ is the mean value probability of transition in SAC rotary excavator, and $\sigma_P =$

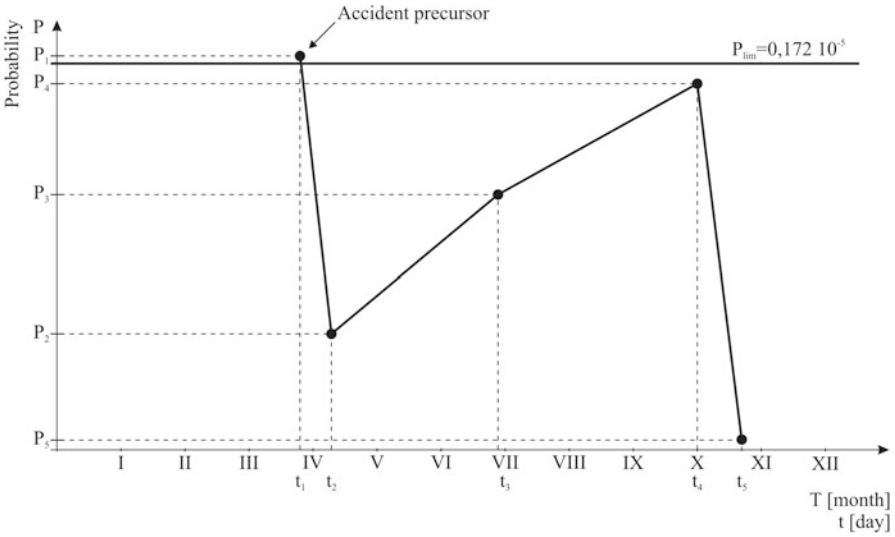

Fig. 8 Safety control chart construction for rotary excavator. SRs 1200×24/4×0(400 kW) + VR

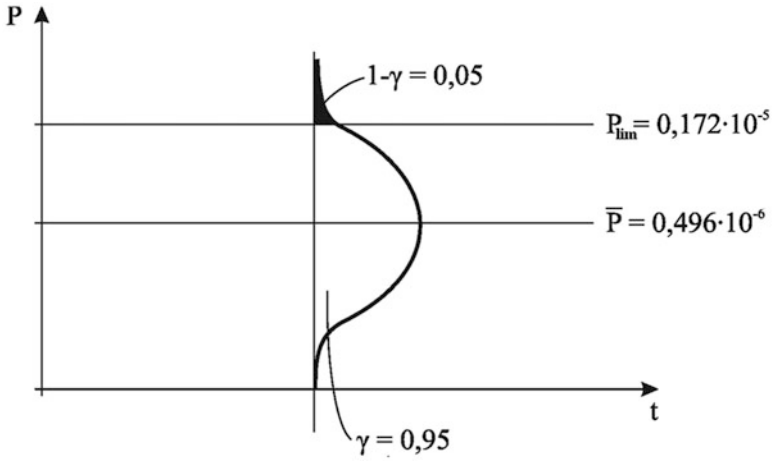

Fig. 9 With consideration of one sided confident level $\gamma = 0.95$

$\sqrt{\frac{1}{5}\sum_{i=1}^{5}(P_i - \overline{P})^2} = \sqrt{\frac{1}{5}\left[(P_1-\overline{P})^2+(P_2-\overline{P})^2+(P_3-\overline{P})^2+(P_4-\overline{P})^2+(P_5-\overline{P})^2\right]}$
$= 0,742 \cdot 10^{-6}$ is the standard deviation of probability transition initial event in the SAC rotary excavator:

$$P_{\text{lim}} = 0.172 \cdot 10^{-5}.$$

From safety control chart, it is obvious that the initial event *breaking at the back gearbox shaft the last (before last) for the hoist of rotor's arrow* is the most difficult from the safety point of view of rotary excavator, because this event has the highest probability (rating) of disturbance transition into SAC $P_1 = 0.1916 \cdot 10^{-5}$ (see Fig. 5), which exceeds limit value of rating $P_{\text{lim}} = 0.172 \cdot 10^{-5}$. Because of that this event is called *accident precursor* of the rotary excavator SRs $1200 \times 24/4 \times 0(400\,\text{kW}) + \text{VR}$.

Thus, the criterion of accident precursor selection is the probability of every SAC larger than limit value. In this case, accident precursor is the initial event, *breaking at the back gearbox shaft the last (before last) for the hoist of rotor's arrow*, for which the rating P_1 exceeds limit value rarely enough, i.e., rating values are below limit value with probability of $\gamma = 0.95$.

For every accident precursor, and even for *breaking at the back gearbox shaft the last (before last) for the hoist of rotor's arrow*, it is necessary to investigate and to eliminate cause of occurrence. For inspection of causes of accident precursor occurrence it is useful to apply the Ishikawa diagram causes-effects method from the aspects of: material, supplier, quality management, etc.

11 A Short Formulation of the Main Results

Results obtained in this paper contribute to increase knowledge which is significant for research in the area of safety statistical analysis in the phase of exploitation of excavator units on open-pit mines. Safety statistical analysis procedure has been given and its role has been explained in the phase of equipment exploitation (operation). Because of the proper implementation of complex risk category indicators, which describe safety during operation (exploitation), a model of disturbance progress is considered which was represented by right dichotomizing tree, i.e., the event tree [10].

After that the system safety analysis was performed on the basis of results of accident precursor selection. The basis for analysis is system exploitation disturbances with the largest rating values, i.e., with the largest probability values of transiting disturbance into an accident for a certain period of exploitation. The disturbance selection—an accident precursor was performed by using safety control chart. The safety control chart implies showing sequence (succession) of disturbances in the form of temporal array of disturbance rating values into the probability—time coordinate system.

The procedure of implementing safety control chart was presented along with forming the databases of accident precursor. That enables the explanation of corrective actions for equipment safety improvement, effectiveness assessment of these measures, and performing of constant monitoring of safety equipment.

It was emphasized that the safety analysis in the phase of excavator units' exploitation should be performed primarily in situations when accident has never happened. Therefore, safety analysis of excavator units is to be performed for

prevention. On that basis it was concluded that data on excavator units' events were not necessary for safety analysis, but is necessary for reliability analysis. The procedure FMECA was performed for rotary excavator Srs 1200×24/4×0 (400 kW) + VR, and results gained represent screening. Through the screening from the list, these initial events were excluded:

- modes of operator errors,
- modes of mechanic errors,
- failures of excavator's certain items,

for which the consequences are not significant.

The conclusion that ensues is that results of the FMECA procedure on rotary excavator maintenance enabled reduction of the initial events list for the event tree analysis from 373 to 8. On that basis comes the conclusion that safety criteria of rotary excavator SRs 1200×24/4×0 (400 kW) + VR represents the SAC of mechanism for the hoist of rotor's arrow, because in that case inevitable comes to an accident (excavator falling on to counterweight) of the entire rotary excavator.

At the end, the safety control chart was gained which shows that the initial event "breaking at the back gearbox shaft (front-end) for the hoist of rotor's arrow" is the hardest from the aspect of rotary excavator safety.

12 Conclusion

For a long time, the main scientific and practical discussions were oriented towards the achievement of the most important characteristics of improvement (effectiveness, capacity and speed increase, new materials, and technology development), without taking into account system accidents and disasters occurrence risk. This led to the fact that practically all industrially developed countries were showed unprepared for the difficult social, economic, and environmental consequences of accidents and disasters, increasing by number and consequences severity. At the same time, the humanmade systems which are doubtless hazard to people and the environment, in most cases, are created using traditional design principles (sequential design) and simplified engineering methods of tests planning (sequential engineering) [14].

This required, in the last decade of the twentieth century, the establishment of new principles and concepts of system safety assurance based on concurring engineering approach [15]. At the same time, undoubtedly, the basic requirement of safety assurance concept, consisting of accidents elimination, is generally accepted. In fact, the large system accidents cause maximum injury. On the other hand, the total accidents and disaster injury depends a lot on system item's failure mode. Therefore, the inclusion of adequate maintenance concept [16] principles in system safety assurance concept proved to be useful.

References

1. Smil, V.: Energy Myths and Realities: Bringing Science to the Energy Policy Debate. American Enterprise Institute for Public Policy Research, Washington, DC (2010). 272 pp
2. Pantelic, M., Papic, L., Aronov, J.: Maintainability and Safety Engineering of Excavator Units (in Serbian with Extended English Summary). DQM Research Center, Prijevor (2011). 289 pp
3. Dhillon, B.S.: Engineering Safety, Fundamentals, Techniques, Applications. World Scientific, Singapore (2003). 239 pp
4. Dhillon, B.S.: Mining Equipment Reliability, Maintainability and Safety. Springer, London (2008). 219 pp
5. Kolubara Metal Company: Internal Maintenance Data Base During Rotary Excavators Life Cycle at Open-Pit Mine. Kolubara, Lazarevac (2013)
6. Papic, L., Pantelic, M.: Implementation methodology for risk minimization into maintenance process of production system at coal mines. Report of Contract No. 4617 (In Serbian), 468 pp. DQM Research Center, Kolubara Metal Company, Prijevor-Vreoci (2009)
7. Dhillon, B.S.: Safety and Human Error in Engineering Systems. Taylor and Francis, Boca Raton (2013). 260 pp
8. Papic, L., Pantelic, M., Aronov, J., Verma, A.K.: Statistical safety analysis of maintenance management process of excavator units. Int. J. Automation Comput. **7**(2), 146–152 (2010)
9. Stamatis, D.H.: Failure Mode and Effects Analysis: FMEA from Theory to Execution. ASQ Quality Press, Milwaukee (2003). 487 pp
10. Zio, E.: An Introduction to the Basis of Reliability and Risk Analysis. World Scientific, Singapore (2007). 234 pp
11. Zio, E.: Computational Methods for Reliability and Risk Analysis. World Scientific, Singapore (2007). 362 pp
12. Aronov, J.: Methodology of operational safety management on the base disturbances statistical analysis during systems operation and assessment methods standardization (in Russian). Ph.D. Thesis, VNIIS, Moscow, 254 pp (1998)
13. Papic, L., Pantelic, M.: Maintenance: Oriented Safety Analysis, Keynote Addresses of 6th International Conference on Quality, Reliability, Infocom Technology (ICQRITITM 2012), Trends and Future Directions, Keynote Address, New Delhi, India, p. 40, 26–28 Nov 2012
14. Nachlas, J.A.: Reliability Engineering, Probabilistic Models and Maintenance Methods. Taylor and Francis, Boca Raton (2005). 403 pp
15. Kusiak, A. (ed.): Concurrent Engineering: Automation, Tools and Techniques. Wiley, New York (1993). 607 pp
16. Papic, L., Aronov, J., Pantelic, M.: Safety based maintenance concept. Int. J. Reliab. Qual. Safety Eng. **16**(6), 1–17 (2009)