Volume 86, Number 8,
August 2025

ISSN 0005-1179
CODEN: AURCAT

AUTOMATION AND REMOTE CONTROL

Automation and Remote Control

Vol. 86, No. 8, August 2025

http://ait.mtas.ru

Automation and Remote Control

ISSN 0005-1179

SCOPE

Automation and Remote Control is one of the first journals on control theory. The scope of the journal is control theory problems and applications. The journal publishes reviews, original articles, and short communications (deterministic, stochastic, adaptive, and robust formulations) and its applications (computer control, components and instruments, process control, social and economy control, etc.).

Automation and Remote Control is abstracted and/or indexed in *ACM Digital Library, BFI List, CLOCKSS, CNKI, CNPIEC Current Contents/Engineering, Computing and Technology, DBLP, Dimensions, EBSCO Academic Search, EBSCO Advanced Placement Source, EBSCO Applied Science & Technology Source, EBSCO Computer Science Index, EBSCO Computers & Applied Sciences Complete, EBSCO Discovery Service, EBSCO Engineering Source, EBSCO STM Source, EI Compendex, Google Scholar, INSPEC, Japanese Science and Technology Agency (JST), Journal Citation Reports/Science Edition, Mathematical Reviews, Naver, OCLC WorldCat Discovery Service, Portico, ProQuest Advanced Technologies & Aerospace Database, ProQuest-ExLibris Primo, ProQuest-ExLibris Summon, SCImago, SCOPUS, Science Citation Index, Science Citation Index Expanded (Sci-Search), TD Net Discovery Service, UGC-CARE List (India), WTI Frankfurt eG, zbMATH.*

Journal website: http://ait.mtas.ru

0005-1179/25. *Automation and Remote Control* (ISSN: 0005-1179 print version, ISSN: 1608-3032 electronic version) is published monthly by Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow 117997, Russia.

Volume 86 (12 issues) is published in 2025.

Publisher: Trapeznikov Institute of Control Sciences, Russian Academy of Sciences.

65 Profsoyuznaya street, Moscow 117997, Russia; e-mail: redacsia@ipu.rssi.ru; http://ait.mtas.ru, http://ait-arc.ru

Contents

Automation and Remote Control

Vol. 86, No. 8, 2025

Special Issue

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 697–698.

SPECIAL ISSUE

To the 90th Anniversary of Boris Polyak

(May 4, 1935–February 3, 2023)

DOI: 10.31857/S0005117925080018

This year commemorates the 90th anniversary of Boris Teodorovich Polyak, a brilliant mathematician and remarkable individual who passed away two years ago.

A pioneer in the theory and methods of optimization in the USSR, he gained worldwide recognition by the mid-1970s and worked for over half a century at the Trapeznikov Institute of Control Sciences. Polyak's research interests expanded far beyond optimization problems. He investigated a diverse range of topics with great enthusiasm, including chaos control, spacecraft stabilization at Lagrange points, peak effects in differential and difference equations, synchronization of oscillators, randomized versions of the PageRank problem, and the stability of power grids.

Polyak's research into mathematical optimization and control theory has exerted a strong influence on the development of these disciplines and has found applications, often unexpected, in practical problems.

Among the multitude of his remarkable results, we highlight some of the most prominent ones:

- the heavy ball method,
- the stochastic approximation method with averaging,
- the conjugate gradient method,
- nonconvex optimization algorithms under the Polyak–Łojasiewicz condition,
- adaptive step-size choice in subgradient methods,
- the sequential projection method,
- the conditional gradient method,
- Newton's method with cubic regularization,
- convexity/nonconvexity certificates for quadratic mappings,
- results on the parametric robustness of linear systems (including the Tsypkin–Polyak plot),

- new versions of random walk methods,
- results on ellipsoidal estimation and the invariant ellipsoid method.

The breadth of Polyak's scientific purview is well illustrated by the diversity of the journals in which he published:

Journal of Computational Mathematics and Mathematical Physics, *Sbornik: Mathematics*, *Journal of Optimization Theory and Applications*, *Mathematical Programming*, *SIAM Journal on Control and Optimization*, *Theory of Probability and Its Applications*, *Automation and Remote Control*, *International Journal of Robust and Nonlinear Control*, *IEEE Transactions on Automatic Control*, *Automatica*, *European Journal of Control*, *Systems and Control Letters*, and many others.

Polyak's personal traits—optimism, genuine humanity, as well as sharp and kind wit—always attracted colleagues and young researchers. Dozens of his students and their followers are now productively working at universities and research centers in many parts of the world. Furthermore, the Traditional Young Scientists School "Control, Information, and Optimization" founded by Boris, has been held with consistent success for over 15 years.

In honor of this anniversary, a special issue of *Automation and Remote Control* has been prepared. It includes articles by Polyak's students and colleagues from Russia (Moscow, St. Petersburg, Nizhny Novgorod, and Syktyvkar) as well as from England, Spain, Italy, France, and the United States. The scope of the special issue is exceptionally broad, ranging from classical optimization methods to probabilistic set approximations on the one hand, and from network flow optimization to robust estimation and robust control of linear systems on the other. Despite this wide range of topics, all of them are connected, in one way or another, to Polyak's research interests.

With deep gratitude, we list the authors who have kindly agreed to contribute their research to this special issue, and the reviewers whose expertise has been invaluable:

A. Ablaev, A. Akhavan, I. Akinfiev, T. Alamo, M. Alkousa, M. Balashov, Y. Bekri, R. Biryukov, S. Boyd, F. Dabbene, P. Dvurechensky, M. Fedotov, A. Gasnikov, O. Granichin, R. Hildebrand, A. Juditsky, M. Kogan, A. Lukashevich, M. Mammarella, N. Mashalov, V. Mirasierra, S. Nazin, A. Nemirovski, S. Parsegov, O. Savchuk, V. Sokolov, F. Stonyakin, E. Tarasova, T. Chikake Mapungwana, A. Tsybakov, D. Yarmoshik, and F. Zhang.

Special thanks go to A. Mazurov who put significant effort to translating the submitted manuscripts to and from English.

Similar to the 2024 special issue dedicated to Boris Polyak, three articles will be published in the next issue due to the monthly page limit of *Automation and Remote Control*.

Editor of the special issue
P.S. Shcherbakov

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 699–717.

SPECIAL ISSUE

A Search Method for Stochastic Non-Stationary Optimization of Functions with Hölder Gradient

I. A. Akinfiev[*,a], **O. N. Granichin**[*,**,b], **and E. Yu. Tarasova**[*,c]

[*] *Saint Petersburg State University, St. Petersburg, Russia*
[**] *Institute for Problems in Mechanical Engineering, Russian Academy of Sciences, St. Petersburg, Russia*
e-mail: [a] *i@iakinfiev.ru,* [b] *o.granichin@spbu.ru,* [c] *elizaveta.tarasova@spbu.ru*

Received June 23, 2025
Revised June 30, 2025
Accepted July 4, 2025

Abstract—We propose a gradient-free method of stochastic optimization with perturbation at the input which is designed to track changes in the minimum point of a function with Hölder gradient, with observations subject to almost arbitrary (unknown-but-bounded) noise. Similar methods are widely used in adaptive control problems (energy, logistics, robotics, goal tracking), optimization of noisy systems (biomodeling, physical experiments), and online learning with drift of the data parameters (finance, streaming analytics). The efficiency of the algorithm is tested under conditions that mimic tracking the evolution of human expectations in reinforcement learning problems based on human feedback when tracking the center of a cluster of problems in queueing systems. Search methods with input perturbations have been actively developed in the works by B.T. Polyak since 1990.

Keywords: tracking, input perturbations, randomization, stochastic optimization, gradient-free methods, reinforcement learning via human feedback, queueing systems, unknown-but-bounded disturbances

DOI: 10.31857/S0005117925080023

1. INTRODUCTION

The problem of minimizing a function (functional) $f(x)$ is at the heart of solving many practical problems, from control of engineering systems to machine learning. Closed-form solutions are often not available due to high dimensionality, nonlinearities, or the lack of an explicit form. Even when the function is defined explicitly, the practical applicability of the existing approaches is limited by computational resources, measurement inaccuracies, or rounding errors. Traditional iterative gradient methods are efficient when finding the minimum of smooth or differentiable functions. However, in real-world problems, situations often arise where computing the gradient is difficult or impossible. Typically, the objective function is subject to stochastic disturbances, or its explicit form is unknown. In practice, the optimized function is often defined by some oracle, and by making requests (function arguments) to this oracle, it is possible to obtain certain realizations. The availability of measurements of the gradient itself is feasible with the implementation of special measuring devices for specific tasks or through finite difference approximations, which are inefficient in the presence of a high-level noise in the obtained measurements. In such cases, alternative approaches are required that do not rely on the information about gradients.

A significant contribution to the development of the theory and methods of stochastic optimization was made by B.T. Polyak and his research group. Their research covers a wide range of issues, including gradient methods [1], pseudo-gradient adaptation and learning algorithms [2–4], and methods for accelerating convergence [5–7]. Even nowadays, the two papers [8, 9] provide

comprehensive answers when analyzing the convergence of general-type iterative stochastic algorithms in terms of mean-square deviations, as well as in the linear case in terms of error covariance matrices.

A new search method of stochastic approximation proposed in the 1990 paper [10] not only develops the overall direction of random search algorithms [11], but also significantly advances the entire general theory of iterative optimization algorithms. This paper shows that, if the observed values of the optimized function are corrupted by noise, the proposed algorithm has the asymptotically optimal rate of convergence in the sense that it is impossible to find a faster algorithm among all possible iterative optimization algorithms for a sufficiently broad class of functions. A similar algorithm was previously proposed in [12], and consistency of estimates generated by it was justified in the presence of almost arbitrary noise in the observations. In the English-language literature, similar methods have been called SPSA (Simultaneous Perturbation Stochastic Approximation), see [13, 14]. A salient feature of these gradient-free methods is that, regardless of the dimensionality of the problem, the oracle needs to be called only once or twice per iteration, with arguments being chosen over a randomly generated line through the current point (it is what is referred to as randomization of the algorithm). A detailed analysis of the history of development of search algorithms of stochastic approximation with perturbation at the input, as well as the properties of the estimates generated by these methods are provided in [15–17].

A limitation of classical iterative zero-order stochastic optimization methods (those which do not use the values of the gradient), such as the Kiefer-Wolfowitz procedure [18] in the multivariate case, is the need to repeatedly compute the function at each iteration. This becomes especially impractical in dynamical environments where the target function $f_n(x)$ changes over time. A similar situation arises, for example, in optimization problems related to real-time systems. It turned out that methods like the previously proposed search algorithms of stochastic optimization with input perturbation remain to be to be efficient in this situation when replacing the decreasing step-sizes over time with constant ones, [19, 20]. Later, it was possible to formulate and justify the properties of a distributed algorithm of this type, combined with a consensus algorithm [20].

In practice, [21, 22], statistical uncertainties are often encountered which do not have second statistical moment. For example, stable distributions, such as Levi–Pareto, are better at describing the prices of stocks and commodities than Gaussian distributions. In [24], the properties of the estimates provided by the SPSA algorithm under such conditions were studied. In the present paper, these studies are extended to the case of optimization of the non-stationary mean-risk functional.

2. STATEMENT OF THE PROBLEM

We consider discrete time $n = 0, 1, \ldots$, defined by the label of step (iteration), and we denote by $\{F_n(\cdot,\cdot)\colon \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}\}$ the set of functions in two vector variables, which are all differentiable with respect to the first argument. At every step n, observations

$$y_n = F_n(x_n, w_n) + v_n \tag{1}$$

are performed at known (chosen) points x_n (experimental design), where the w_ns are uncontrollable disturbances defined over a probabilistic space Ω and having identical unknown distribution $P_w(\cdot)$, and v_n is the (perhaps non-random) observation noise.

Let $\mathcal{F}_{n-1}$ denote the σ-algebra of all random events that have been realized up to the time instant n; $\mathbb{E}$ be the symbol of mathematical expectation; $\mathbb{E}_{\mathcal{F}_{n-1}}$ denote the conditional mathematical expectation relative to the σ-algebra $\mathcal{F}_{n-1}$.

We are interested in the minimization of the following nonstationary mean risk functional:

$$f_n(x) = \mathbb{E}_{\mathcal{F}_{n-1}} F_n(x, w) = \int\limits_{\mathbb{R}^q} F_n(x, w) P_w(dw) \to \min_x . \tag{2}$$

The goal is to evaluate the minimum point θ_n of the function $f_n(x)$; i.e., to find

$$\theta_n = \arg\min_x f_n(x).$$

Accuracy of the estimate x of the points θ_n is addressed through use of the scalar Lyapunov functions

$$V_n(x) = \|x - \theta_n\|^{\rho+1} = \sum_{i=1}^{n} |x^{(i)} - \theta_n^{(i)}|^{\rho+1},$$

where θ_n are the vectors to be found, and $\rho \in (0, 1]$ is the Hölder exponent for the gradients of the functions $V_n(x)$. In the sequel, we write $\|\cdot\|_{\rho+1}$ to denote the $l_{\rho+1}$-norm and $\langle\cdot,\cdot\rangle$ for the inner product in $\mathbb{R}^d$.

To characterize the behavior of the estimates of the minimum points of the non-stationary functional (2), we present two definitions.

Definition 1. The sequence $\hat{\theta}_n$ of the estimates of the minimum points θ_n is said to be $l_{\rho+1}$-*stabilized*, if there exists $C > 0$ such that

$$\mathbb{E}V_n(\hat{\theta}_n) \le C \quad \forall\, n.$$

Definition 2. The number L is referred to as the *asymptotic upper bound* for the estimation errors in the $l_{\rho+1}$-norm, if the sequence of estimates $\{\hat{\theta}_n\}$ of the minimum points θ_n satisfy

$$\overline{\lim_{n\to\infty}} \mathbb{E}V_n(\hat{\theta}_n) \le L < \infty.$$

In what follows, we construct the sequence of stabilizing estimates $\{\hat{\theta}_n\}$ in the spirit of Definition 2 under the following conditions satisfied for all $n > 0$:

(A) *The functions $f_n(\cdot)$ are strongly convex in the first argument:*

$$\langle \nabla V_n(x), \nabla f_n(x)\rangle \ge \mu V_n(x).$$

(B) *For all admissible w, the gradients $\nabla F_n(\cdot, w)$ satisfy the condition*

$$\|\nabla F_n(x, w) - \nabla F_n(y, w)\|_1 \le M\|x - y\|_\rho^\rho$$

for a certain constant M.

(C) *The local Lebesgue property:* For every point $x \in \mathbb{R}^d$ there exists a neighborhood U_x and a function $\Phi_x(w)$ such that $\mathbb{E}\Phi_x(w) < \infty$ and $\|\nabla F_n(x', w)\|_2 \le \Phi_x(w)\ \forall x' \in U_x$.

(D) *The rate of drift of the minimum point satisfies the following conditions:*

a: $\|\theta_n - \theta_{n-1}\|_1 \le A$;

alternatively, if $\{\theta_n\}$ is a sequence of random variables, then

$$\mathbb{E}_{\mathcal{F}_{n-1}}\|\theta_n - \theta_{n-1}\|_{\rho+1}^{\rho+1} \le A^{\rho+1},$$

b: $\mathbb{E}_{\mathcal{F}_{n-1}} \|\nabla_x F_n(x, w) - \nabla_x F_{n-1}(x, w)\|_1 \leqslant B\|x - \theta_{n-1}\|_1^\rho$,

c: $\mathbb{E}_{\mathcal{F}_{n-1}} \|\nabla_x F_n(\theta_n, w_n)\|_{\rho+1}^{\rho+1} \leqslant C$,

d: $\mathbb{E}_{\mathcal{F}_{2n-2}} |F_{2n}(x, w_{2n}) - F_{2n-1}(x, w_{2n-1})|^{\rho+1} \le DV_{2n-2}(x) + E$.

(E) *The observation noise v_n satisfies the condition*

$$|v_{2n} - v_{2n-1}| \le \sigma_v,$$

or

$$\mathbb{E}_{\mathcal{F}_{2n-2}}\{|v_{2n} - v_{2n-1}|^{\rho+1}\} \le \sigma_v^{\rho+1}$$

if it has random nature.

Note that the last condition is valid for arbitrary deterministic bounded sequences $\{v_n\}$. Condition (C) allows for interchanging the integration and differentiation operations when justifying the stabilizability of the estimates. Conditions of the form (D) cover both the random walk drift and directed drift in a certain direction. For instance, the following condition based on (D) is presented in [1]:

$$\theta_n = \theta_{n-1} + a + \xi_n,$$

where ξ_n is a zero-mean random variable, and a is trend. Stabilizability of the estimates generated by the algorithm under conditions (D) shows its applicability to a wide range of problems.

3. A SEARCH RANDOMIZED ESTIMATION ALGORITHM

Assume that the sequence $\{\Delta_n\}$ of trial simultaneous perturbations fed to the input of the algorithm is a realization of a sequence of independent Bernoulli vectors in $\mathbb{R}^d$ with components being independent random variables taking values $\pm\frac{1}{\sqrt{d}}$ with probability 0.5. Let us pick an initial vector $\theta_0 \in \mathbb{R}^d$. We will estimate the sequence $\{\theta_n\}$ of the minimum points by the sequence $\{\hat{\theta}_n\}$ defined by the following stochastic optimization algorithm with trial simultaneous input perturbations:

$$\begin{cases} \hat{\theta}_{2n-1} = \hat{\theta}_{2n-2} \\ x_{2n} = \hat{\theta}_{2n-2} + \beta\Delta_n,\ x_{2n-1} = \hat{\theta}_{2n-2} - \beta\Delta_n \\ \hat{\theta}_{2n} = \hat{\theta}_{2n-2} - \frac{\alpha}{2\beta}\Delta_n(y_{2n} - y_{2n-1}), \end{cases} \tag{3}$$

where α and β are the step-size parameters. To substantiate the stabilizability property of the estimates generated by algorithm (3), we adopt yet another assumption:

(F) *The random vectors Δ_n and w_{2n}, w_{2n-1} are independent of each other as well as of $\mathcal{F}_{n-1}$. If $\{v_n\}$ are assumed to have random nature, then Δ_n do not depend on v_{2n}, v_{2n-1}.*

4. STABILIZATION OF ESTIMATES

Denote $H = A + \alpha\beta M$, where A and M are constant bounds on the rate of drift and change of gradients, respectively.

Theorem 1. *Let conditions (A)–(F) be satisfied and let the parameters α, β be chosen in such a way as to guarantee the constant $K > 0$ defined later in the proof to be less than unity.*

Then, for any initial choice $\hat{\theta}_0$ with $E\|\hat{\theta}_0 - \theta_0\|^{\rho+1} < \infty$, the estimates generated by algorithm (3) *are being stabilized in the following sense:*

$$\overline{\lim_{n\to\infty}} \mathbb{E}\|\hat{\theta}_n - \theta_n\|_{\rho+1} \leqslant \left(\frac{L}{K}\right)^{\frac{1}{\rho+1}},$$

where L is also defined at the end of the proof.

Conditions (A)–(C) and (E)–(F) are standard when proving the consistency of estimates generated by stochastic optimization algorithms with input perturbations; see [18]. Mean-square stabilizability of the estimates provided by algorithm (3) has been earlier proved in [19] under more stringent assumptions.

Proof of Theorem 1 and the precise definition of the constants K and L are presented in the Appendix.

5. SIMULATION IN THE RLHF-SCENARIO

In reinforcement learning based on feedback from humans (Reinforcement Learning from Human Feedback, RLHF), a key challenge is working with noisy and unstable data, [26, 27]. Human evaluations often contain random errors and may change over time, hence complicating the optimization process. In particular, in tasks related to fine tuning of language models (Large Language Models, LLM), RLHF is used to improve the quality of text generation, align with user preferences, and minimize undesirable model behavior. However, the subjectivity and variability of human evaluations create significant difficulties for traditional optimization methods.

5.1. The Model

In the simulations, we examine the efficiency of the search algorithm under conditions close to reality; i.e., in the presence of heavy-tailed noise (Pareto distribution) and preference drift ([28, 30]), which mimics the evolution of human expectations. Three scenarios are considered: Moderate drift, near-stationary preferences, and stationary preferences with asymmetric noise. This allows for the assessment of stability and adaptability of the algorithm under RLHF conditions and checking its applicability to tasks related to LLM training and other systems where human feedback plays a key role.

The goal of simulations is to test the ability of RLHF agents to adapt to a reward model shaped from noisy and changing human evaluations. We then

— model heavy-tailed noise (Pareto distribution) describing uncertainty and rare but significant deviations in estimates;

— introduce a preference drift model that simulates the gradual change in human expectations;

— note that all functions and parameters are formulated in conditions (A)–(F) in Section 2.

Each agent has to minimize the discrepancy between its own estimate of the parameter and the true value set by the reward model, despite noise and dynamics of target preferences; a search algorithm is used for the minimization.

The RLHF-based reward model is specified as follows:

$$F_n(\mathbf{x}) = -\sum_{i=1}^{m}(x_i - x_n^*)^{1.35}, \tag{4}$$

where the target parameter x_n^* drifts in time n as

$$x_n^* = x_{n-1}^* + \delta, \quad x_0^* = 5,$$

thus, reflecting a change in preferences.

Choosing x_n from the feedback, we obtain

$$y_n = F_n(\mathbf{x}) + v_n,$$

where v_n is noise that models uncertainty in the feedback channel. Two types of noise were used in the simulations:

— symmetric noise $v_i = Z_i \cdot \mathrm{sgn}_i$, where $Z_i \sim \mathrm{Pareto}(\beta, \sigma)$, $\mathrm{sgn}_i \sim \mathrm{Uniform}(\{-1, 1\})$;

— asymmetric noise $v_i = Z_i$, where $Z_i \sim \mathrm{Pareto}(\beta, \sigma)$, which potentially reflects a tendency to overestimate.

Table 1 presents the basic parameters of the numerical simulation. They cover the structure of the experiment, settings of the algorithm (so-called hyper-parameters), as well as the characteristics of noise and drift scenarios, which model feedback instabilities.

Table 1. Parameters of simulations

Parameter	Description	Value
Agent's initial estimate	Initial point for learning	$\hat{\theta}_0 = 0$
Number of iterations	Number of adaptation steps	$N = 1000$
Number of runs	Amount of independent experiments	$m = 1000$
Hyper-parameters		
Adaptation step	Conservative step (for stability)	$\gamma = 0.05$
Level of perturbation	Amplitude to estimate the gradient	$c = 0.1$
Characteristics of noise		
Shape parameter	Defines weights of tails	$\beta = 1.6$
Scale	Intensity of deviations	$\sigma = 2.0$
Rate of drift	Moderate drift	$\delta = 0.01$
	Near-stationary mode	$\delta = 0.0001$
Type of noise	Random deviations	Symmetric
	Systematic bias	Asymmetric

5.2. Simulation Scenarios

To analyze the adaptability of the algorithm, we consider three scenarios:

1. Moderate drift of preferences ($\delta = 0.01$) and symmetric noise (referred to as noise with symmetric distribution). This scenario mimics gradual changes in target parameters in the presence of random errors in the estimates.
2. Near-stationary preferences ($\delta = 0.0001$) and symmetric noise. Within this scenario we test accuracy of tuning under conditions close to stable ones.
3. Stationary preferences ($\delta = 0.0001$) and asymmetric noise (referred to as noise with asymmetric distribution). This scenario corresponds to a systematic distortion of feedback; i.e., a permanent overestimation.

5.3. Agent Adaptation Process

The agent updates its estimate $\hat{\theta}$ of the parameter based on the observed values of y (rewards) obtained from the model. The algorithm follows the iterative scheme described in (3).

Namely, at every even iteration $k = 2n$, $n = 1, 2, \ldots$.

1. The estimate $\hat{\theta}_{2n-2}$ obtained at the previous even iteration is used (for $n = 1$, $\hat{\theta}_0$ is used).
2. A random vector Δ_n of perturbations is generated, with every component independently taking values $+1$ or -1 with probability 0.5.
3. Two points are considered according to (3):

$$x_{2n} = \hat{\theta}_{2n-2} + \beta\Delta_n, \qquad x_{2n-1} = \hat{\theta}_{2n-2} - \beta\Delta_n.$$

4. The values of the reward are then observed at the perturbed points: y_{2n} (associated with x_{2n}) and y_{2n-1} (associated with x_{2n-1}). These two quantities include both the true value of the function and the noise; i.e., $y_n = F_n(x_n, w_n) + v_n$ in terms of the notation of this paper.
5. The estimate $\hat{\theta}$ updates similarly to the formula given by the third line of system (3); however, with sign "+", since the maximization is performed:

$$\hat{\theta}_{2n} \leftarrow \hat{\theta}_{2n-2} + \frac{\alpha}{2\beta}\Delta_n(y_{2n} - y_{2n-1}).$$

At every odd iteration $k = 2n - 1$, the estimate is being copied: $\hat{\theta}_{2n-1} \leftarrow \hat{\theta}_{2n-2}$.

5.4. Checking Conditions (A)–(F) *for Simulations in the RLHF-Scenario*

(A) *Strong convexity of* $f_n(\mathbf{x})$.

$$\nabla f_n(\mathbf{x}) = -\nabla F_n(\mathbf{x}) = -\left[1.35(x_1 - x_n^*)^{0.35}, \dots, 1.35(x_m - x_n^*)^{0.35}\right]^\top,$$
$$\nabla V_n(\mathbf{x}) = [(\rho+1)\mathrm{sgn}(x_1 - x_n^*)|x_1 - x_n^*|^\rho, \dots, (\rho+1)\mathrm{sgn}(x_m - x_n^*)|x_m - x_n^*|^\rho]^\top,$$
$$\langle \nabla V_n(\mathbf{x}), \nabla f_n(\mathbf{x})\rangle = -1.35(\rho+1)\sum_{i=1}^m |x_i - x_n^*|^{\rho+0.35}.$$

Using the inequality $|x_i - x_n^*|^{\rho+0.35} \geq |x_i - x_n^*|^{\rho+1} a^{-0.65}$ with $a \leq |x_i - x_n^*|$, we obtain

$$\sum_{i=1}^m |x_i - x_n^*|^{\rho+0.35} \geq a^{-0.65}\sum_{i=1}^m |x_i - x_n^*|^{\rho+1} = a^{-0.65} V_n(\mathbf{x}).$$

Therefore,

$$\langle \nabla V_n(\mathbf{x}), \nabla f_n(\mathbf{x})\rangle \leq -1.35(\rho+1)a^{-0.65} V_n(\mathbf{x});$$

i.e., the condition of the form $\langle \nabla V_n(\mathbf{x}), \nabla f_n(\mathbf{x})\rangle \geq \mu V_n(\mathbf{x})$ holds for $\mu = -1.35(\rho+1)a^{-0.65} < 0$. In the minimization of $f_n(\mathbf{x})$, the strong convexity condition in the sense of the scalar inequality above is satisfied with $\mu < 0$.

(B) *The Hölder continuity of the gradient.*

The gradient of the reward function $F_n(x)$ writes

$$\nabla F_n(x) = -1.35\left[(x_1 - x_n^*)^{0.35}, \dots, (x_m - x_n^*)^{0.35}\right]^\top.$$

Then the components of the difference of the gradients have the form

$$\left|(x_i - x_n^*)^{0.35} - (y_i - x_n^*)^{0.35}\right| \leq M'|x_i - y_i|^{0.35},$$

where M' is the Hölder constant, which exist for the function $s \mapsto s^{0.35}$ over bounded intervals.

Substitution to the norm gives

$$\|\nabla F_n(x) - \nabla F_n(y)\|_2^2 = 1.35^2 \sum_{i=1}^m \left|(x_i - x_n^*)^{0.35} - (y_i - x_n^*)^{0.35}\right|^2$$
$$\leq 1.35^2 M'^2 \sum_{i=1}^m |x_i - y_i|^{0.7} \leq M^2 \|x - y\|_2^{0.7},$$

where $M^2 = 1.35^2 M'^2 m^{1-0.7/2}$ is a generalized constant.

Then we have

$$\|\nabla F_n(x) - \nabla F_n(y)\|_2 \leq M\|x - y\|_2^{0.35},$$

which corresponds to condition (B) with $\rho = 0.35$ and $M = 1.35M'm^{0.325}$.

(C) *The local Lebesgue condition.*

Let us fix the point x and consider its neighborhood $U_x = B(x, \varepsilon)$ for some $\varepsilon > 0$. Then, for any $x' \in U_x$ we have

$$\|\nabla F_n(x', w)\|_2^2 = 1.35^2 \sum_{i=1}^m |x_i' - x_n^*|^{0.7} \leq 1.35^2 m R^{0.7},$$

where $R = \sup_{x' \in U_x} \max_i |x_i' - x_n^*| < \infty$, and it is finite by the construction of U_x.

We then can set $\Phi_x(w) = 1.35\sqrt{m}R^{0.35}$, which is independent of w, so that $\mathbb{E}\Phi_x(w) = \Phi_x(w) < \infty$. Condition (C) is satisfied.

(D.a) *Boundedness of the drift of the minimum point.*

Since $\theta_n = x_n^*\mathbf{1}$ and $x_n^* = x_{n-1}^* + \delta$, we have $\|\theta_n - \theta_{n-1}\|_2 = \|\delta\mathbf{1}\|_2 = \delta\sqrt{m}$. Hence, condition (D.a) is satisfied for $A = \delta\sqrt{m}$.

(D.b) *Boundedness of change in the gradient.*

Let $r_i = x_i - x_{n-1}^*$, then

$$|\partial_i F_n(x) - \partial_i F_{n-1}(x)| \le 1.35M'|\delta|^{0.35},$$

where M' is the Hölder constant of the function $s^{0.35}$ over the fesible compact.

Summing up over i we obtain

$$\|\nabla_x F_n(x) - \nabla_x F_{n-1}(x)\|_1 \le 1.35M'm|\delta|^{0.35}.$$

Denote $R = \inf_{x\neq\theta_{n-1}} \|x - \theta_{n-1}\|_1 > 0$; then $\|x - \theta_{n-1}\|_1^\rho \ge R^\rho$, and condition (D.b) is satisfied for

$$B = \frac{1.35M'm\delta^{0.35}}{R^{0.35}}.$$

(D.c) *Boundedness of the gradient at the minimum point.*

Since $\theta_n = x_n^*\mathbf{1}$, we have $\nabla_x F_n(\theta_n) = \mathbf{0}$; therefore, $\|\nabla_x F_n(\theta_n, w_n)\|_{\rho+1}^{\rho+1} = 0$, so that the condition holds for $C = 0$.

(D.d) *Boundedness of change in the function at a step.*

$$F_{n-1}(x,w) = -\sum_{i=1}^m (x_i - x_{n-1}^*)^{1.35},$$

$$F_n(x,w) - F_{n-1}(x,w) = \sum_{i=1}^m \left[(x_i - x_{n-1}^*)^{1.35} - (x_i - x_n^*)^{1.35}\right],$$

$$\left|(x_i - x_n^*)^{1.35} - (x_i - x_{n-1}^*)^{1.35}\right| \le M'|\delta|^{1.35}$$

$$|F_n(x,w) - F_{n-1}(x,w)| \le mM'\delta^{1.35}.$$

Since the noise v is subject to the Pareto distribution with parameter $\beta = 1.6 > \rho + 1 = 1.35$, the moment of order 1.35 does exist, and $\mathbb{E}|v_n - v_{n-1}|^{\rho+1} \le \tilde{E} < \infty$. Therefore, for $D = 0$ and $E = (mM'\delta^{1.35} + \tilde{E})$ condition (D.d) is satisfied:

$$\mathbb{E}_{\mathcal{F}_{2n-2}}|F_{2n}(x, w_{2n}) - F_{2n-1}(x, w_{2n-1})|^{\rho+1} \le DV_{2n-2}(x) + E.$$

(E) *Boundedness of change in the observed noise.*

Consider the observation noise v_n defined via the Pareto noise:

$$v_n = \begin{cases} Z_n \mathrm{sgn}_n, & \text{symmetric noise,} \\ Z_n, & \text{asymmetric noise,} \end{cases}$$

where $Z_n \sim \mathrm{Pareto}(\beta = 1.6, \sigma = 2.0)$, $\mathrm{sgn}_n \sim \mathrm{Uniform}\{-1, 1\}$.

Condition (E) requires the fulfillment of the inequality

$$\mathbb{E}_{\mathcal{F}_{2n-2}} |v_{2n} - v_{2n-1}|^{\rho+1} \le \sigma_v^{\rho+1},$$

where $\rho + 1 = 1.5 < \beta$; i.e., the moment of order 1.5 does exist.

Since v_{2n} and v_{2n-1} are independent, the difference $v_{2n} - v_{2n-1}$ is also a random variable with finite moment of order $\rho + 1$. For the symmetric case (with alternating signs) numerical simulation over 10^6 realizations results in $\mathbb{E}\,|v_{2n} - v_{2n-1}|^{1.5} \approx 53.73$, which allows to admit $\sigma_v^{1.5} = 53.73$. Hence, condition (E) is satisfied with explicitly defined constant $\sigma_v^{\rho+1} = 53.73$.

(F) *Independence of perturbations* Δ_n.

By the construction of the search algorithm and the simulations with the RLHF-model, the vectors Δ_n are generated to be independent of all exogenous factors. The noise v_n is incorporated afterwards and does not depend on the chosen direction of perturbation.

5.5. Metrics for the Estimates and the Results of Simulations

We use a system of empirical metrics to quantify the behavior of the algorithm under the conditions of the optimum drift and the presence of noise with heavy tails. These metrics account for both the accuracy and stability of the estimates and the dynamics of adaptation to changing conditions. The metrics are selected in such a way as to cover both the steady-state characteristics of the algorithm and its behavior throughout optimization. This makes it possible to identify the strengths and weaknesses of the method in various scenarios, from stationary to rapidly changing and noisy ones.

The assessment of the average accuracy of tracking a drifting parameter on the later stages of the algorithm is performed through the average absolute error over the last iterations. The stability of the behavior of the algorithm is determined by the standard deviation of these errors. The range of fluctuations within a single run is characterized by the average minimum and maximum errors across runs, which allows for an evaluation of both the achievable potential and worst-case cases.

The dynamical characteristics of the algorithm are reflected in the metrics of the average time to achieve a given level of accuracy; this provides insight into the rate of adaptation under constraints on the error. The connection to theoretical definitions of stability is ensured through two moments of error: The moment of order ρ, which assesses convergence on average, and the corresponding asymptotic bound that normalizes the error according to the chosen order of the moment. The order used is selected based on the noise parameters to ensure the existence of the corresponding mathematical expectations.

Table 2. Basic metrics of the algorithm

Metrics	Expression
Mean absolute deviation over last 100 iterations	$\mu_{\text{last100}} = \frac{1}{100m} \sum\limits_{n=N-100}^{N-1} \sum\limits_{i=1}^{m} \lvert x_{n,i} - x_n^* \rvert$
Standard deviation of errors over last 100 iterations	$\sigma_{\text{last100}} = \sqrt{\frac{1}{100m-1} \sum\limits_{n=N-100}^{N-1} \sum\limits_{i=1}^{m} \left(\lvert x_{n,i} - x_n^* \rvert - \mu_{\text{last100}} \right)^2}$
Minimum mean deviation over the runs	$\bar{D}_{\min} = \frac{1}{m} \sum\limits_{i=1}^{m} \min_{0 \le n < N} \lvert x_{n,i} - x_n^* \rvert$
Maximum mean deviation over the runs	$\bar{D}_{\max} = \frac{1}{m} \sum\limits_{i=1}^{m} \max_{0 \le n < N} \lvert x_{n,i} - x_n^* \rvert$
Mean convergence time to threshold ϵ	$\bar{T}_\epsilon = \frac{1}{m} \sum\limits_{i=1}^{m} T_{i,\epsilon}$, $T_{i,\epsilon} = \min\{\{n \mid 0 \le n < N, \lvert x_{n,i} - x_n^* \rvert < \epsilon\} \cup \{N\}\}$
$l_{\rho+1}$-metrics of the estimation error	$\mu_{\text{def2,last100}} = \frac{1}{100} \sum\limits_{n=N-100}^{N-1} \frac{1}{m} \sum\limits_{i=1}^{m} \left(\lvert x_{n,i} - x_n^* \rvert^{\rho+1} \right)^{1/2}$

The definitions of the metrics are given in Table 2, a comparison of the results for different metrics is presented in Table 3, and their dynamics are plotted in Fig. 1.

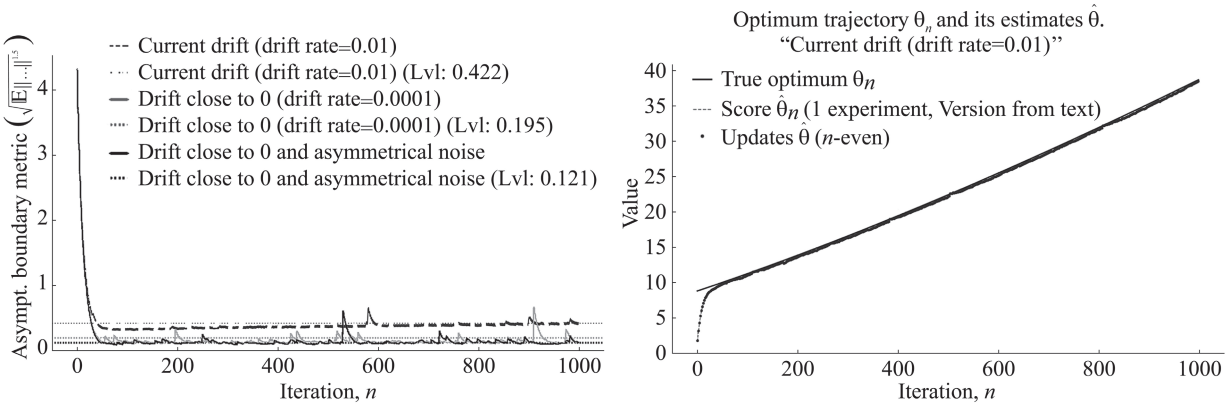

Fig. 1. Quality of the estimates. Left: Plot of the $l_{\rho+1}$ estimate of the error as function of the iteration number n; right: True trajectory of the optimum x_n^* and its estimate x_{n,i_0}.

The results of simulations presented in Table 3 testify to the influence of the environmental parameters on the performance of the algorithm (based on 1000 experiments, $\rho = 0.50$; statistics for the last 100 iterations is presented). With moderate drift ($\delta = 0.01$) and symmetric noise, the agent does track the goal, but with a noticeable average error (0.3012), moderate stability (Std = 0.1682), and rare but significant outliers (maximum 19.7897). High-order metrics take values 0.1788 and 0.4219, with convergence achieved in 20 iterations.

Table 3. Comparison of the results for different types of drift and noise

Metrics	Moderate drift $\delta = 0.01$ (symm)	Near-stationary $\delta = 0.0001$ (symm)	Asymmetric noise $\delta = 0.0001$ (asymm)
Stability metrics $E[\|x - x^*\|^{1.5}]$	0.1788	0.0524	0.0153
$l_{\rho+1}$-metrics of the estimation error	0.4219	0.1954	0.1206
Mean distance $E[\|x - x^*\|]$	0.3012	0.0520	0.0356
Standard deviation of the estimate	0.1682	0.2776	0.0989
Minimum deviation	0.0002	0.0000	0.0000
Maximum deviation	19.7897	57.1922	10.3462
Convergence time (< 1.0)	20	18	18

Decrease of drift down to $\delta = 0.0001$ (near-stationary environment) improves the mean error (0.0520); at the same time it increases instability. Namely, the standard deviation reaches the value 0.2776, and the maximum error attains the level of 57.19. This indicates an increase in sensitivity to noise with heavy tails under weakened drift.

The best results were achieved for asymmetric noise under conditions of weak drift. The error decreases to 0.0356, the variability is bounded (Std = 0.0989), and the maximum deviations are significantly lower (10.3462). Stability metrics (0.0153, 0.1206) and convergence time (18 iterations) also improve.

Hence, decrease of rate of drift increases the accuracy; however, robustness to noise depends on its type. Thus, asymmetric noise implies a better control over extreme errors, perhaps due to the specifics of gradient estimate. This effect requires further analysis.

6. SIMULATION OF THE TASK DISTRIBUTION SYSTEM IN QUEUEING SYSTEM PROBLEMS

Queueing systems, such as modern call centers, are characterized by an incoming flow of tasks having processing times that are often subject to heavy-tailed distributions [31]. This indicates

the presence of a statistically significant share of tasks that require disproportionately large processing times, distinguishing them from systems described by classical exponential or Gaussian distributions. The Pareto distribution can be thought of as a suitable model for describing such phenomena [32], since it accounts for rare but lengthy operations which affect the overall performance of the system [33].

To efficiently control such queueing systems, one has to adaptively evaluate the characteristics of the flow and time of service. Below we analyze an application of our stochastic optimization search algorithm (3) to the model of dynamical tuning the estimated expected processing time for different types of tasks; see [34] for a detailed description of the model. We use our method to iteratively optimize the parameters $\hat{\theta}_k, \hat{\theta}_m$, which are adaptive estimates of time of service for each task cluster m and for the system as a whole, k.

A simulation model of a call center was presented in [34]. Task service time in the model is generated from the Pareto distribution, with the parameters being calibrated for each cluster based on the characteristics of lognormal distributions that approximate historical data. The search algorithm (3) is used to refine the estimates $\hat{\theta}_k$, $\hat{\theta}_m$, which in turn are used for the assignment of incoming tasks to agents. The simulation shows the satisfactory performance of the method in the stochastic environment and heavy-tail nature of the task processing time.

6.1. The Model

We consider a system of agents having identical resources and performance. The load of agent i, denoted by q^i, corresponds to the number of tasks in its queue. Each task x_k is characterized by type m and the predicted execution time, calculated via the formula

$$x_{km} = \alpha\hat{\theta}^i_k + (1-\alpha)\hat{\theta}^i_m, \quad \alpha = \frac{\chi|\lambda_m|}{N_m+1},$$
$$\lambda_m(\hat{\theta}_m) = \frac{1}{N_m}\sum_{k\in N_m}\omega_k.\frac{\hat{\theta}_m - t_{km}}{\hat{\theta}_m} \to \min,$$

where $\hat{\theta}^i_k$ is the individual forecast of agent i for task k, $\hat{\theta}^i_m$ is the average predicted time to complete tasks of type m (taking local history into account), α is the weight factor that determines the contribution of the individual forecast and aggregated statistics, and χ is the convergence coefficient. The quantity λ_m characterizes the accuracy of the model prediction for problems of type m type and it is corrected when new observations are received. Here, N_m is the amount of completed tasks of type m, ω_k is the weight of the corresponding error, and t_{km} is the actual time to complete task k of type m.

Such a mechanism for calculating predictions and accuracy let the model adapt to the current quality of forecasts reducing the impact of unreliable data and strengthening the contribution of accumulated statistics with high confidence.

As a new task x_k arrives at step k, it is assigned to the following agent i_k in order to balance the load of the agents:

$$i_k = \arg\min_i \sum_j \left|\frac{q^i_k + x_{km} - q^j_k}{d_{ij}+1}\right|, \tag{5}$$

where q^j is the load of agent j, d_{ij} is the "distance" between agents (for example, based on load or physical location). The agents are connected in a fully connected topology, where each agent interacts with all the others. This ensures global communication with varying influence of agents depending on their relative proximity.

6.2. Description of the Data Set and the Primary Analysis

To demonstrate the efficiency of the developed method, a modeling of the load distribution system was conducted based on real data from an operator call center for September 2023 (over 2.3 million calls). For each inquiry, we recorded the instant of arrival, response time (wait time), the actual duration of call (ACD Time), and customer's segment.

Figure 2 presents two complementary visualizations that reveal the key characteristics of the incoming flow of some clusters and the human resources potential of the call center. The plot on the left shows the hourly intensity of tasks over the ten largest clusters, with the peak load observed for one of the clusters between 10 AM and noon. The plot on the right shows the distribution of active operators (those who received more than 50 calls in a two-hour interval), with maximum values occurring between 8 AM and 4 PM. At the same time, the personnel resources do not always keep up with the sharp fluctuations in incoming traffic. Simulation delays aggregated by time of day generally replicate the dynamics of the actual wait times, including a morning rise around 8–9 AM and an evening peak after 5 PM.

The diagrams presented in Fig. 3 display the distribution of conversation durations for 14 client segments. For the sake of anonymization, all segments have been renamed to numerical identifiers from 1 to 14 (see Table 4). The greatest variability and extended tails of the distribution are observed in segments 11 and 13, whereas segment 2 is characterized by an exclusively short duration range. Segments 3 and 14 also demonstrate a relatively narrow distribution with short medians.

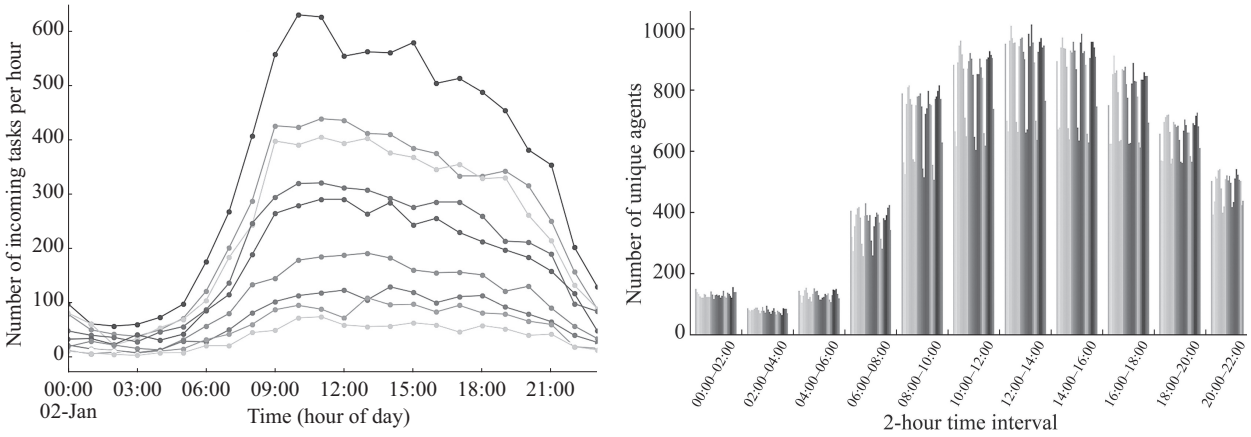

Fig. 2. Dynamics of load and the personnel time commitment. Left: Hourly intensity of tasks (top ten clusters); right: Amount of active agents over two-hour intervals.

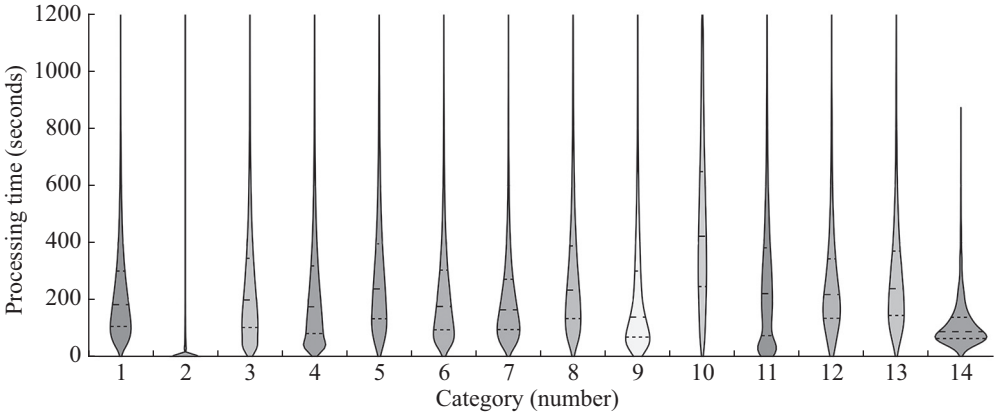

Fig. 3. Duration of calls for the top 14 clusters (max ACD = 1200 sec).

6.3. Results of Simulations

To assess the performance of the proposed method, a simulation of the call center operations was conducted using real data. The results allowed for the evaluation of both the dynamics of task wait times throughout the day and the stability of the load distribution. Figure 4 presents the results of a specific simulation session.

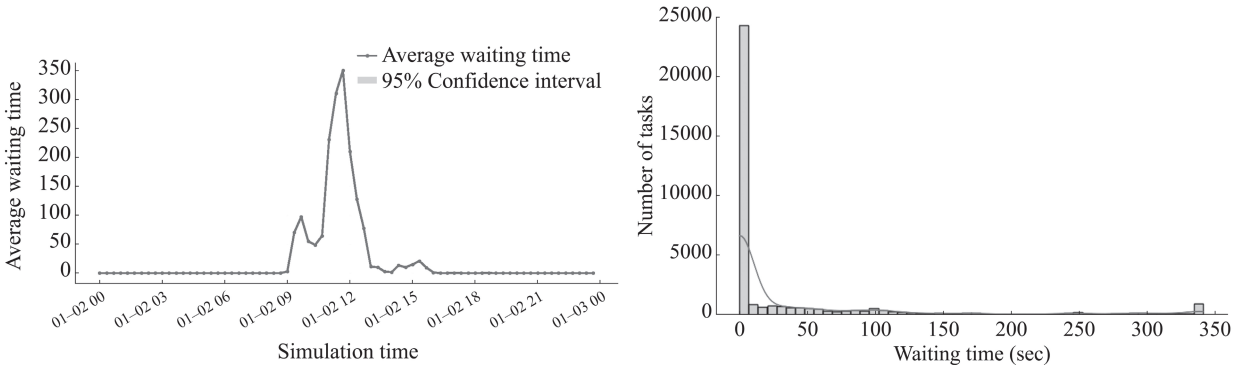

Fig. 4. The results of the simulation model: Analysis of delays in the service. Left: Mean wait time (20-minute intervals), right: The distribution of wait time (98th percentile).

The plot on the left presents the mean wait time for tasks over twenty-minute intervals showing a salient peak during work hours associated with high load. The model efficiently adapts to the changing environment; namely, after a sharp growth of delays at around noon, the mean wait time quickly decreases due to the redistribution of tasks.

The histogram on the right represents the 98-percentile distribution of the wait times. Most of the tasks have been processed in less than 50 sec, which corresponds well to the target SLA-indicators for typical scenarios.

For key clusters, Table 4 presents the values of the predicted processing time z, the amount k of completed requests, the mean actual time t_{avg}, and the maximum duration t_{max}. Clusters 1 to 14 correspond to those presented in Fig. 3, whereas cluster 0 accumulates all other segments outside of the top-14. The quantities z are seen to fit well the empirical means, despite the different statistics for different clusters, which confirms robustness properties of the adaptive prediction based on the developed search algorithm.

Table 4. Results for different key clusters

	0	1	2	3	4	5	6	7
z	143.76	163.14	3.73	174.75	147.62	196.05	159.75	151.75
k	36858	8180	6166	5764	4461	3857	2523	1707
t_{avg}	124.34	158.94	3.71	163.51	135.06	174.91	161.16	157.93
$t_{\max}$	200	200	200	200	200	200	200	200
	8	9	10	11	12	13	14	
z	219.63	151.74	321.53	144.92	147.55	191.31	49.54	
k	1329	943	906	484	265	223	44	
t_{avg}	233.06	153.30	330.74	158.08	156.68	198.90	87.34	
$t_{\max}$	200	200	200	200	200	200	44	

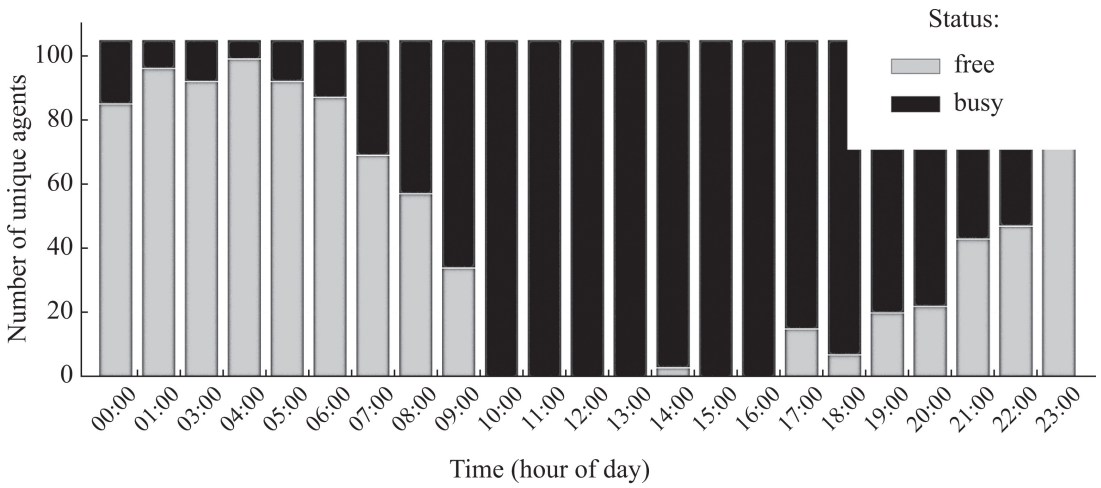

Fig. 5. Hourly load of agents: Comparison of vacant and occupied resources.

The plot presented in Fig. 5 illustrates the hourly workload of operators during the simulation. During the night and morning hours (until 8 AM), a significant portion of agents remains free; however, between 9 AM and 3 PM, there is full utilization of all resources: The number of free agents drops to zero. This coincides with the peak of incoming task flow and stresses the need for an accurate prediction of the duration of processing. In the evening and at night, the load gradually decreases, and the system returns to a balanced state.

Overall, the model demonstrates the ability to correctly adapt to the load, ensuring the mitigation of wait times and an even distribution of tasks throughout the day. The proposed approach allows for efficient resource utilization under conditions of high variability in requests and can be recommended for implementation in distributed support systems with intensive and irregular loads.

7. CONCLUSIONS

In this paper we proposed and thoroughly analyzed a method for estimating the minimum of a functional which varies in time, under conditions where the measurements are subject to noise. This method is based on the pseudogradient approach with randomization and it does not rely on the knowledge of the gradient of the objective function and uses a small number of observations at every iteration. An assumption was made on the boundedness of rate of change (drift) of the extremum of the functional. It is proved that the asymptotic estimation error is bounded from above by $\frac{L}{K}$, where L and K are found from the properties of the objective function, noise characteristics, and the parameters of the algorithm. The validity of the theoretical conclusions was confirmed by the results of numerical simulations which testified to efficient adaptation of RLHF-agents to noisy and dynamical feedback (in particular, heavy-tailed noise and different preference drift rate). The experiments showed that the search algorithm ensures the convergence of the estimates to the target value region.

According to the simulations, the steady-state error and oscillations in the estimates resulting from noise and drift are consistent with theoretical predictions about the boundedness of the asymptotic error. Furthermore, the proposed method was tested through simulations based on real data from an operator call center. Use of empirical characteristics of the flow of requests and processing times demonstrated reliable applicability of the algorithm in dynamical load distribution problems and in predicting service parameters in real service systems.

FUNDING

The theoretical results presented in Sections 1–4 were obtained in the Institute for Problems in Mechanical Engineering, Russian Academy of Science, under the financial support of the Russian Science Foundation (project No. 23-41-00060); the applied part presented in Sections 5 and 6 was implemented under the financial support of the Saint Petersburg State University, project No. 121061000159-6.

APPENDIX

Proof of Theorem 1. Denote the estimation error by $\mathrm{err}_n = \hat{\theta}_n - \theta_n$.

Step 1: Recursive relation for the estimation error. By algorithm (3) we have

$$\hat{\theta}_{2n} = \hat{\theta}_{2n-2} - \frac{\alpha}{2\beta}\Delta_{2n}(y_{2n} - y_{2n-1}),$$

hence,

$$\mathrm{err}_{2n} = \mathrm{err}_{2n-2} - \underbrace{(\theta_{2n} - \theta_{2n-2})}_{\mathrm{drift}_n} - \underbrace{\frac{\alpha}{2\beta}\Delta_{2n}(y_{2n} - y_{2n-1})}_{\mathrm{step}_n}.$$

Step 2: Recursive relation for the estimate of the Lyapunov function $V(x)$. For the vectors $a = \hat{\theta}_{2n-2}$ and $b = \mathrm{drift}_n + \mathrm{step}_n$ we have

$$V_{2n}(\hat{\theta}_{2n}) = V_{2n-2}(\hat{\theta}_{2n} - \mathrm{drift}_n) = V_{2n-2}(a - b) = \|a - b - \theta_{2n-2}\|_{\rho+1}^{\rho+1}$$

by definition. Using the Taylor series expansion of the function $V_{2n-2}(a - b)$ at the point a in the direction $-b$, we obtain

$$V_{2n-2}(a - b) = V_{2n-2}(a) - \langle \nabla V_{2n-2}(a - \delta b), b\rangle, \quad \delta \in [0, 1], \tag{A.1}$$

noting that the gradient $\nabla V_{2n-2}(a - \delta b)$ is computed according to

$$\nabla V_{2n-2}(a - \delta b) = (\rho + 1) \cdot \mathrm{sgn}(\delta) \odot |a - \theta_{2n-2} - \delta b|^{\rho},$$

where $\mathrm{sgn}_n^{(i)}(\delta) = 0$ or ± 1 depending on the sign of the ith component of the vector $a - \theta_{2n-2} - \delta b$; $|a - \theta_{2n-2} - \delta b|^{\rho}$ is the vector of the absolute values of the components of $a - \theta_{2n-2} - \delta b$ to the power ρ, and $\odot$ denotes the componentwise multiplication. The second term in (A.1) can be evaluated as

$$-\langle \nabla V_{2n-2}(a - \delta b), b\rangle \leqslant -\langle (\rho + 1) \cdot \mathrm{sgn}(0) \odot |a - \theta_{2n-2}|^{\rho}, b\rangle + 2^{1-\rho}\delta^{\rho}\|b\|_{\rho+1}^{\rho+1} \leqslant$$

$$-\langle \nabla V_{2n-2}(a), b\rangle + 2^{1-\rho}\|b\|_{\rho+1}^{\rho+1}$$

(see proof of Theorem 1 in [24], p. 93).

Keeping the considerations above and using condition (D.a), we have

$$V_{2n}(\hat{\theta}_{2n}) \leqslant V_{2n-2}(\hat{\theta}_{2n-2}) - \langle \nabla V_{2n-2}(\hat{\theta}_{2n-2}), \mathrm{drift}_n + \mathrm{step}_n\rangle + 2(A^{\rho+1} + \|\mathrm{step}_n\|_{\rho+1}^{\rho+1}). \tag{A.2}$$

Step 3: Expansion of the correcting term. According to the model of observations, represent the term step_n as the sum

$$\mathrm{step}_n = \underbrace{\frac{\alpha}{2\beta}\Delta_n\Big(F_{2n}(x_{2n}, w_{2n}) - F_{2n-1}(x_{2n-1}, w_{2n-1})\Big)}_{\text{almost pseudogradient term}} + \underbrace{\frac{\alpha}{2\beta}\Delta_n(v_{2n} - v_{2n-1})}_{\text{noise}}.$$

a. Almost pseudogradient term. Denote $n^{\pm} = 2n - \frac{1}{2} \pm \frac{1}{2}$.

Using the Taylor formula, we first add and subtract the quantity $\sum_{n^{\pm}} \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle$, then the quantity $\langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle$, and finally $\langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle$, to obtain

$$\begin{aligned}
\sum_{n^{\pm}} \pm F_{n^{\pm}}(x_{n^{\pm}}, w_{n^{\pm}}) &= \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) + \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&= \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) + \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&= \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) + \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle,
\end{aligned}$$

where $\delta_{n^{\pm}} \in [0, 1]$.

Now take the conditional mathematical expectation with respect to the σ-algebra $\mathcal{F}_{2n-2}$. By condition (F), the vectors Δ_n are independent of $w_{n^{\pm}}$ and the σ-algebra $\mathcal{F}_{2n-2}$, hence we have

$$\frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}} \left\{ \Delta_n \sum_{n^{\pm}} \pm F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}) \right\} = 0,$$

since Δ_n are centered, and

$$\frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}} \left\{ \Delta_n \sum_{n^{\pm}} \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \right\} = 0,$$

since $\mathbb{E}_{\mathcal{F}_{2n-2}}\{\nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}})\} = \nabla_x f_{2n-2}(\theta_{2n-1})$ by condition (C), and the gradient of $f_{2n-2}(\cdot)$ at the minimum point θ_{2n-2} is equal to zero.

As a result, by condition (C) we obtain

$$\mathbb{E}_{\mathcal{F}_{2n-2}} \left\{ \frac{\alpha}{2\beta}\Delta_n \sum_{n^{\pm}} \pm F_{n^{\pm}}(x_{n^{\pm}}, w_{n^{\pm}}) \right\} = \frac{\alpha}{d}\nabla f_{2n}(\hat{\theta}_{2n-2}) + \frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}}\mathrm{corr}_n$$

for the almost pseudogradient term, where

$$\begin{aligned}
\mathrm{corr}_n = \sum_{n^{\pm}} &\langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2} \pm \delta_{n^{\pm}}\beta\Delta_n, w_{n^{\pm}}), \beta\Delta_n \rangle - \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\theta_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle \\
&+ \langle \nabla_x F_{n^{\pm}}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle - \langle \nabla_x F_{2n-2}(\hat{\theta}_{2n-2}, w_{n^{\pm}}), \beta\Delta_n \rangle.
\end{aligned}$$

By conditiond (B) and (D.b), the following estimate holds:

$$\begin{aligned}
\|\mathrm{corr}_n\| &\leqslant M\beta^{\rho}\|\Delta_n\| \left(2\|\Delta_n\|^{\rho} + 2\|\hat{\theta}_{2n-2} - \theta_{2n-2}\|^{\rho}) + 3B\|\hat{\theta}_{2n-2} - \theta_{2n-2}\|^{\rho} \right) \\
&= 2M\beta^{\rho} + (2 + 3B)\|\hat{\theta}_{2n-2} - \theta_{2n-2}\|^{\rho}.
\end{aligned}$$

b. Noise. Take the conditional mathematical expectation with respect to the σ-algebra $\mathcal{F}_{2n-2}$. By the independence of Δ_n on v_{2n}, v_{2n-1} and $\mathcal{F}_{2n-2}$, we obtain

$$\mathbb{E}_{\mathcal{F}_{2n-2}}\left\{\frac{\alpha}{2\beta}\Delta_n(v_{2n}-v_{2n-1})\right\}=0.$$

c. The final estimate of the second term on the right-hand side of Ineq (A.2). By the strong convexity (see condition (A)), we obtain

$$\begin{aligned}
-\mathbb{E}_{\mathcal{F}_{2n-2}}\{\langle\nabla V_{2n-2}(\hat\theta_{2n-2}),\mathrm{drift}_n+\mathrm{step}_n\rangle\} &\leqslant -\frac{\mu\alpha}{d}V_{2n-2}(\hat\theta_{2n-2})\\
-\frac{\alpha}{2\beta}\mathbb{E}_{\mathcal{F}_{2n-2}}\langle\nabla V_{2n-2}(\hat\theta_{2n-2}),\mathrm{drift}_n+\mathrm{corr}_n\rangle &\leqslant -\frac{\mu\alpha}{d}V_{2n-2}(\hat\theta_{2n-2})\\
+2(A+\alpha M\beta^{\rho-1})^2+\left(2+\frac{\alpha}{2\beta}(2+3B)\right)\sum_{i=1}^{d}|\hat\theta^i_{2n-2}-\theta^i_{2n-2}|^{2\rho}&\\
\leqslant -\frac{\mu\alpha}{d}V_{2n-2}(\hat\theta_{2n-2})+\varepsilon V_{2n-2}(\hat\theta_{2n-2})+c_1,&
\end{aligned}$$

where $\varepsilon>0$ and

$$c_1=2(A+\alpha M\beta^{\rho-1})^2+\varepsilon^{\rho-1}\left(2+\frac{\alpha}{2\beta}(2+3B)\right)^{\frac{1-\rho}{\rho+1}}.$$

Step 4: Estimate of the third term on the right-hand side of inequality (A.2). Similarly to the derivations at Step 3 above, the term step_n can be represented as

$$\mathrm{step}_n=\frac{\alpha}{2\beta}\Delta_n\sum_{i=1}^{8}a_i,$$

where

- $a_1=\sum_{n^\pm}\pm F_{n^\pm}(\hat\theta_{2n-2},w_{n^\pm})$;
- $a_2=a_3=\langle\nabla_x F_{n^\pm}(\hat\theta_{2n-2}\pm\delta_{n^\pm}\beta\Delta_n,w_{n^\pm}),\beta\Delta_n\rangle-\langle\nabla_x F_{n^\pm}(\hat\theta_{2n-2},w_{n^\pm}),\beta\Delta_n\rangle$;
- $a_4=a_5=\langle\nabla_x F_{n^\pm}(\hat\theta_{2n-2},w_{n^\pm}),\beta\Delta_n\rangle-\langle\nabla_x F_{n^\pm}(\theta_{n^\pm},w_{n^\pm}),\beta\Delta_n\rangle$;
- $a_6=a_7=\langle\nabla_x F_{n^\pm}(\theta_{n^\pm},w_{n^\pm}),\beta\Delta_n\rangle$;
- $a_8=v_{2n}-v_{2n-1}$.

Respectively, we have

- for a_1: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_1|^{\rho+1}\leqslant DV_{2n-2}(\hat\theta_{2n-2})+E$ by condition (D.d);
- for a_2,a_3: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_i|^{\rho+1}\leqslant M^{\rho+1}\beta^{2\rho+2}$, $i=2,3$, by condition (B);
- for a_4,a_5: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_i|^{\rho+1}\leqslant(M\beta\|\hat\theta_{2n-2}-\theta_{n^\pm}\|_2^\rho)^{\rho+1}\leqslant M^{\rho+1}\beta^{\rho+1}d^{\frac{\rho-1}{2}}V_{n^\pm}(\hat\theta_{2n-2})$, $i=4,5$, by condition (B) and Jensen's inequality;
- for a_6,a_7: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_i|^{\rho+1}\leqslant C$, $i=6,7$, by condition (D.c);
- for a_8: $\mathbb{E}_{\mathcal{F}_{2n-2}}|a_8|^{\rho+1}\leqslant\sigma_v^{\rho+1}$ by condition (E).

Overall, by Jensen's inequality we obtain

$$\left(\frac{\sum_{i=1}^{8}|a_i|}{8}\right)^{\rho+1}\leqslant\frac{1}{8}\sum_{i=1}^{8}|a_i|^{\rho+1},$$

so that

$$2A^{\rho+1} + 2\mathbb{E}_{\mathcal{F}_{2n-2}}\|\mathrm{step}_n\|_{\rho+1}^{\rho+1} \leqslant 2A^{\rho+1} + 2\cdot 8^{\rho}\left(\frac{\alpha}{2\beta}\right)^{\rho+1}\sum_{i=1}^{7}|a_i|^{\rho+1}$$

$$\leqslant 2A^{\rho+1} + 2^{2\rho}\alpha^{\rho+1}\left(2M^{\rho+1}(\beta^{\rho+1} + d^{\frac{\rho-1}{2}}\sum_{n^{\pm}}V_{n^{\pm}}(\hat{\theta}_{2n-2})) + \frac{2C + DV_{2n-2}(\hat{\theta}_{2n-2}) + E + \sigma_v^{\rho+1}}{\beta^{\rho+1}}\right)$$

$$\leqslant c_2\alpha^{\rho+1}V_{2n-2}(\hat{\theta}_{2n-2}) + c_3,$$

where

$$c_2 = 2^{3\rho+1}M^{\rho+1}\left(d^{\frac{\rho-1}{2}} + \frac{D}{\beta^{\rho+1}}\right)$$

and

$$c_3 = 2A^{\rho+1} + 2^{2\rho}\alpha^{\rho+1}\left(2M^{\rho+1}(\beta^{\rho+1} + 3\cdot 2^{\rho}d^{\frac{\rho-1}{2}}) + \frac{E + 2C + \sigma_v^{\rho+1}}{\beta^{\rho+1}}\right).$$

Step 5: Shaping the recursive inequality for the Lyapunov function. Collecting all estimates obtained above, we arrive at

$$V_{2n} \leqslant V_{2n-2} - (\mu\alpha d^{-1} - \varepsilon - c_2\alpha^{\rho+1})V_{2n-2} + c_1 + c_3.$$

Introducing the notation

$$K = 1 - \mu\alpha d^{-1} + \varepsilon + c_2\alpha^{\rho+1}, \quad L = c_1 + c_3,$$

we obtain

$$V_{2n} \leqslant (1-K)V_{2n-2} + L.$$

By choosing α and ε sufficiently small, the inequality $K < 1$ can be achieved, which implies the assertion of Theorem 1. □

REFERENCES

1. Polyak, B.T., *Introduction to Optimization*, New York, Optimization Software, 1987.
2. Polyak, B.T. and Tsypkin, Ya.Z., Pseudogradient Adaptation and Training Algorithms, *Autom. Remote Control*, 1973, vol. 34, no. 3, pp. 377–397.
3. Polyak, B.T. and Tsypkin, Ya.Z., Adaptive Estimation Algorithms (Convergence, Optimality, Stability), *Autom. Remote Control*, 1979, vol. 40, no. 3, pp. 378–389.
4. Poliak, B.T. and Tsypkin, J.Z., Optimal Pseudogradient Adaptation Algorithms, *Autom. Remote Control*, 1981, vol. 41, no. 8, pp. 1101–1110.
5. Polyak, B.T., Some Methods of Speeding up the Convergence of Iteration Methods, *USSR Comput. Math. Math. Phys.*, 1964, vol. 4, no. 5, pp. 1–17.
6. Polyak, B.T., New Method of Stochastic Approximation Type, *Autom. Remote Control*, 1990, vol. 51, no. 7, pp. 937–946.
7. Polyak, B.T. and Yuditsky, A.B., Acceleration of Stochastic Approximation by Averaging, *SIAM J. Control Optim.*, 1992, vol. 30, no. 4, pp. 838–855.
8. Polyak, B.T., Convergence and Convergence Rate of Iterative Stochastic Algorithms. I. General Case, *Autom. Remote Control*, 1976, vol. 37, no. 12, pp. 1858–1868.
9. Polyak, B.T., Convergence and Convergence Rate of Iterative Stochastic Algorithms. II. The Linear Case, *Autom. Remote Control*, 1977, vol. 38, no. 4, pp. 537–542.
10. Polyak, B.T. and Tsybakov, A.B., Optimal Order of Accuracy for Search Algorithms in Stochastic Optimization, *Problems Inform. Transmiss.*, 1990, vol. 26, no. 2, pp. 126–133.
11. Rastrigin, L.A., *Statisticheskie metody poiska* (Statistical Search Methods), Moscow, Nauka, 1968.
12. Granichin, O.N., Stochastic Approximation with Input Perturbation under Dependent Observation Noises, *Vestn. Leningr. Univ.*, 1989, pp. 27–31.

13. Spall, J.C., Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Trans. Autom. Control*, 1992, vol. 37, no. 3, pp. 332–341.
14. Spall, J.C., A One-measurement Form of Simultaneous Perturbation Stochastic Approximation, *Automatica*, 1997, vol. 33, no. 1, pp. 109–112.
15. Granichin, O.N. and Polyak, B.T., *Randomizirovannye algoritmy otsenivaniya i optimizatsii pri pochti proizvol'nykh pomekhakh* (Randomized Algorithms for Estimation and Optimization under Almost Arbitrary Disturbances), Moscow: Nauka, 2003.
16. Granichin, O., Volkovich, V., and Toledano-Kitai, D., *Randomized Algorithms in Automatic Control and Data Mining*, Springer, 2015.
17. Popkov, A.Yu., Gradient Methods for Nonstationary Unconstrained Optimization Problems, *Autom. Remote Control*, 2005, vol. 66, no. 6, pp. 883–891.
18. Kiefer, J. and Wolfowitz, J., Stochastic Estimation of the Maximum of a Regression Function, *Ann. Math. Stat.*, 1952, vol. 23, no. 3, pp. 462–466.
19. Vakhitov, A.T., Granichin, O.N., and Gurevich, L.S., Algorithm for Stochastic Approximation with Trial Input Perturbation in the Nonstationary Problem of Optimization, *Autom. Remote Control*, 2009, vol. 70, no. 11, pp. 1827–1835.
20. Granichin, O. and Amelina, N., Simultaneous Perturbation Stochastic Approximation for Tracking under Unknown but Bounded Disturbances, *IEEE Trans. Autom. Control*, 2015, vol. 60, no. 6, pp. 1653–1658.
21. Shibaev, I.A., *Bezgradientnye metody optimizatsii dlya funktsii s gel'derovym gradientom* (Gradient-free Optimization Methods for Functions with Hölder Gradient), PhD Dissertation, MIPT, 2024, Dolgoprudny.
22. Shibaev, I., Dvurechensky, P., and Gasnikov, A., Zeroth-order Methods for Noisy Hölder-gradient Functions, *Optim. Lett.*, 2022, vol. 16, pp. 2123–2143.
23. Mandelbrot, B., New Methods in Statistical Economics, *J. Polit. Econ.*, 1963, vol. 71, no. 5, pp. 421–440.
24. Vakhitov, A.T., Granichin, O.N., and Sysoev, S.S., A Randomized Stochastic Optimization Algorithm: Its Estimation Accuracy, *Autom. Remote Control*, 2006, vol. 67, no. 4, pp. 589–597.
25. Granichin, O.N., Stochastic Approximation Search Algorithms with Randomization at the Input, *Autom. Remote Control*, 2015, vol. 76, no. 5, pp. 762–775.
26. Min, T. et al., Understanding Impact of Human Feedback via Influence Functions, *ArXiv preprint arXiv:2501.05790*, 2025.
27. Shen, W. et al., Loose Lips Sink Ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback, *Find. Assoc. Comput. Linguist.: EMNLP*, 2023, pp. 2859–2873.
28. Christiano, P.F. et al., Deep Reinforcement Learning from Human Preferences, *Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–9.
29. Stiennon, N. et al., Learning to Summarize with Human Feedback, *Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 3008–3021.
30. Ouyang, L. et al., Training Language Models to Follow Instructions with Human Feedback, *Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 27730–27744.
31. Gans, N., Koole, G., and Mandelbaum, A., Telephone Call Centers: Tutorial, Review, and Research Prospects, *Manuf. Serv. Oper. Manag.*, 2003, vol. 5, no. 2, pp. 79–141.
32. Anderson, C., *The Long Tail: Why the Future of Business is Selling Less of More*, Hyperion, 2006.
33. Goel, S., Broder, A., Gabrilovich, E., and Pang, B., Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes, *Proc. 3rd ACM Int. Conf. Web Search Data Min. (WSDM)*, New York, Feb. 4–6, 2010, pp. 201–210.
34. Akinfiev, I. and Tarasova, E., Cluster-Aware LVP: Enhancing Task Allocation with Growth Dynamics, *Proc. 15th IFAC Workshop Adapt. Learn. Control Syst. (ALCOS)*, Mexico City, Jul. 2–4, 2025.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 718–739.

SPECIAL ISSUE

On Robust Recovery of Signals from Indirect Observations

Ya. Bekri[*,a], **A. Nemirovski**[**,b], **and A. Juditsky**[*,c]

[*] *LJK, Université Grenoble Alpes, Campus de Saint-Martin-d'Hères, 38401 France*
[**] *Georgia Institute of Technology, Atlanta, Georgia, 30332, USA*
e-mail: [a] *yannis.bekri@univ-grenoble-alpes.fr,* [b] *nemirovs@isye.gatech.edu,*
[c] *anatoli.juditsky@univ-grenoble-alpes.f*

Received March 3, 2025
Revised May 20, 2025
Accepted June 27, 2025

Abstract—We consider an *uncertain linear inverse problem* as follows. Given observation $\omega = Ax_* + \zeta$ where $A \in \mathbf{R}^{m\times p}$ and $\zeta \in \mathbf{R}^m$ is observation noise, we want to recover unknown signal x_*, known to belong to a convex set $\mathcal{X} \subset \mathbf{R}^n$. As opposed to the "standard" setting of such a problem, we suppose that the model noise ζ is "corrupted"—contains an uncertain (deterministic dense or singular) component. Specifically, we assume that ζ decomposes into $\zeta = N\nu_* + \xi$ where ξ is the random noise and $N\nu_*$ is the "adversarial contamination" with known $\mathcal{N} \subset \mathbf{R}^n$ such that $\nu_* \in \mathcal{N}$ and $N \in \mathbf{R}^{m\times n}$. We consider two "uncertainty setups" in which $\mathcal{N}$ is either a convex bounded set or is the set of sparse vectors (with at most s nonvanishing entries). We analyse the performance of "uncertainty-immunized" *polyhedral estimates*—a particular class of nonlinear estimates as introduced in [19, 20]—and show how "presumably good" estimates of the sort may be constructed in the situation where the signal set is an *ellitope* (essentially, a symmetric convex set delimited by quadratic surfaces) by means of efficient convex optimization routines.

Keywords: robust estimation, linear inverse problems with contaminated observations, signal estimation in singular noise

DOI: 10.31857/S0005117925080038

1. SITUATION AND GOALS

1.1. Introduction

Since the term was coined in the 1950s, the problem of robust estimation has received much attention in the classical statistical literature. It is impossible to give an overview of the existing literature on robust estimation, and we do not try to do it here; for the "classical" framework one may refer to early references in [39], the foundational manuscript [16], or a recent survey [41].[1]

In this paper, our focus is on robust estimation of a signal from indirect linear observations. Specifically, suppose that our objective is to recover a linear image $w_* = Bx_*$ of unknown signal x_*, known to belong to a given convex set $\mathcal{X} \subset \mathbf{R}^p$, given $B \in \mathbf{R}^{q\times p}$, $A \in \mathbf{R}^{m\times p}$, and a noisy observation

$$\omega = Ax_* + \eta_* + \xi \in \mathbf{R}^m \tag{1}$$

of x_*, perturbed by a mixed—random-and-deterministic noise $\xi + \eta_*$. Here ξ is the random noise component, while η_* is the "adversarial" deterministic noise. Recently, this problem attracted much attention in the context of robust recovery of sparse (with at most $s \ll p$ nonvanishing entries) signal x_*. In particular, robust sparse regression with an emphasis on contaminated design was

[1] An important contribution to the development of robust statistics has been the line of work on distributionally robust algorithms of stochastic optimization and system identification by B. Polyak and Ya. Tsypkin, see [31–36].

investigated in [1, 5, 9, 25, 29]; methods based on penalizing the vector of outliers were studied in [7, 13], see also [3, 37]. We refer to the monograph [8] for the description of the present state of the art.

In this paper, our emphasis is on rather different assumptions about the structure of the signal x_* to be recovered and on the contamination η_* what precludes direct comparison with the cited results. Namely, we assume that the set $\mathcal{X}$ of signals is an *ellitope*—a convex compact symmetric w.r.t. the origin subset of $\mathbf{R}^p$ delimited by quadratic surfaces.[2] Our interest in ellitopes is motivated by the fact that these signal sets are well suited for the problem of estimating unknown signal x_* from observation (1) in the Gaussian no-nuisance case ($\eta_* = 0$, $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$). Specifically, let us consider linear estimate $\widehat{w}_{\rm lin}(\omega) = G_{\rm lin}^T\omega$ and *polyhedral estimate* $\widehat{w}_{\rm poly}(\omega) = B\widehat{x}_{\rm poly}(\omega)$ where

$$\widehat{x}_{\rm poly}(\omega) \in \operatorname*{Argmin}_{x\in\mathcal{X}} \|G_{\rm poly}^T(Ax - \omega)\|_\infty$$

of w_*. Let $\mathcal{X}$ be an ellitope, and let the estimation error be measured in a co-ellitopic norm $\|\cdot\|$ (i.e., such that the unit ball $\mathcal{B}_*$ of the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$ is an ellitope). In this situation, one can point out (cf. [17, 19, 20]) efficiently computable *contrast matrices* $G_{\rm lin} \in \mathbf{R}^{m\times q}$ and $G_{\rm poly} \in \mathbf{R}^{m\times m}$ such that estimates $\widehat{w}_{\rm lin}(\cdot)$ and $\widehat{w}_{\rm poly}(\cdot)$ attain nearly minimax-optimal performance.

We suppose that the adversarial perturbation η_* has a special structure: we are given a "nuisance set" $\mathcal{N} \subset \mathbf{R}^n$ such that $\nu_* \in \mathcal{N}$ and a $m \times n$ matrix N such that $\eta_* = N\nu_*$. We consider two types of assumptions about $\mathcal{N}$: $\mathcal{N}$ is either 1) a (nonempty) compact convex set, or, more conventionally, 2) $\mathcal{N}$ is the set of sparse disturbances (with at most $s \leqslant n$ nonvanishing components). Our focus is on the design and performance analysis of the "uncertainty immunized" polyhedral estimate $\widehat{w}(\omega)$ of $w_* = Bx_*$ in the presence of the contaminating signal, and solving the problem in the first case leads to a "presumably good" solution for the second.

We would like to emphasize the principal feature of the approach we promote: in this paper, A, B, and N are "general" matrices of appropriate dimensions, while $\mathcal{X}$ and $\mathcal{N}$ are rather general sets. As a consequence, we adopt here an "operational" approach[3] initiated in [10] and further developed in [18–20, 22], within which both the estimates and their risks are yielded by efficient computation, rather than by an explicit analytical analysis, seemingly impossible under the circumstances. The term "efficient" in the above is essential and is also responsible for the principal limitations of the results to follow. First of all, it imposes restrictions on the structure of the set of signals of interest and on the norm quantifying the estimation error. As it is shown in [20], the maximum of a quadratic form over an ellitope admits a "reasonably tight" efficiently computable upper bound, leading to tight bounds on the risk of linear and polyhedral estimates when the signal set is an ellitope. Furthermore, while in the case of convex compact set $\mathcal{N}$ of contaminations, constructing risk bounds for the polyhedral estimate $\widehat{w}_G(\omega)$ associated with a *given* contrast matrix G is possible under rather weak assumptions about the nuisance set $\mathcal{N}$ (essentially, the computational tractability[4] of this set is sufficient), the fundamental problem of *contrast synthesis*—minimizing these bounds over contrast matrices—allows for efficiently computable solution only when $\mathcal{N}$ is either an ellitope itself, or is a "co-ellitope" (the polar of an ellitope).

To complete this section, we would like to mention another line of research on the problem of estimating signal x_* from observation (1) under purely deterministic disturbance (case of $\xi = 0$),

[2] See [20, Section 4.2.1] or Section 1.3 below; as of now, an instructive example of ellitope is an intersection of a finite family of ellipsoids/elliptic cylinders with a common center.

[3] As opposed to the classical "descriptive" approach to solving the estimation problem in question via deriving and optimizing, w.r.t. estimate parameters, closed-form analytical expressions for the risk of a candidate estimate.

[4] For most practical purposes, computational tractability of a est means that we can model the set constraint using the CVX [15]. For an "executive summary" of what these words actually mean, we refer the reader to [20, Appendix A].

the standard problem of optimal recovery [26, 27] and guaranteed estimation in dynamical systems under uncertain-but-bounded perturbation [6, 12, 14, 23, 24, 28, 38]. The present work may be seen as an attempt to extend the corresponding framework to the case in which both deterministic and random observation noises are present.

Organization of the paper. We introduce the exact statement of the estimation problem to be considered and the entities that are relevant for the analysis to follow in Section 1.2. Analysis and design of the polyhedral estimate in the case of uncertain-but-bounded contamination are presented in Section 2. Then in Section 3 we describe the application of the proposed framework to the case of (unbounded) singular contamination using the sparse model of the nuisance vector. Finally, we recall some results on ℓ_1 recovery used in the paper in Appendix.

Notation. In the sequel, order relations between vectors are understood entry-wise; e.g., $t \geqslant t'$ for $t, t' \in \mathbf{R}^n$ means that vector $t - t'$ has nonnegative entries. $[A; B]$ and $[A, B]$ stand for vertical and horizontal concatenation of matrices A and B of appropriate dimensions. We denote $\mathbf{S}^m$ the space of symmetric $m \times m$ matrices, $\mathbf{S}^m_+$ denotes the positive semidefinite cone of $\mathbf{S}^m$; notation $A \succeq B$ ($A \succ B$) means that matrix $A - B$ is positive semidefinite (respectively, positive definite). In what follows, for a nonempty compact set $\mathcal{Z} \subset \mathbf{R}^N$

$$\phi_{\mathcal{Z}}(\zeta) := \max_{z \in \mathcal{Z}} \zeta^T t$$

is the support function of $\mathcal{Z}$. We denote $\mathrm{n}[G]$ the maximum of Euclidean norms of the columns of a matrix G.

1.2. The Problem

The estimation problem we are interested in is as follows:

Recall that we are given observation (cf. (1))

$$\omega = Ax_* + N\nu_* + \xi \in \mathbf{R}^m \tag{2}$$

where

- $N \in \mathbf{R}^{m \times n}$, $A \in \mathbf{R}^{m \times p}$ are given matrices,
- $\nu_* \in \mathbf{R}^n$ is unknown *nuisance signal*, $\nu_* \in \mathcal{N}$, a known subset of $\mathbf{R}^n$,
- x_* is an unknown signal of interest known to belong to a given convex compact set $\mathcal{X} \subset \mathbf{R}^p$ symmetric w.r.t. the origin,
- $\xi \sim P_\xi$ is a random observation noise.

Given ω, *our objective* is to recover the linear image $w_* = Bx_*$ of x_*, B being a given $q \times p$ matrix.

Given $\epsilon \in (0, 1)$, we quantify the quality of the recovery $\widehat{w}(\cdot)$ by its ϵ-risk[5]

$$\mathrm{Risk}_\epsilon[\widehat{w}] = \inf \left\{ \rho : \mathrm{Prob}_{\xi \sim P_\xi}\{\xi : \|Bx_* - \widehat{w}(\omega)\| > \rho\} \leqslant \epsilon \quad \forall (\nu_* \in \mathcal{N} \text{ and } x_* \in \mathcal{X}) \right\}$$

where $\|\cdot\|$ is a given norm.

Observation noise assumption. In the sequel, we assume that the observation noise ξ is sub-Gaussian with parameters $(0, \sigma^2 I_m)$, that is,

$$\mathbf{E}\left\{e^{h^T\xi}\right\} \leqslant \exp\left(\frac{1}{2}\sigma^2\|h\|_2^2\right).$$

[5] The ϵ-risk of an estimate depends, aside of ϵ and the estimate, on the "parameters" $\|\cdot\|$, $\mathcal{X}$, $\mathcal{N}$; these entities will always be specified by the context, allowing us to omit mentioning them in the notation $\mathrm{Risk}_\epsilon[\cdot]$.

1.3. Ellitopes

Risk analysis of a candidate polyhedral estimate heavily depends on the geometries of the signal set $\mathcal{X}$ and norm $\|\cdot\|$. In the sequel, we restrict ourselves to the case where $\mathcal{X}$ and the *polar* $\mathcal{B}_*$ of the unit ball of $\|\cdot\|$ are *ellitopes*.

By definition [17, 20], *a basic ellitope* in $\mathbf{R}^n$ is a set of the form

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists, t \in \mathcal{T} : x^T T_\ell x \leqslant t_\ell, \ell \leqslant L\}, \tag{3}$$

where $T_\ell \in \mathbf{S}^k$, $T_\ell \succeq 0$, $\sum_\ell T_\ell \succ 0$, and $\mathcal{T} \subset \mathbf{R}^L_+$ is a convex compact set with a nonempty interior which is monotone: whenever $0 \leqslant t' \leqslant t \in \mathcal{T}$ one has $t' \in \mathcal{T}$. An ellitope is an image of a basic ellitope under a linear mapping. We refer to L as *ellitopic dimension* of $\mathcal{X}$.

Clearly, every ellitope is a convex compact set symmetric w.r.t. the origin; a *basic ellitope*, in addition, has a nonempty interior.

Examples.

A. Bounded intersection $\mathcal{X}$ of L centered at the origin ellipsoids/elliptic cylinders $\{x \in \mathbf{R}^n : x^T T_\ell x \leqslant 1\}$ $[T_\ell \succeq 0]$ is a basic ellitope:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} := [0,1]^L : x^T T_\ell x \leqslant t_\ell, \ell \leqslant L\}$$

In particular, the unit box $\{x \in \mathbf{R}^n : \|x\|_\infty \leqslant 1\}$ is a basic ellitope.

B. A $\|\cdot\|_p$-ball in $\mathbf{R}^n$ with $p \in [2, \infty]$ is a basic ellitope:

$$\{x \in \mathbf{R}^n : \|x\|_p \leqslant 1\} = \left\{ x : \exists t \in \mathcal{T} = \{t \in \mathbf{R}^n_+, \|t\|_{p/2} \leqslant 1\} : \underbrace{x_\ell^2}_{x^T T_\ell x} \leqslant t_\ell, \ell \leqslant n \right\}.$$

Ellitopes admit fully algorithmic "calculus"—this family is closed with respect to basic operations preserving convexity and symmetry w.r.t. the origin, e.g., taking finite intersections, linear images, inverse images under linear embedding, direct products, arithmetic summation (for details, see [20, Section 4.6]).

Main assumption. We assume from now on that the signal set $\mathcal{X}$ and the *polar* $\mathcal{B}_*$ of the unit ball of the norm $\|\cdot\|$ are *basic ellitopes*:[6]

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^T T_k x \leqslant t_k, k \leqslant K\}, \tag{4a}$$

$$\mathcal{B}_* = \{y \in \mathbf{R}^q : \exists s \in \mathcal{S} : y^T S_\ell y \leqslant s_\ell, \ell \leqslant L\} \tag{4b}$$

where $\mathcal{T} \subset \mathbf{R}^K_+$, $\mathcal{S} \in \mathbf{R}^L_+$ are monotone convex compact sets with nonempty interiors, $T_k \succeq 0$, $\sum_k T_k \succ 0$, $S_\ell \succeq 0$, and $\sum_\ell S_\ell \succ 0$.

2. UNCERTAIN-BUT-BOUNDED NUISANCE

In this section, we consider the case of *uncertain-but-bounded* nuisance. Specifically, we assume that $\mathcal{N} \subset \mathbf{R}^n$ is a convex compact set, symmetric w.r.t. the origin, and specify $\pi(\cdot)$ as the semi-norm on $\mathbf{R}^m$ given by

$$\pi(h) = \sup_u \left\{ (Nu)^T h : u \in \mathcal{N} \right\}.$$

[6] The results to follow straightforwardly extend to the case where $\mathcal{X}$ and $\mathcal{B}_*$ are "general" ellitopes; we assume them to be basic to save notation.

2.1. Bounding the ϵ-Risk of Polyhedral Estimate

In this section, a polyhedral estimate is specified by $m \times I$ *contrast matrix* G and is as follows: given observation ω (see (2)), we find an optimal solution $\widehat{x}_G(\omega)$ to the (clearly solvable) optimization problem

$$\min_{x,\nu} \left\{ \|G^T(Ax + N\nu - \omega)\|_\infty : x \in \mathcal{X},\, \nu \in \mathcal{N} \right\}. \tag{5}$$

Given a $m \times I$ matrix $G = [g_1, \ldots, g_I]$, let $\Xi_\epsilon[G]$ be the set of all realizations of ξ such that

$$|[g_i^T \xi]_i| \leqslant \underbrace{\sigma\sqrt{2\ln[2I/\epsilon]}}_{=:\varkappa(\epsilon)} \|g_i\|_2, \quad \forall i \leqslant I. \tag{6}$$

Note that

$$\mathrm{Prob}_{\xi \sim \mathcal{SG}(0,\sigma^2 I_m)} \{\xi \notin \Xi_\epsilon(G)\} \leqslant \epsilon. \tag{7}$$

Indeed, we have $\mathbf{E}_\xi \left\{ e^{\gamma g^T \xi} \right\} \leqslant e^{\frac{1}{2}\gamma^2 \|g\|_2^2 \sigma^2}$, implying that for all $a \geqslant 0$,

$$\mathrm{Prob}\{g^T\xi > a\} \leqslant \inf_{\gamma > 0} \exp\left\{ \frac{1}{2}\gamma^2 \|g\|_2^2 \sigma^2 - \gamma a \right\} = \exp\left\{ -\frac{1}{2} a^2 \sigma^2 \|g\|_2^2 \right\},$$

so that

$$\mathrm{Prob}\left\{ \exists i \leqslant I : |g_i^T \xi| > \varkappa(\epsilon)\|g_i\|_2 \right\} \leqslant 2I \exp\left\{ -\frac{\varkappa^2(\epsilon)}{2\sigma^2} \right\} \leqslant \epsilon.$$

Given a $m \times I$ contrast matrix $G = [g_1, \ldots, g_I]$, consider the optimization problem

$$\mathrm{Opt}[G] = \min_{\lambda,\mu,\gamma} \left\{ f_G(\lambda,\mu,\gamma) :\ \lambda \geqslant 0,\ \mu \geqslant 0,\ \gamma \geqslant 0, \right. \tag{8}$$
$$\left. \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T \left[\sum_i \gamma_i g_i g_i^T\right] A \end{array} \right] \succeq 0 \right\}$$

where

$$f_G(\lambda,\mu,\gamma) = \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\psi^2[G] \sum_i \gamma_i$$

with

$$\psi[G] = \max_i \pi(g_i) + \varkappa(\epsilon)\mathrm{n}[G].$$

Proposition 1. *Let (λ, μ, γ) be a feasible solution to the problem in* (8)*. Then*

$$\mathrm{Risk}_\epsilon[\widehat{w}_G] \leqslant f_G(\lambda,\mu,\gamma),$$

i.e., the ϵ-risk of the estimate $\widehat{w}_G$ is upper bounded with $f_G(\lambda,\mu,\gamma)$.

Proof. Let us fix $\xi \in \Xi_\epsilon[G]$, $x_* \in \mathcal{X}$, and $\eta_* \in \mathcal{N}$. Let also $\widehat{x} = \widehat{x}_G(\omega)$ be the x-component of some optimal solution $[\widehat{x}; \widehat{\nu}]$, $\widehat{\nu} \in \mathcal{N}$, to (5) and, finally, let $\Delta = \widehat{x} - x_*$. Observe that $[x, \nu] = [x_*, \nu_*]$ is feasible for (5) and $\|G^T[Ax_* + N\nu_* - \omega]\|_\infty = \|G^T\xi\|_\infty \leqslant \varkappa(\epsilon)\mathrm{n}[G]$, implying that $\|G^T[A\widehat{x} + N\widehat{\nu} - \omega]\|_\infty \leqslant \varkappa(\epsilon)\mathrm{n}[G]$ as well. Therefore,

$$\|G^T A\Delta\|_\infty \leqslant 2\varkappa(\epsilon)\mathrm{n}[G] + \|G^T N[\widehat{\nu} - \nu_*]\|_\infty.$$

Taking into account that $\widehat{\nu}, \nu_* \in \mathcal{N}$, we have $\|G^T N[\widehat{\nu} - \nu_*]\|_\infty \leqslant 2\max_i \pi(g_i)$, and we arrive at

$$|g_i^T A\Delta| \leqslant 2\psi[G],\ i = 1, \ldots, I. \tag{9}$$

Now, we have $\Delta \in 2\mathcal{X}$, that is, for some $t \in \mathcal{T}$ and all k it holds $\Delta^T T_k \Delta \leqslant 4t_k$, and let $v \in \mathcal{B}_*$, so that for some $s \in \mathcal{S}$ for all ℓ it holds $v^T S_\ell v \leqslant s_\ell$. By the semidefinite constraint of (8) we have

$$\begin{aligned} v^T B\Delta &\leqslant v^T \left[\sum_\ell \lambda_\ell S_\ell\right] v + \Delta^T \left[\sum_k \mu_k T_k\right] \Delta + [A\Delta]^T \sum_i \gamma_i g_i g_i^T A\Delta \\ &\leqslant \sum_\ell \lambda_\ell s_\ell + 4\sum_k \mu_k t_k + \sum_i \gamma_i (g_i^T A\Delta)^2 \\ &\leqslant \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + \sum_i \gamma_i (g_i^T A\Delta)^2 \\ &\leqslant \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\psi^2[G] \sum_i \gamma_i. \end{aligned}$$

Maximizing the left-hand side of the resulting inequality over $v \in \mathcal{B}_*$, we arrive at $\|B\Delta\| \leqslant f_G(\lambda, \mu, \gamma)$. □

Note that the optimization problem in the right-hand side of (8) is an explicit convex optimization problem, so that $\mathrm{Opt}[G]$ is efficiently computable, provided that $\phi_{\mathcal{S}}, \phi_{\mathcal{T}}$ and π are so. Thus, Proposition 1 provides us with an efficiently computable upper bound on the ϵ-risk of the polyhedral estimate stemming from a given contrast matrix G and as such gives us a computation-friendly tool to *analyse* the performance of a polyhedral estimate. Unfortunately, this tool does not allow to *design* a "presumably good" estimate, since an attempt to make G a variable, rather than a parameter, in the right-hand side problem in (8) results in a nonconvex, and thus difficult to solve, optimization problem. We now look at two situations in which this difficulty can be overcome.

2.2. Synthesis of "Presumably Good" Contrast Matrices

We consider here two types of assumptions about the set $\mathcal{N}$ of nuisances which allow for a computationally efficient design of "presumably good" contrast matrices. Namely,

1) "ellitopic case:" $\mathcal{N}$ is a basic ellitope;
2) "co-ellitopic case:" the set $N\mathcal{N} = \{N\nu : \nu \in \mathcal{N}\}$ is the *polar* of the ellitope

$$\begin{aligned} &\mathcal{N}_* = \{w \in \mathbf{R}^m : \exists \overline{r} \in \overline{\mathcal{R}} : w^T \overline{R}_j w \leqslant \overline{r}_j, j \leqslant \overline{J}\} \\ &\left[\ \overline{R}_j \succeq 0,\ \textstyle\sum_j \overline{R}_j \succ 0;\ \overline{\mathcal{R}} \subset \mathbf{R}^{\overline{J}}_+,\ \mathrm{int}\overline{\mathcal{R}} \neq \emptyset, \text{ is a monotone convex compact}\right] \end{aligned}$$

Note that $\mathcal{N}_*$ is exactly the unit ball of the norm $\pi(g) = \max_{\nu \in \mathcal{N}} g^T N\nu$.

2.2.1. Ellitopic case. An immediate observation is that *the ellitopic case can be immediately reduced to the no-nuisance case*. Indeed, when $\mathcal{N}$ is an ellitope, so is the direct product $\overline{\mathcal{X}} = \mathcal{X} \times \mathcal{N}$. Thus, setting $\overline{A}[x; \nu] = Ax + N\nu$, $\overline{B}[x; \nu] = Bx$, observation (2) becomes

$$\omega = \overline{A}\overline{x}_* + \xi, \qquad [\overline{x_*} = [x_*; \nu_*] \in \overline{\mathcal{X}}]$$

and our objective is to recover from this observation the linear image $w_* = \overline{B}\overline{x}_*$ of the new signal $\overline{x}_*$. Design of presumably good (and near-minimax-optimal when $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$) polyhedral estimates in this setting is considered in [20]. It makes sense to sketch the construction here since it explains the idea used throughout the rest of the paper.

Thus, consider the case when $\mathcal{N} = \{0\}$, and let the signal set and the norm $\|\cdot\|$ still be given by (4). In this situation problem (8) becomes

$$\mathrm{Opt}[G] = \min_{\lambda,\mu,\gamma} \left\{ \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\varkappa^2(\epsilon)\mathrm{n}^2[G]\sum_i \gamma_i : \ \lambda \geqslant 0,\ \mu \geqslant 0,\ \gamma \geqslant 0, \right. \tag{10}$$
$$\left. \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T\left[\sum_i \gamma_i g_i g_i^T\right] A \end{array} \right] \succeq 0 \right\}.$$

Note that when $\theta > 0$, we have $\mathrm{Opt}[G] = \mathrm{Opt}[\theta G]$. Indeed, (λ, μ, γ) is a feasible solution to the problem specifying $\mathrm{Opt}[G]$ if and only if $\lambda, \mu, \theta^2\gamma)$ is a feasible solution to the problem specifying $\mathrm{Opt}[\theta G]$, and the values of the respective objectives at these solutions are the same. It follows that as far as optimization of $\mathrm{Opt}[G]$ in G is concerned, we lose nothing when restricting ourselves to contrast matrices G with $\varkappa(\epsilon)\mathrm{n}[G] = 1$. In other words, by setting

$$\theta(g) = \varkappa(\epsilon)\|g\|_2 \tag{11}$$

and augmenting variables λ, μ, and γ in (10) by variables g_i, $\theta(g_i) \leqslant 1$, $i = 1, \ldots, I$ (recall that we want to make G variable rather than parameter and to minimize $\mathrm{Opt}[G]$ over G), we arrive at the problem

$$\mathrm{Opt} = \min_{\lambda,\mu,\gamma,\{g_i\},\rho} \left\{ \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\rho : \ \lambda \geqslant 0,\ \mu \geqslant 0,\ \gamma \geqslant 0, \right. \tag{12}$$
$$\left. \theta(g_i) \leqslant 1,\ \sum_i \gamma_i \leqslant \rho,\ \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T\left[\sum_i \gamma_i g_i g_i^T\right] A \end{array} \right] \succeq 0 \right\}.$$

Now, aggregating variables γ, $g_1, \ldots, g_I$ into the matrix $\Theta = \sum_i \gamma_i g_i g_i^T$ and denoting by $\mathfrak{T}$ the set of the pairs $(\Theta \in \mathbf{S}_+^m, \rho)$ for which there exists decomposition $\Theta = \sum_{i \leqslant I} \gamma_i g_i g_i^T$ with $\theta(g_i) \leqslant 1$ and $\gamma_i \geqslant 0$, $\sum_i \gamma_i \leqslant \rho$, (12) can be rewritten as the optimization problem

$$\mathrm{Opt} = \min_{\lambda,\mu,\Theta,\rho} \left\{ \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\rho : \ \lambda \geqslant 0,\ \mu \geqslant 0, \right. \tag{13}$$
$$\left. (\Theta, \rho) \in \mathfrak{T},\ \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T\Theta A \end{array} \right] \succeq 0 \right\}.$$

Observe that when $I \geqslant m$, $\mathfrak{T}$ is a simple convex cone:

$$\mathfrak{T} = \{(\Theta, \rho) : \ \Theta \succeq 0,\ \rho \geqslant \varkappa^2(\epsilon)\mathrm{Tr}(\Theta)\},$$

so that (13) is an explicit (and clearly solvable) convex optimization program. To convert an optimal solution $(\lambda^*, \mu^*, \Theta^*, \rho^*)$ to (13) into an optimal solution to (12), it suffices to subject Θ^* to the eigenvalue decomposition $\Theta^* = \sum_{i=1}^I \upsilon_i e_i e_i^T$ with $\|e_i\|_2 = 1$ and $\upsilon_i \geqslant 0$, $i \in \{1, \ldots, m\}$, and $e_i = 0$, $\upsilon_i = 0$, $i \in \{m, \ldots, I\}$, and set $g_i^* = \varkappa^{-1}(\epsilon)e_i$, $\gamma_i^* = \varkappa^2(\epsilon)\upsilon_i$, thus arriving at an optimal solution $(\lambda^*, \mu^*, \{g_i^*, \gamma_i^*\}_{i \leqslant I}, \rho^*)$ to problem (12).

2.2.2. Co-ellitopic case. The just outlined approach to reducing the nonconvex problem (12) responsible for the design of the best, in terms of $\mathrm{Opt}[G]$, contrast matrix G to an explicit convex optimization problem heavily utilizes the fact that the unit ball of the norm $\theta(\cdot)$ (cf. (11)) is a simple ellitope—a multiple of the unit Euclidean ball; this was the reason for $\mathfrak{T}$ to be a computationally tractable convex cone. Our future developments are built on the fact that when the unit ball of $\theta(\cdot)$ is a basic ellitope, something similar takes place: the associated set $\mathfrak{T}$, while not necessarily convex and computationally tractable, can be tightly approximated by a computationally tractable convex cone. The underlying result (which is [21, Proposition 3.2], up to notation) is as follows:

Proposition 2. *Let $I \geqslant m$, and let $\mathcal{W} \subset \mathbf{R}^m$ be a basic ellitope:*

$$\mathcal{W} = \{w \in \mathbf{R}^m : \exists r \in \mathcal{R} : w^T R_j w \leqslant r_j,\ j \leqslant J\}$$
$$\left[R_j \succeq 0,\ \textstyle\sum_j R_j \succ 0;\ \mathcal{R} \subset \mathbf{R}^J_+,\ \mathrm{int}\mathcal{R} \neq \emptyset,\ \text{is a monotone convex compact}\right]$$

Let us associate with $\mathcal{W}$ the closed convex cone[7]

$$\mathbf{K} = \{(\Theta, \rho) : \exists r \in \mathcal{R} : \mathrm{Tr}(\Theta R_j) \leqslant \rho r_j,\ 1 \leqslant j \leqslant J,\ \Theta \succeq 0,\ \rho \geqslant 0\}.$$

Whenever a matrix $\Theta \in \mathbf{S}^m_+$ is representable as $\sum_i \gamma_i w_i w_i^T$ with $\gamma_i \geqslant 0$ and $w_i \in \mathcal{W}$, one has $(\Theta, \sum_{i=1}^I \gamma_i) \in \mathbf{K}$, and nearly vice versa: whenever $(\Theta, \rho) \in \mathbf{K}$, one can find efficiently (via a randomized algorithm) vectors $w_i \in \mathcal{W}$, and reals $\gamma_i \geqslant 0$, $i \leqslant I$, such that $\Theta = \sum_i \gamma_i w_i w_i^T$ and and

$$\sum_i \gamma_i \leqslant 2\sqrt{2}\ln(4m^2J)\rho.$$

We are now ready to outline a "presumably good" contrast design in the co-ellitopic case. Let us put $R_j = \frac{1}{4}\overline{R}_j$, $j \leqslant \overline{J}$, and $R_{\overline{J}+1} = \frac{\varkappa^2(\epsilon)}{4} I_m$ and consider the ellitope

$$\begin{aligned}\mathcal{W} &= 2\left[\mathcal{N}_* \cap \{w :\ \varkappa(\epsilon)\|w\|_2 \leqslant 1\}\right] \\ &= \left\{w \in \mathbf{R}^m :\ \exists r \in \mathcal{R} = \overline{\mathcal{R}} \times [0,1] :\ w^T R_j w \leqslant r_j,\ j \leqslant J = \overline{J}+1\right\},\end{aligned} \tag{14}$$

and let $\theta(\cdot)$ be the norm on $\mathbf{R}^n$ with the unit ball $\mathcal{W}$. Note that $\theta(\cdot) = 2\max\left[\pi(\cdot),\ \varkappa(\epsilon)\|\cdot\|_2\right]$, so that for every $G = [g_1, .., g_I]$, the quantity $\psi[G]$, see (8), is upper-bounded by $\max_i \theta(g_i)$, and this bound is tight within the factor 2. Consequently, Proposition 1 states that the ϵ-risk of the polyhedral estimate with contrast matrix G is upper-bounded by the quantity

$$\begin{aligned}\overline{\mathrm{Opt}}[G] = \min_{\lambda,\mu,\gamma}\Bigg\{&\phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\left[\max_i \theta(g_i)\right]^2 \sum_i \gamma_i :\ \lambda \geqslant 0,\ \mu \geqslant 0,\ \gamma \geqslant 0, \\ &\left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T\left[\sum_i \gamma_i g_i g_i^T\right] A \end{array}\right] \succeq 0\Bigg\}\end{aligned} \tag{15}$$

and $\mathrm{Opt}[G] \leqslant \overline{\mathrm{Opt}}[G] \leqslant 2\mathrm{Opt}[G]$. As in the previous section, the problem of minimizing $\overline{\mathrm{Opt}}[G]$ over G can be reformulated in the form (13). A computationally efficient way to get a tight approximation to the optimal solution of the latter problem is given by the following result.

Let $I \geqslant m$, $\alpha = 2\sqrt{2}\ln(4m^2J)$, and let

$$\mathbf{K} = \{(\Theta, \rho) :\ \exists r \in \mathcal{R} :\ \mathrm{Tr}(\Theta R_j) \leqslant \rho r_j,\ 1 \leqslant j \leqslant J,\ \Theta \succeq 0,\ \rho \geqslant 0\}$$

(see (14)). Consider the convex optimization problem

$$\begin{aligned}\mathrm{Opt}_* = \min_{\lambda,\mu,\gamma,\Theta,\rho}\Bigg\{&\phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + 4\alpha\rho :\ \lambda \geqslant 0,\ \mu \geqslant 0, \\ &(\Theta,\rho) \in \mathbf{K},\ \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T\Theta A \end{array}\right] \succeq 0\Bigg\}.\end{aligned} \tag{16}$$

[7] This indeed is a closed convex cone—the conic hull of the convex compact set $\{\Theta \succeq 0 :\ \exists r \in \mathcal{R} :\ \mathrm{Tr}(\Theta R_j) \leqslant r_j,\ 1 \leqslant j \leqslant J\} \times \{1\}$.

Theorem 1. *One can convert, in a computationally efficient way, the Θ-component Θ^* of an optimal solution to the (clearly solvable) problem* (16) *into the contrast matrix G^* such that*

$$\overline{\mathrm{Opt}}[G^*] \leqslant \sqrt{\alpha}\min_G \overline{\mathrm{Opt}}[G] \leqslant 2\sqrt{\alpha}\min_G \mathrm{Opt}[G].$$

In particular, the ϵ-risk of the polyhedral estimate with contrast matrix G^* (this risk is upper-bounded by $\overline{\mathrm{Opt}}[G^*]$) does not exceed $2\sqrt{\alpha}\min_G \mathrm{Opt}[G]$.

Proof. When repeating the reasoning in the previous section, we conclude that $\overline{\mathrm{Opt}} := \inf_G \overline{\mathrm{Opt}}[G]$ is equal to

$$\inf_{g_1,\ldots,g_I}\left\{\overline{\mathrm{Opt}}([g_1,\ldots,g_I]):\ \max_i \theta(g_i)=1\right\}.$$

The latter inf is clearly attained at certain collection $g_i^+,\ldots,g_I^+$ with $\max_i \theta(g_i^+)=1$. Let $G^+=[g_1^+,\ldots,g_I^+]$, let $\lambda^+,\mu^+,\gamma_i^+$, $i\leqslant I$, be an optimal solution to the problem in the right-hand side of (15) associated with $g_i=g_i^+$, $i\leqslant I$, and let $\Theta^+=\sum_i \gamma_i^+[g_i^+][g_i^+]^T$ and $\rho^+=\sum_i\gamma_i^+$. We clearly have

$$\overline{\mathrm{Opt}} = \overline{\mathrm{Opt}}[G^+] = \phi_{\mathcal{S}}(\lambda^+)+4\phi_{\mathcal{T}}(\mu^+)+4\rho^+.$$

Besides this, we are in the case where $\theta(g)\leqslant 1$ is equivalent to $g\in\mathcal{W}$, and therefore, by the first claim in Proposition 2, $(\Theta^+,\rho^+)\in\mathbf{K}$, implying that $(\lambda^+,\mu^+,\Theta^+,\rho^+)$ is a feasible solution to the optimization problem in (16). Due to the structure of the latter problem, for $\kappa>0$ the collection $(\kappa^{-1}\lambda^+,\kappa\mu^+,\kappa\Theta^+,\kappa\rho^+)$ is feasible for (16) with the corresponding value of the objective $\kappa^{-1}\phi_{\mathcal{S}}(\lambda^+)+\kappa[\phi_{\mathcal{T}}(\mu^+)+4\alpha\rho^+]$. It follows that

$$\begin{aligned}\mathrm{Opt}_* &\leqslant \inf_{\kappa>0}\left[\kappa^{-1}\phi_{\mathcal{S}}(\lambda^+)+\kappa[4\phi_{\mathcal{T}}(\mu^+)+4\alpha\rho^+]\right]\\ &= 2\Big(\phi_{\mathcal{S}}(\lambda^+)\underbrace{[4\phi_{\mathcal{T}}(\mu^+)+4\alpha\rho^+]}_{\leqslant\alpha[4\phi_{\mathcal{T}}(\mu^+)+4\rho^+]}\Big)^{1/2} \leqslant 2\sqrt{\phi_{\mathcal{S}}(\lambda^+)[4\phi_{\mathcal{T}}(\mu^+)+4\rho^+]}\sqrt{\alpha}\\ &\leqslant \sqrt{\alpha}[\phi_{\mathcal{S}}(\lambda^+)+4\phi_{\mathcal{T}}(\mu^+)+4\rho^+] = \sqrt{\alpha}\,\overline{\mathrm{Opt}}.\end{aligned}$$

Finally, let $\overline{\lambda},\overline{\mu},\overline{\Theta},\overline{\rho}$ be an optimal solution to (16). As $(\overline{\Theta},\overline{\rho})\in\mathbf{K}$, the second claim in Proposition 2 states that there exists (and can be efficiently found) decomposition $\overline{\Theta}=\sum_i\overline{\gamma}_i[\overline{g}_i][\overline{g}]_i^T$ with $\overline{g}_i\in\mathcal{W}$ (i.e., $\theta(\overline{g}_i)\leqslant 1$), $i\leqslant I$, $\overline{\gamma}_i\geqslant 0$, and $\sum_i\overline{\gamma}_i\leqslant\alpha\overline{\rho}$. The ϵ-risk of the polyhedral estimate with the contrast matrix $\overline{G}=[\overline{g}_1,\ldots,\overline{g}_I]$ is then upper-bounded by $\overline{\mathrm{Opt}}[\overline{G}]$. However, $\overline{\lambda},\overline{\mu}$, and $\{\overline{\gamma}_i\}$ form a feasible solution to the problem specifying $\overline{\mathrm{Opt}}[\overline{G}]$, and the value of the objective at this solution is upper bounded with

$$\phi_{\mathcal{S}}(\overline{\lambda})+4\phi_{\mathcal{T}}(\overline{\mu})+4[\max_i\theta(\overline{g}_i)]\sum_i\overline{\gamma}_i \leqslant \phi_{\mathcal{S}}(\overline{\lambda})+4\phi_{\mathcal{T}}(\overline{\mu})+4\alpha\overline{\rho}=\mathrm{Opt}_*.$$

Thus, the ϵ-risk of the polyhedral estimate with contrast matrix $\overline{G}$ does not exceed

$$\mathrm{Opt}_*\leqslant\sqrt{\alpha}\,\overline{\mathrm{Opt}}\leqslant 2\sqrt{\alpha}\min_G\mathrm{Opt}[G].$$ □

3. OBSERVATIONS WITH OUTLIERS

In this section, we consider the estimation problem posed in Section 1.2 in the situation where the nuisance ν_* in (2) is sparse—has at most a given number s of nonzero entries.

Estimate construction. Let $\epsilon\in(0,1)$ be a given reliability tolerance. We consider the polyhedral estimate specified by two contrast matrices $H=[h_1,\ldots,h_n]\in\mathbf{R}^{m\times n}$ and $G=[g_1,\ldots,g_I]\in\mathbf{R}^{n\times I}$ which is as follows. Given observation ω (see (2)) we solve the optimization problem

$$\min_{\nu,x}\left\{\|\nu\|_1:\ x\in\mathcal{X},\ \begin{array}{ll}|h_k^T[N\nu+Ax-\omega]|\leqslant\overline{\varkappa}(\epsilon)\|h_k\|_2, & k=1,\ldots,n,\\ |g_i^T[N\nu+Ax-\omega]|\leqslant\overline{\varkappa}(\epsilon)\|g_i\|_2, & i=1,\ldots,I\end{array}\right\}, \tag{17}$$

where

$$\overline{\varkappa}(\epsilon) = \sigma\sqrt{2\ln[2(n+I)/\epsilon]}.$$

Let $(\widehat{\nu}, \widehat{x}) = (\widehat{\nu}(\omega), \widehat{x}(\omega))$ be an optimal solution to the problem when the problem is feasible, otherwise we put $(\widehat{\nu}, \widehat{x}) = (0, 0)$. Vector

$$\widehat{w}_{G,H}(\omega) = B\widehat{x}(\omega)$$

is the estimate of $w_* = Bx_*$ we intend to use.

3.1. Risk Analysis

Let us denote $\Xi_\epsilon(G, H)$ the set of realizations of ξ such that

$$|h_k^T\xi| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2,\ k = 1, \ldots, n, \quad |g_i^T\xi| \leqslant \overline{\varkappa}(\epsilon)\|g_i\|_2,\ i = 1, \ldots, I, \quad \forall \xi \in \Xi_\epsilon(G, H). \tag{18}$$

For the same reasons as in (7), one has

$$\mathrm{Prob}_{\xi\sim\mathcal{SG}(0,\sigma^2 I_m)}(\Xi_\epsilon(G, H)) \geqslant 1 - \epsilon.$$

Let us now fix $x_* \in \mathcal{X}$, s-sparse ν_*, and $\xi \in \Xi_\epsilon(G, H)$, so that our observation is $\omega = Ax_* + N\nu_* + \xi$. **A.** By (18) we have $|h_k^T\xi| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2$ and $|g_i^T\xi| \leqslant \overline{\varkappa}(\epsilon)\|g_i\|_2$ for all $k \leqslant n$ and $i \leqslant I$, while (17) becomes the problem

$$\min_{\nu,x}\left\{\|\nu\|_1 :\ x \in \mathcal{X},\ \begin{array}{ll} |h_k^T[N[\nu - \nu_*] + A[x - x_*] - \xi]| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2, & k = 1, \ldots, n, \\ |g_i^T[N[\nu - \nu_*] + A[x - x_*] - \xi]| \leqslant \overline{\varkappa}(\epsilon)\|g_i\|_2, & i = 1, \ldots, I \end{array}\right\}. \tag{19}$$

We conclude that $(\nu, x) = (\nu_*; x_*)$ is a feasible solution to (17). Thus, we are in the case where $\widehat{\nu}$, $\widehat{x}$ are feasible for (19), and

$$\|\widehat{\nu}\|_1 \leqslant \|\nu_*\|_1.$$

B. Assume from now on that $(H, \|\cdot\|_\infty)$ satisfies Condition $Q_\infty(s, \kappa)$ of Section 3.5 with $\kappa < \frac{1}{2}$, that is,[8]

$$\|w\|_\infty \leqslant \|H^T N w\|_\infty + \frac{\kappa}{s}\|w\|_1 \quad \forall w \in \mathbf{R}^n. \tag{20}$$

Since $\widehat{\nu}$ and $\widehat{x}$ are feasible for (19), we have

$$|h_k^T[N[\widehat{\nu} - \nu_*] + A[\widehat{x} - x_*] - \xi]| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2,\ \forall k \leqslant n.$$

Invoking (18) and the fact that $A[\widehat{x} - x_*] \in 2A\mathcal{X}$ (since $\mathcal{X}$ is symmetric w.r.t. the origin), we conclude that

$$\|H^T N[\widehat{\nu} - \nu_*]\|_\infty \leqslant \max_k\left[\overline{\varkappa}(\epsilon)\|h_k\|_2 + 2\max_{x\in\mathcal{X}}|h_k^T Ax|\right],$$

and besides this, ν_* is s-sparse and $\|\widehat{\nu}\|_1 \leqslant \|\nu_*\|_1$. Now Proposition 5 with ν_* in the role of ν implies that

$$\|\widehat{\nu} - \nu_*\|_q \leqslant \frac{(2s)^{\frac{1}{q}}}{1 - 2\kappa}\max_k\left[\overline{\varkappa}(\epsilon)\|h_k\|_2 + 2\max_{x\in\mathcal{X}}|h_k^T Ax|\right], \quad 1 \leqslant q \leqslant \infty, \tag{21}$$

[8] Condition $Q_\infty(s, \kappa)$ is the simplest (and the most restrictive) member of the family $Q_q(s, \kappa)$, $q \in [1, \infty]$ of conditions used to establish the properties of ℓ_1-recovery of sparse signals. The property of this condition crucial here is that it can be efficiently verified. We refer to [20, Section 1.3] for the discussion of efficiently verifiable conditions in sparse recovery and their relation to other conditions used (Restricted Isometry Property (RIP) [4], Restricted Eigenvalue (RE) [2], Mutual Incoherence (MI) [11], and Compatibility [40]).

in particular, that

$$\|\widehat{\nu}-\nu_*\|_\infty \leqslant \frac{1}{1-2\kappa}\max_k\left[\overline{\varkappa}(\epsilon)\|h_k\|_2+2\max_{x\in\mathcal{X}}|h_k^TAx|\right] =: \rho_H, \tag{22a}$$

$$\|\widehat{\nu}-\nu_*\|_1 \leqslant \frac{2s}{1-2\kappa}\max_k\left[\overline{\varkappa}(\epsilon)\|h_k\|_2+2\max_{x\in\mathcal{X}}|h_k^TAx|\right] = 2s\rho_H. \tag{22b}$$

In addition, [20, Proposition 1.10] states that the set $\mathcal{H}$ of the pairs (H,κ) with $m\times n$ matrices H satisfying Condition $Q_\infty(s,\kappa)$ is the computationally tractable convex set

$$\mathcal{H}=\left\{(H,\kappa)\in\mathbf{R}^{m\times n}\times\mathbf{R}:\ |[I_n-N^TH]_{ij}|\leqslant s^{-1}\kappa,\ 1\leqslant i,j\leqslant n\right\}. \tag{23}$$

C. Since $\widehat{\nu}$ and $\widehat{x}$ are feasible for (19), we have

$$|g_i^T\left(N[\widehat{\nu}-\nu_*]+A[\widehat{x}-x_*]-\xi\right)|\leqslant\overline{\varkappa}(\epsilon)\|g_i\|_2,\quad i=1,\dots,I,$$

while $|g_i^T\xi|\leqslant\overline{\varkappa}(\epsilon)\|g_i\|_2\ \forall i$ due to $\xi\in\Xi_\epsilon(G,H)$. We conclude that

$$|g_i^TA[\widehat{x}-x_*]|\leqslant 2\overline{\varkappa}(\epsilon)\|g_i\|_2+|g_i^TN[\widehat{\nu}-\nu_*]|,\ i\leqslant I. \tag{24}$$

Let $\|z\|_{k,1}$, $z\in\mathbf{R}^n$, be the sum of $\min[k,n]$ largest magnitudes of entries in z; note that $\|\cdot\|_{k,1}$ is the norm conjugate to the norm with the unit ball $\{u:\|u\|_\infty\leqslant 1,\|u\|_1\leqslant k\}$. Consequently, (22) implies that

$$|g_i^TN[\widehat{\nu}-\nu_*]|\leqslant\rho_H\|N^Tg_i\|_{2s,1}, \tag{25}$$

and, therefore, by (24)

$$|g_i^TA[\widehat{x}-x_*]|\leqslant\psi_H[G],\ \psi_H[G]=\max_i\left[2\overline{\varkappa}(\epsilon)\|g_i\|_2+\rho_H\|N^Tg_i\|_{2s,1}\right]. \tag{26}$$

Let

$$f_{G,H}(\lambda,\mu,\gamma)=\phi_{\mathcal{S}}(\lambda)+4\phi_{\mathcal{T}}(\mu)+\psi_H^2[G]\sum_i\gamma_i,$$

and let us consider the optimization problem (cf. (8))

$$\mathrm{Opt}[G,H]=\min_{\lambda,\mu,\gamma}\left\{f_{G,H}(\lambda,\mu,\gamma):\right.$$
$$\left.\lambda\geqslant 0,\ \mu\geqslant 0,\ \gamma\geqslant 0,\ \left[\begin{array}{c|c}\sum_\ell\lambda_\ell S_\ell & \frac{1}{2}B\\ \hline \frac{1}{2}B^T & \sum_k\mu_kT_k+A^T\left[\sum_i\gamma_ig_ig_i^T\right]A\end{array}\right]\succeq 0.\right\} \tag{27}$$

Applying the same argument as in the proof of Proposition 1, with (26) in the role of (9), we arrive at the following result:

Proposition 3. *In the situation of this section given $\kappa\in(0,1/2)$ and $m\times n$ matrix H satisfying $(H,\kappa)\in\mathcal{H}$, see* (23)*, let (λ,μ,γ) be a feasible solution to* (27)*. Then*

$$\mathrm{Risk}_\epsilon[\widehat{w}_{G,H}]\leqslant f_{G,H}(\lambda,\mu,\gamma).$$

3.2. Synthesis of Contrast Matrices

Our present objective is to design contrast matrices H and G with a small value of the bound $\mathrm{Opt}[G,H]$ for the ϵ-risk of the estimate $\widehat{w}_{G,H}$.

D. Building the contrast matrix $H \in \mathbf{R}^{m\times n}$ is straightforward: the risk bound $\mathrm{Opt}[G,H]$, depends on $H=[h_1,\dots,h_n]$ solely through the quantity

$$\rho_H = \frac{1}{1-2\kappa}\max_{k\leqslant n}\left[\overline{\varkappa}(\epsilon)\|h_k\|_2 + 2\max_{x\in\mathcal{X}}\|h_k^TAx\|_\infty\right]$$

and is smaller the smaller is ρ_H. For a fixed $\kappa\in(0,1/2)$, a presumably good choice of $H=[h_1,\dots,h_n]$ is then given by optimal solutions to n optimization problems

$$h_k = \operatorname*{argmin}_h\left\{\overline{\varkappa}(\epsilon)\|h\|_2 + 2\max_{x\in\mathcal{X}}\|h^TAx\|_\infty:\ h\in\mathbf{R}^m,\ \|\mathrm{Col}_i[I_n-N^Th]\|_\infty\leqslant s^{-1}\kappa\right\} \tag{28}$$

which, when recalling what $\mathcal{X}$ is, by conic duality, are equivalent to the problems

$$h_k = \operatorname*{argmin}_{h,v,\chi}\left\{\overline{\varkappa}(\epsilon)\|h\|_2 + v + \phi_{\mathcal{T}}(\chi):\ h\in\mathbf{R}^m,\chi\geqslant 0,\right.$$
$$\left.\left[\begin{array}{c|c} v & h^TA \\ \hline A^Th & \sum_k\chi_kT_k\end{array}\right]\succeq 0,\ \|\mathrm{Col}_i[I_n-N^Th]\|_\infty\leqslant s^{-1}\kappa\right\},\quad 1\leqslant k\leqslant n.$$

E. The proposed construction of G is less straightforward. We proceed as follows. Let $G=[G_1,G_2]$ where $G_2,G_1\in\mathbf{R}^{m\times m}$ (so that $I=2m$).

E.1 Notice that as $\xi\in\Xi_\epsilon(G,H)$, problem (19) is feasible, and $(\widehat{x},\widehat{\nu})$ is its feasible solution. For a column g of G, by the constraints of the problem, we have

$$|g^TA[\widehat{x}-x_*]|\leqslant 2\overline{\varkappa}(\epsilon)\|g\|_2 + |g^TN[\widehat{\nu}-\nu_*]|\leqslant 2\overline{\varkappa}(\epsilon)\|g\|_2 + 2s\rho_H\|N^Tg\|_\infty, \tag{29}$$

(we have used (24) and (25)), implying that

$$\left(g^TA[\widehat{x}-x_*]\right)^2\leqslant 2\left(4\overline{\varkappa}^2(\epsilon)\|g\|_2^2 + 4s^2\rho_H^2\|N^Tg\|_\infty^2\right),\quad i=1,\dots,m. \tag{30}$$

Note that the set

$$\mathcal{M}=\left\{g\in\mathbf{R}^m:\ 8\overline{\varkappa}^2(\epsilon)\|g\|_2^2 + 8s^2\rho_H^2\|N^Tg\|_\infty^2\leqslant 1\right\}$$

is an ellitope: when denoting $N=[\mathrm{n}_1,\dots,\mathrm{n}_n]$ we have

$$\mathcal{M}=\left\{g\in\mathbf{R}^m:\ \exists r\in[0,1]^n:\ g^T\underbrace{\left(8\overline{\varkappa}^2(\epsilon)I_m + 8s^2\rho_H^2\mathrm{n}_j\mathrm{n}_j^T\right)}_{M_j}g\leqslant r_j,\ j=1,\dots,n\right\}.$$

E.2 Next, observe that when $\xi\in\Xi_\epsilon(G,H)$, by (21) one has

$$\|\widehat{\nu}-\nu_*\|_2\leqslant\frac{\sqrt{2s}}{1-2\kappa}\max_{k\leqslant n}\left[\overline{\varkappa}(\epsilon)\|h_k\|_2 + \max_{x\in\mathcal{X}}|h_k^TAx|\right] = \sqrt{2s}\rho_H.$$

Then by (29), for a column g of G it holds

$$\left(g^TA[\widehat{x}-x_*]\right)^2\leqslant\left(2\overline{\varkappa}(\epsilon)\|g\|_2 + |g^TN[\widehat{\nu}-\nu_*]|\right)^2\leqslant\left(2\overline{\varkappa}(\epsilon)\|g\|_2 + \sqrt{2s}\rho_H\|N^Tg\|_2\right)^2$$
$$\leqslant g^T\left(8\overline{\varkappa}^2(\epsilon)I_m + 4s\rho_H^2NN^T\right)g. \tag{31}$$

Now, let us put

$$Q = (8\overline{\varkappa}^2(\epsilon) I_m + 4s\rho_H^2 NN^T)^{-1/2}, \tag{32}$$

and consider the optimization problem

$$\mathrm{Opt} = \min_{\lambda,\mu,\Theta_1,\Theta_2,\rho} \left\{ f_H(\lambda,\mu,\Theta_1,\Theta_2,\rho) : \ \lambda \geqslant 0,\ \mu \geqslant 0,\ \Theta_1 \succeq 0,\ \Theta_2 \succeq 0, \right. \tag{33a}$$

$$\left. \mathrm{Tr}(M_j\Theta_1) \leqslant \rho,\ j = 1,\dots,n,\ \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T(\Theta_1 + Q\Theta_2 Q^T)A \end{array}\right] \succeq 0 \right\}$$

where

$$f_H(\lambda,\mu,\Theta_1,\Theta_2,\rho) = \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + \mathrm{Tr}(\Theta_2) + 2\sqrt{2}\ln(4m^2n)\rho. \tag{33b}$$

Note that the constraints on Θ_1 and ρ of the problem (33a) say exactly that (Θ_1,ρ) belongs to the cone $\mathbf{K}$ associated, as explained in Proposition 2, with the ellitope $\mathcal{M}$ in the role of $\mathcal{W}$.

Theorem 2. *Given a feasible solution $(\lambda,\mu,\tau,\Theta_1,\Theta_2)$ to (33), let us build $m\times m$ contrast matrices G_1, G_2 as follows.*

- *To build G_1, we apply the second part of Proposition 2 to $\Theta_1,\rho,\mathcal{M}$ in the roles of $\Theta,\rho,\mathcal{W}$, to get, in a computationally efficient way, a decomposition $\Theta_1 = \sum_{i=1}^m \gamma_i g_{1,i} g_{1,i}^T$ with $g_{1,i} \in \mathcal{M}$ and $\gamma_i \geqslant 0$, $\sum_i \gamma_i \leqslant 2\sqrt{2}\ln(4m^2n)\rho$. We set $G_1 = [g_{1,1},\dots,g_{1,m}]$.*
- *To build G_2, we subject Θ_2 to eigenvalue decomposition $\Theta_2 = \Gamma\mathrm{Diag}\{\chi\}\Gamma^T$ and set $G_2 = [g_{2,1},\dots,g_{2,m}] = Q\Gamma$.*

Note that $\Theta_1 + Q\Theta_2 Q = \sum_i \gamma_i g_{1,i} g_{1,i}^T + \sum_i \chi_i g_{2,i} g_{2,i}^T$.

For the resulting polyhedral estimate $\widehat{w}_{G,H}$ and for all $x_* \in \mathcal{X}$, s-sparse ν_* and $\xi \in \Xi_\epsilon(G,H)$ it holds

$$\|\widehat{w}_{G,H}(Ax_* + N\nu_* + \xi) - Bx_*\| \leqslant f_H(\lambda,\mu,\Theta_1,\Theta_2,\rho) \tag{34}$$

implying that the ϵ-risk of the estimate is upper-bounded by $f_H(\lambda,\mu,\Theta_1,\Theta_2,\rho)$ (due to $\xi \in \Xi_\epsilon(G,H)$ with probability $\geqslant 1-\epsilon$).

Proof. Let us fix $x_* \in \mathcal{X}$, s-sparse ν_*, $\xi \in \Xi_\epsilon(G,H)$, and let $w = Ax_* + N\nu_* + \xi$. By **A**, problem (17) is feasible, so that $(\widehat{x},\widehat{\nu}) = (\widehat{x}(\omega),\widehat{\nu}(\omega))$ is its optimal solution, and $\widehat{w} = B\widehat{x}$ is the estimate $\widehat{w}_{G,H}(\omega)$. Let $\Delta = \widehat{x} - x_*$, and let $e_1,\dots,e_m$ be the columns of the orthonormal matrix Γ. By construction of G_2, we have for all $j \leqslant m$ (see (31))

$$(g_{2,j}^T A\Delta)^2 \leqslant g_{2,j}^T \left(8\overline{\varkappa}^2(\epsilon) I_m + 4s\rho_H^2 NN^T\right) g_{2,j} = e_j^T \left[Q\left(8\overline{\varkappa}^2(\epsilon) I_m + 4s\rho_H^2 NN^T\right)Q\right] e_j = e_j^T e_j = 1.$$

Furthermore, due to $g_{1,i} \in \mathcal{M}$ one has (see (30))

$$\left(g_{1,i}^T A\Delta\right)^2 \leqslant 8\overline{\varkappa}^2(\epsilon)\|g\|_2^2 + 8s^2\rho_H^2\|N^T g\|_\infty^2 \leqslant 1 \quad \forall i \leqslant m.$$

Now, by the semidefinite constraint of (33a) and due to $\Theta_1 + Q\Theta_2 Q = \sum_i \gamma_i g_{1,i} g_{1,i}^T + \sum_i \chi_i g_{2,i} g_{2,i}^T$, for every $v \in \mathcal{B}_*$ we have

$$\begin{aligned} v^T B\Delta &\leqslant v^T \left[\sum_\ell \lambda_\ell S_\ell\right] v + \Delta^T \left[\sum_k \mu_k T_k\right] \Delta + [A\Delta]^T \left[\sum_i \gamma_i g_{1,i} g_{1,i}^T + \sum_i \chi_i g_{2,i} g_{2,i}^T\right] A\Delta \\ &\leqslant \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + \sum_i \chi_i (g_{1,i}^T A\Delta)^2 + \sum_j \gamma_j (g_{2,j}^T A\Delta)^2 \\ &\left[\text{as } [v^T S_1 v;\dots;v^T S_L v] \in \mathcal{S} \text{ due to } v \in \mathcal{B}_* \text{ and } [\Delta^T T_1\Delta;\dots;\Delta^T T_L\Delta] \in 4\mathcal{T} \text{ due to } \Delta \in 2\mathcal{X}\right] \\ &\leqslant \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + \sum_i \chi_i + \sum_j \gamma_j \leqslant f_H(\lambda,\mu,\tau,\Theta_1,\Theta_2) \end{aligned}$$

due to $\sum_i \gamma_i \leqslant 2\sqrt{2}\ln(4m^2n)\rho$ and $\sum_i \chi_i = \mathrm{Tr}(\Theta_2)$. Taking the supremum over $v \in \mathcal{B}_*$ in the resulting inequality, we arrive at (34). □

3.3. An Alternative

Our objective in this section is to refine risk bounds (27) and (33a) to produce more efficient contrasts. Our course of action is as follows. First, to extend the possible choice of H-contrasts" responsible" for the perturbation recovery, we refine the bounds (22) for the accuracy of sparse recovery, notably, to allow using contrasts not satisfying Condition $Q_\infty(s,\kappa)$. Second, we improve the bounding of the risk of the estimate $\widehat{w}(\omega)$ by taking into account the contribution of the H-component of the "complete" contrast matrix $[H,G]$ when optimizing the G-component of the contrast.

In the sequel, we consider the estimate described at the beginning of Section 3, the only difference being in the sizes of contrast matrices G and H: now $H = [h_1,\dots,h_M] \in \mathbf{R}^{m\times M}$, and $G = [g_1,\dots,g_{2m}]$. Thus, in our present setting, given observation ω, we solve the optimization problem

$$\min_{\nu,x}\left\{\|\nu\|_1 : x\in\mathcal{X},\ \begin{array}{ll}|h_k^T(N\nu+Ax-\omega)| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2, & k=1,\dots,M,\\ |g_i^T(N\nu+Ax-\omega)| \leqslant \overline{\varkappa}(\epsilon)\|g_{2,i}\|_2, & i=1,\dots,2m,\end{array}\right\} \tag{35}$$

with

$$\overline{\varkappa}(\epsilon) = \sigma\sqrt{2\ln[(2M+4m)/\epsilon]},$$

specify $\widehat{x}(\omega),\widehat{\nu}(\omega)$ as an optimal solution to the problem when the problem is feasible, otherwise set $(\widehat{x}(\omega),\widehat{\nu}(\omega)) = (0,0)$, and take $\widehat{w}_{G,H}(\omega) = B\widehat{x}(\omega)$ as the estimate of Bx_*.

3.3.1. Risk analysis. The above problem can be rewritten equivalently as

$$\min_{\nu,x}\left\{\|\nu\|_1 : x\in\mathcal{X},\ \begin{array}{ll}|h_k^T(N[\nu-\nu_*]+A[x-x_*]-\xi)| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2, & k=1,\dots,M,\\ |g_i^T(N[\nu-\nu_*]+A[x-x_*]-\xi)| \leqslant \overline{\varkappa}(\epsilon)\|g_i\|_2, & i=1,\dots,2m,\end{array}\right\} \tag{36}$$

and when setting

$$\Xi_\epsilon(G,H) := \left\{\xi\in\mathbf{R}^m :\ \begin{array}{ll}|h_k^T\xi| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2, & k=1,\dots,M,\\ |g_i^T\xi| \leqslant \overline{\varkappa}(\epsilon)\|g_i\|_2, & i=1,\dots,2m,\end{array}\right\} \tag{37}$$

we have

$$\mathrm{Prob}_{\xi\sim\mathcal{SG}(0,\sigma^2I_m)}(\Xi_\epsilon(G,H)) \geqslant 1-\epsilon.$$

Let us fix $\xi\in\Xi_\epsilon(G,H)$ and set $\omega = Ax_* + N\nu_* + \xi$. As $(\widehat{\nu},\widehat{x})$ is a feasible for (36), $\widehat{x} := \widehat{x}(\omega)$, $\widehat{\nu} := \widehat{\nu}(\omega)$ is feasible as well, $\|\widehat{\nu}\|_1 \leqslant \|\nu_*\|_1$. Thus, same as in the proof of Proposition 5, for $z = \widehat{\nu}-\nu_*$ it holds

$$\|z\|_1 \leqslant 2\|z\|_{s,1}$$

implying that

$$\|z\|_1 \leqslant 2s\|z\|_\infty, \quad \|z\|_2 \leqslant \sqrt{2s}\|z\|_\infty. \tag{38}$$

Now denote $\Delta = \widehat{x} - x_*$, and consider n pairs of convex optimization problems

$$\mathrm{Opt}_2[i] = \max_{v,t,w}\left\{\sqrt{w_i t} : v\in 2\mathcal{X},\ \begin{array}{l}\|w\|_\infty \leqslant w_i,\ \|w\|_1 \leqslant t,\ t\leqslant 2sw_i,\\ |h_k^T(Nw+Av)| \leqslant 2\overline{\varkappa}(\epsilon)\|h_k\|_2,\ k=1,\dots,M\end{array}\right\} \tag{$P_2[i]$}$$

$$\mathrm{Opt}_\infty[i] = \max_{v,w}\left\{w_i : v\in 2\mathcal{X},\ \begin{array}{l}\|w\|_\infty \leqslant w_i,\ \|w\|_1 \leqslant 2sw_i,\\ |h_k^T(Nw+Av)| \leqslant 2\overline{\varkappa}(\epsilon)\|h_k\|_2,\ k=1,\dots,M\end{array}\right\}. \tag{$P_\infty[i]$}$$

Observe that a feasible solution (v,t,w) to $(P_2[i])$ satisfies $\|w\|_\infty \leqslant w_i$ and $\|w\|_1 \leqslant t$, whence

$$\|w\|_2 \leqslant \sqrt{w_i t} \leqslant \mathrm{Opt}_2[i]. \tag{39}$$

Now, let $\iota = \iota_z$ be the index of the largest in magnitude entry in z. Taking into account that $\xi \in \Xi_\epsilon(G,H)$ and recalling that $\Delta \in 2\mathcal{X}$, we conclude that when $z_\iota \geqslant 0$, $(v,t,w) = (\Delta, \|z\|_1, z)$ is feasible for $(P_2[\iota])$ and $(v,w) = (\Delta, z)$ is feasible for $(P_\infty[\iota])$, while when $z_\iota < 0$ the same holds true for $(v,t,w) = (-\Delta, \|z\|_1, -z)$ and $(v,w) = (-\Delta, -z)$. Indeed in the first case $v = \Delta \in \mathcal{X}$, $|h_k^T[A\widehat{x} + N\widehat{\nu} - \omega]| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2$ and $|h_k^T[Ax_* + N\nu_* - \omega]| \leqslant \overline{\varkappa}(\epsilon)\|h_k\|_2$ as both pairs $(\widehat{x}, \widehat{\nu})$ and (x_*, ν_*) are feasible for (35), implying the second line constraints of $(P_2[i])$. Note that we are in the case of $z_\iota = \|z\|_\infty$, that is, constraints in the first line of $(P_2[i])$ are satisfied for $w = z$ due to (38). Thus, $(\Delta, \|z\|_1, z)$ indeed is feasible for $(P_2[i])$. As a byproduct of our reasoning, (Δ, z) is feasible for $(P_\infty[i])$. In the second case, the reasoning is completely similar.

Next, setting

$$\mathrm{Opt}_2 = \max_i \mathrm{Opt}_2[i], \quad \mathrm{Opt}_\infty = \max_i \mathrm{Opt}_\infty[i], \tag{40}$$

and recalling that $(\Delta, \|z\|_1, z)$ or $(-\Delta, \|z\|_1, -z)$ is feasible for some of the problems $(P_2[i])$, and (Δ, z) or $(-\Delta, -z)$ is feasible for some of the problems $(P_\infty[i])$, when invoking (39) we get for all $\xi \in \Xi_\epsilon(G,H)$

$$\|z\|_\infty \leqslant \mathrm{Opt}_\infty, \quad \|z\|_2 \leqslant \mathrm{Opt}_2, \quad \|z\|_1 \leqslant 2s\mathrm{Opt}_\infty.$$

Consequently, for all $d \in \mathbf{R}^m$

$$\begin{aligned} |d^T N z| &\leqslant \max_z \left\{ d^T N z : \|z\|_\infty \leqslant \mathrm{Opt}_\infty, \|z\|_2 \leqslant \mathrm{Opt}_2, \|z\|_1 \leqslant 2s\mathrm{Opt}_\infty \right\} \\ &= \underbrace{\min_{u,v,w} \left\{ \|u\|_1 \mathrm{Opt}_\infty + \|v\|_2 \mathrm{Opt}_2 + 2s\|w\|_\infty \mathrm{Opt}_\infty, \, u+v+w = N^T d \right\}}_{=:\overline{\pi}(N^T d)}. \end{aligned} \tag{41}$$

Now, recalling that $\widehat{x}, \widehat{\nu}$ is feasible for (36) and that $\xi \in \Xi_\epsilon(G,H)$, we conclude that columns d_i, $i = 1, \ldots, M+2m$ of the *"aggregated" contrast matrix* $D = [G, H]$ satisfy

$$|d_i^T A\Delta| \leqslant |d_i^T N z| + |d_i^T \xi| + \overline{\varkappa}(\epsilon)\|g\|_2,$$

whence

$$|d_i{}^T A\Delta| \leqslant \underbrace{\overline{\pi}(N^T d_i) + 2\overline{\varkappa}(\epsilon)\|d_i\|_2}_{=:\psi_H(d_i)}, \quad i \leqslant M+2m. \tag{42}$$

Next, let us put

$$\bar{f}_{G,H}(\lambda, \mu, \gamma) = \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + \sum_i \gamma_i \psi_H^2(d_i),$$

and consider optimization problem (cf. (27))

$$\begin{aligned} \mathrm{Opt}[G,H] = \min_{\lambda,\mu,\gamma} \Bigg\{ & \bar{f}_{G,H}(\lambda,\mu,\gamma) : \lambda \geqslant 0, \ \mu \geqslant 0, \ \gamma \geqslant 0, \\ & \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T \left[\sum_i \gamma_i d_i d_i{}^T \right] A \end{array} \right] \succeq 0. \Bigg\} \end{aligned} \tag{43}$$

Applying the same argument as in the proof of Proposition 1, with (42) in the role of (9), we arrive at the following result:

Proposition 4. *In the situation of this section, let* (λ, μ, γ) *be a feasible solution to* (43). *Then*

$$\mathrm{Risk}_\epsilon[\widehat{w}_{G,H} | \mathcal{X}, \mathcal{N}] \leqslant \bar{f}_{G,H}(\lambda, \mu, \gamma).$$

3.3.2. Contrast matrix synthesis. We continue our analysis of the estimate $\widehat{w}_{G,H}$ in the situation when the observation is $\omega = Ax_* + N\nu_* + \xi$ with $\xi \in \Xi_\epsilon(G,H)$, see (37). By (41), for $z = \widehat{\nu} - \nu_*$ and all $g \in \mathbf{R}^m$ we have

$$|g^T N z| \leqslant \min\left\{\|N^T g\|_2 \mathrm{Opt}_2,\ \sqrt{2s}\|N^T g\|_2 \mathrm{Opt}_\infty,\ 2s\|N^T g\|_\infty \mathrm{Opt}_\infty\right\}$$

what implies (cf. (42)) that for all $i \leqslant 2m$

$$|g_i^T A\Delta| \leqslant 2\overline{\varkappa}(\epsilon)\|g_i\|_2 + \min\left\{\|N^T g_i\|_2 \mathrm{Opt}_2,\ \sqrt{2s}\|N^T g_i\|_2 \mathrm{Opt}_\infty,\ 2s\|N^T g_i\|_\infty \mathrm{Opt}_\infty\right\}. \tag{44}$$

Note that the right-hand side in (44) is nonconvex in g, making our design techniques inapplicable. To circumvent this difficulty, we intend to utilize the following important feature of polyhedral estimates: one may easily "aggregate" several estimates of this type to yield an estimate with the risk which is nearly as small as the smallest of the risks of the estimates combined.

Here is how it works in the present setting. We split the $m \times 2m$ contrast G into two $m \times m$ blocks $G_\chi = [g_{\chi,1}, \ldots, g_{\chi,m}]$, $\chi = 1, 2$, and design the blocks utilizing the respective inequalities inherited from (44), specifically, the inequalities

$$\begin{aligned}
|g_{1,i}^T A\Delta| &\leqslant 2\overline{\varkappa}(\epsilon)\|g_{1,i}\|_2 + 2s\|N^T g_{1,i}\|_\infty \mathrm{Opt}_\infty,\\
|g_{2,i}^T A\Delta| &\leqslant 2\overline{\varkappa}(\epsilon)\|g_{2,i}\|_2 + \|N^T g_{2,i}\|_2 \underbrace{\min\{\mathrm{Opt}_2,\ \sqrt{2s}\mathrm{Opt}_\infty\}}_{=:\varrho_{2,H}}
\end{aligned}$$

where $\Delta = \widehat{x} - x_*$. We weaken these inequalities to

$$\begin{aligned}
|g_{1,i}^T A\Delta|^2 &\leqslant \pi_1^2(g_{1,i}), \quad \pi_1(g) = \sqrt{8\overline{\varkappa}^2(\epsilon)\|g\|_2^2 + 8s^2\mathrm{Opt}_\infty^2\|N^T g\|_\infty^2},\\
|g_{2,i}^T A\Delta|^2 &\leqslant \pi_2^2(g_{2,i}), \quad \pi_2(g) = \sqrt{8\overline{\varkappa}^2(\epsilon)\|g\|_2^2 + 2\varrho_{2,H}^2\|N^T g\|_2^2}.
\end{aligned}$$

Notice that norms π_χ, $\chi = 1, 2$, are ellitopic, so we can use in our present situation the techniques from Section 3.2, thus arriving at an analogue of Theorem 2. To this end, denote by $\mathrm{n}_1, \ldots, \mathrm{n}_n$ the columns of N and set

$$\overline{M}_j = 8\overline{\varkappa}^2(\epsilon) I_m + 8s^2\mathrm{Opt}_\infty^2 \mathrm{n}_j \mathrm{n}_j^T,\ j \leqslant m, \ \text{ and } \ \overline{Q} = \left(8\overline{\varkappa}^2(\epsilon) I_m + 2\varrho_{2,H}^2 NN^T\right)^{-1/2}.$$

Next, observe that the unit ball of $\pi_1(\cdot)$ is the ellitope

$$\overline{\mathcal{M}} = \left\{w \in \mathbf{R}^m : \exists r \in [0,1]^M : w^T \overline{M}_j w \leqslant \rho_j,\ j = 1, \ldots, M\right\}$$

and the unit ball of π_2 is the ellipsoid $w^T \overline{Q}^{-2} w \leqslant 1$. Now, let us consider the optimization problem

$$\begin{aligned}
\mathrm{Opt} = \min_{\lambda,\mu,\tau,\Theta_1,\Theta_2,\rho} \Bigg\{ & \overline{f}_H(\lambda,\mu,\tau,\Theta_1,\Theta_2,\rho) :\ \lambda \geqslant 0,\ \mu \geqslant 0,\ \tau \geqslant 0, && \text{(45a)}\\
& \Theta_1 \succeq 0,\ \Theta_2 \succeq 0,\ \mathrm{Tr}(\overline{M}_j \Theta_1) \leqslant \rho,\ j = 1, \ldots, n,\\
& \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & \sum_k \mu_k T_k + A^T\left[\sum_i \tau_i h_i h_i^T\right]A + A^T(\Theta_1 + Q\Theta_2 Q^T)A \end{array}\right] \succeq 0 \Bigg\}
\end{aligned}$$

where

$$\overline{f}_H(\lambda,\mu,\tau,\Theta_1,\Theta_2,\rho) = \phi_{\mathcal{S}}(\lambda) + 4\phi_{\mathcal{T}}(\mu) + \sum_i \tau_i \psi_H^2(h_i) + \mathrm{Tr}(\Theta_2) + 2\sqrt{2}\ln(4m^2 n)\rho. \tag{45b}$$

Note that the constraints on Θ_1 and ρ in this problem say exactly that (Θ_1, ρ) belongs to the cone $\mathbf{K}$ associated, according to Proposition 2, with the ellitope $\overline{\mathcal{M}}$ in the role of $\mathcal{W}$.

Theorem 3. *Given a feasible solution $(\lambda, \mu, \tau, \Theta_1, \Theta_2, \rho)$ to (45), let us build $m \times m$ contrast matrices G_1, G_2 as follows.*

- *To build G_1, we apply the second part of Proposition 2 to $(\Theta_1, \rho, \overline{\mathcal{M}})$ in the role of $(\Theta, \rho, \mathcal{W})$, to get, in a computationally efficient way, a decomposition $\Theta_1 = \sum_{i=1}^m \gamma_i g_{1,i} g_{1,i}^T$ with $g_{1,i} \in \overline{\mathcal{M}}$ and $\gamma_i \geqslant 0$, $\sum_i \gamma_i \leqslant 2\sqrt{2}\ln(4m^2n)\rho$. We set $G_1 = [g_{1,1}, \ldots, g_{1,m}]$.*
- *To build G_2, we subject Θ_2 to the eigenvalue decomposition $\Theta_2 = \Gamma \mathrm{Diag}\{\chi\}\Gamma^T$ and set $G_2 = [g_{2,1}, \ldots, g_{2,m}] = Q\Gamma$.*

Note that $\Theta_1 + Q\Theta_2 Q = \sum_i \gamma_i g_{1,i} g_{1,i}^T + \sum_i \chi_i g_{2,i} g_{2,i}^T$.

For the resulting polyhedral estimate $\widehat{w}_{G,H}$ and for all $x_* \in \mathcal{X}$, s-sparse ν_*, and $\xi \in \Xi_\epsilon(G, H)$ if holds

$$\|\widehat{w}_{G,H}(Ax_* + N\nu_* + \xi) - Bx_*\| \leqslant \overline{f}_H(\lambda, \mu, \tau, \Theta_1, \Theta_2, \rho) \tag{46}$$

implying that the ϵ-risk of the estimate is upper-bounded by $f_H(\lambda, \mu, \tau, \Theta_1, \Theta_2, \rho)$ (as $\xi \in \Xi_\epsilon(G, H)$ with probability $\geqslant 1 - \epsilon$).

Proof of the theorem follows that of Theorem 2 and is omitted.

3.4. Putting Things Together

Finally, we can "aggregate" polyhedral estimates from Sections 3.2 and 3.3 in the following construction (cf. [20, Section 5.1.6]):

Let us put

$$\overline{\varkappa}(\epsilon) = \sigma\left(2\ln[(2n + 2M + 8m)/\epsilon]\right)^{1/2},$$

and let $\widetilde{H} = [\widetilde{h}_1, \ldots, \widetilde{h}_n] \in \mathbf{R}^{m\times n}$, $\widetilde{G} = [\widetilde{g}_1, \ldots, \widetilde{g}_{2m}] \in \mathbf{R}^{m\times 2m}$, and $\overline{H} = [\overline{h}_1, \ldots, \overline{h}_M] \in \mathbf{R}^{m\times M}$, $\overline{G} = [\overline{g}_1, \ldots, \overline{g}_{2m}] \in \mathbf{R}^{m\times 2m}$ be the contrast matrices specified according to the synthesis recipes of Sections 3.2 and 3.3, respectively. We define the aggregated estimate $\widehat{w}$ of w_* as $\widehat{w}(\omega) = B\widehat{x}(\omega)$ where $\widehat{x}(\omega)$ is the x-component of

$$(\widehat{x}(\omega), \widehat{\nu}(\omega)) \in \operatorname*{Argmin}_{x,\nu} \left\{ \|\nu\|_1 : x \in \mathcal{X}, \begin{array}{ll} |\widetilde{h}_k^T[N\nu + Ax - \omega]| \leqslant \overline{\varkappa}(\epsilon)\|\widetilde{h}_k\|_2, & k = 1, \ldots, n, \\ |\overline{h}_k^T[N\nu + Ax - \omega]| \leqslant \overline{\varkappa}(\epsilon)\|\overline{h}_k\|_2, & k = 1, \ldots, M, \\ |\widetilde{g}_i^T[N\nu + Ax - \omega]|_\infty \leqslant \overline{\varkappa}(\epsilon)\|\widetilde{g}_i\|_2, & i = 1, \ldots, 2m, \\ |\overline{g}_i^T[N\nu + Ax - \omega]|_\infty \leqslant \overline{\varkappa}(\epsilon)\|\overline{g}_i\|_2, & i = 1, \ldots, 2m, \end{array} \right\}$$

when the problem is feasible, and $\widehat{x}(\omega) = 0$ otherwise.

Let us denote $G = [\widetilde{G}, \overline{G}] \in \mathbf{R}^{m\times 4m}$, let also $(\widetilde{\lambda}, \widetilde{\mu}, \widetilde{\gamma})$ be a feasible solution to the problem (27) with $H = \widetilde{H}$, and let $(\overline{\lambda}, \overline{\mu}, \overline{\gamma})$ be a feasible solution to the problem (43) with $H = \overline{H}$. Let $f_{G,H}$ and $\bar{f}_{G,H}$ be specified in (27) and (43) respectively. From Propositions 3 and 4 it immediately follows that for every s-sparse ν_* and every $x_* \in \mathcal{X}$ the error bound

$$\mathrm{Risk}_\epsilon[\widehat{w}(\cdot)|\mathcal{X}, \mathcal{N}] \leqslant \min\left[f_{G,\widetilde{H}}(\widetilde{\lambda}, \widetilde{\mu}, \widetilde{\gamma}), \bar{f}_{G,\overline{H}}(\overline{\lambda}, \overline{\mu}, \overline{\gamma})\right] \tag{47}$$

holds true.

Note that the resulting estimate can be efficiently optimized w.r.t. all parameters involved, except for $\overline{H}$, by specifying

- $\widetilde{H}$ as (near) minimizer of $\rho[H]$ over $H \in \mathcal{H}$ (23),
- $\widetilde{G}$ as a result of the decomposition of the (Θ_1, Θ_2)-component of a (near-) optimal solution to the problem (33a), (33b) (see Theorem 2) associated with $\widetilde{H}$,
- $\overline{G}$ as a result of the decomposition of the (Θ_1, Θ_2)-component of a (near-) optimal solution to the problem (45) (see Theorem 3) associated with $\overline{H}$.

3.5. Numerical Illustration

In our "proof of concept" experiment we compare three estimates of x_*: 1) estimate $\widehat{x}_{HG}$ with contrast matrix $[H, G]$ computed according to the recipe of Section 3.2, 2) estimate $\widehat{x}_{IG}$ with contrast $[\overline{H}, \overline{G}] = [I_m, \overline{G}]$ with $\overline{G}$ conceived utilizing the synthesis routine of Section 3.3.2, and 3) "aggregated" estimate $\widehat{x}_{HIG}$ with combined contrast $[H, I_m, G, \overline{G}]$. We solve adopted versions of optimization problems in (28) and (33a), (33b) to compute contrasts H and G of the estimate $\widehat{x}_{HG}$, and solve (45), to build the contrast $\overline{G}$ of $\widehat{x}_{IG}$. For instance, when computing the contrast $\overline{G}$, we set $\overline{\varkappa}(\epsilon) = \sqrt{2}\sigma\mathrm{erfcinv}\left(\frac{\epsilon}{2n}\right)$ where $\mathrm{erfcinv}(\cdot)$ is the inverse complementary Gaussian error function; when processing problem (45) numerically, Θ_1 was set to 0; the resulting problem can be rewritten as

$$\mathrm{Opt} = \min_{\lambda,\mu,\gamma,\Theta} \left\{ \lambda + 4\sum_{k=1}^{p}\mu_k + \sum_{j=1}^{n}\gamma_j + \mathrm{Tr}(\Theta) : \lambda \geqslant 0,\ \mu \geqslant 0,\ \gamma \geqslant 0,\ \Theta \succeq 0, \right.$$
$$\left. \left[\begin{array}{c|c} \lambda I_p & \frac{1}{2}I_p \\ \hline \frac{1}{2}I_p & \overline{A}^T\Theta\overline{A} + P^T\mathrm{Diag}\{\mu\}P + \overline{A}^T\mathrm{Diag}\{\gamma\}\overline{A} \end{array}\right] \succeq 0 \right\} \tag{48}$$

where $\overline{A} = A/\rho_2$ with $\rho_2 = 2\overline{\varkappa}(\epsilon) + \varrho_{2,\overline{H}}$, the subsequent entries in Pz being z_1, $[z_2 - z_1]/h$, $[z_{i-2} - 2z_{i-1} + z_i]/h^2$, $3 \leqslant i \leqslant p$, and $h = 2\pi/p$. The corresponding risk bounds are evaluated by computing solutions to (43). Optimization problems involved are processed using `Mosek` commercial solver [30] via `CVX` [15].[9]

In our illustration,

- $m = n = 256$, $q = p = 32$, $N = I_n$, $B = I_p$, A is a $n \times p$ random matrix with Gaussian entries such that $A^TA = I_p$;
- $\mathcal{X}$ is the restriction on the p-point equidistant grid on the segment $\Delta = [0, 2\pi]$ of functions f satisfying $|f(0)| \leqslant 4$, $|f'(0)| \leqslant 1$, $|f''(t)| \leqslant 4$, $t \in \Delta$;
- the norm $\|\cdot\|$ quantifying the recovery error is the standard Euclidean norm on $\mathbf{R}^p$;
- $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ with $\sigma = 0.1$, $\epsilon = 0.05$, and $s = 8$.

Figure 1 illustrates the results of the computation. In each experiment, we compute $n_S = 100$ recoveries $\widehat{x}_{HG}$, $\widehat{x}_{IG}$, and $\widehat{x}_{HIG}$ of randomly selected signals $x_* \in \mathcal{X}$ with generated at random

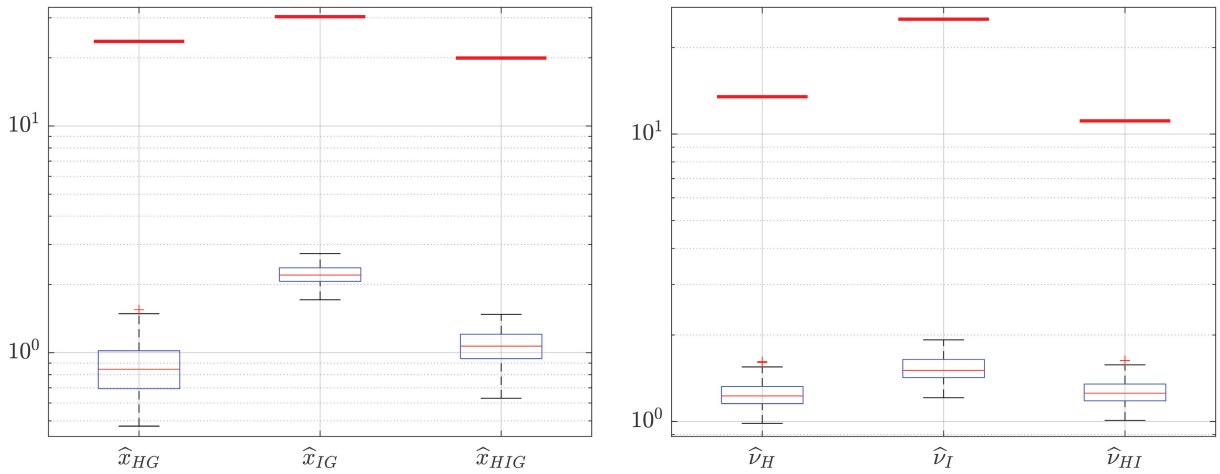

Fig. 1. Left plot: distributions of $\|\cdot\|_2$-errors of recovery of x_* and theoretical upper bounds on $\mathrm{Risk}_{0.05}$ (red horizontal bars); right plot: distributions of $\|\cdot\|_2$-errors and theoretical upper bounds on $\mathrm{Risk}_{0.05}$ of recovery of ν_*.

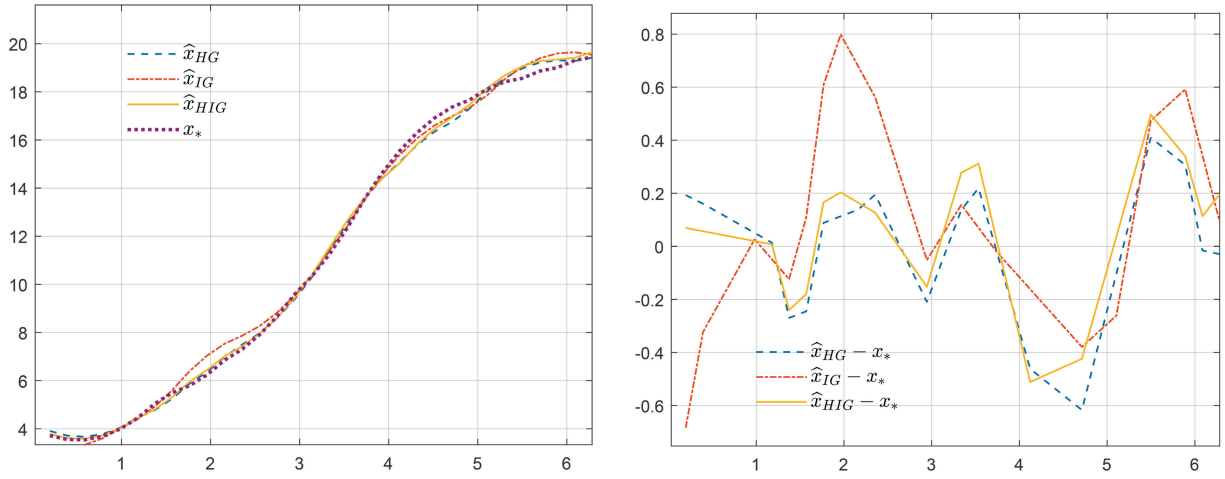

Fig. 2. A typical signal/estimates realization and recovery errors.

sparse nuisances ν_*. The results are presented in the left plot in Fig. 1. The right plot displays the boxplots of errors of recovery of the nuisance ν_* along with the upper risk bound Opt_2 of (40). Figure 2 illustrates a typical realization of the signal and the recovery errors; the values of $\|\cdot\|_2$-recovery errors are $\|\widehat{x}_{HG} - x_*\|_2 = 1.48\ldots$, $\|\widehat{x}_{IG} - x_*\|_2 = 2.02\ldots$, and $\|\widehat{x}_{HIG} - x_*\|_2 = 1.43\ldots$, the corresponding $\|x_*\|_2 = 72.2\ldots$.

APPENDIX. Error bound for ℓ_1 recovery

Condition $Q_\infty(s, \kappa)$

Given an $m \times n$ sensing matrix N, positive integer $s \leqslant n$, and $\kappa \in (0, 1/2)$, we say that $m \times p$ matrix H satisfy condition $\mathbf{Q}_\infty(s, \kappa)$ if

$$\|w\|_\infty \leqslant \|H^T N w\|_\infty + \frac{\kappa}{s}\|w\|_1 \quad \forall w \in \mathbf{R}^n. \tag{A.1}$$

For $y \in \mathbf{R}^n$, let y^s stand for the vector obtained from y by zeroing our all but the s largest in magnitude entries.

Proposition 5. *Given N and integer $s > 0$, assume that H satisfies the condition $\mathbf{Q}_\infty(s, \kappa)$ with $\kappa < \frac{1}{2}$. Then for all $\nu, \widehat{\nu} \in \mathbf{R}^n$ such that $\|\widehat{\nu}\|_1 \leqslant \|\nu\|_1$ it holds:*

$$\|\widehat{\nu} - \nu\|_q \leqslant \frac{(2s)^{\frac{1}{q}}}{1 - 2\kappa}\left[\|H^T N[\widehat{\nu} - \nu]\|_\infty + \frac{\|\nu - \nu^s\|_1}{s}\right], \ 1 \leqslant q \leqslant \infty. \tag{A.2}$$

Proof. Let us denote $\rho = \|H^T N[\widehat{\nu} - \nu]\|_\infty$, and let $z = \widehat{\nu} - \nu$.

1°. Let $I \subset \{1, \ldots, n\}$ of cardinality $\leqslant s$ and let $\overline{I}$ be its complement in $\{1, \ldots, n\}$. When denoting by x_I the vector obtained from a vector x by zeroing out the entries with indexes not belonging to I, we have

$$\|\widehat{\nu}_{\overline{I}}\|_1 = \|\widehat{\nu}\|_1 - \|\widehat{\nu}_I\|_1 \leqslant \|\nu\|_1 - \|\widehat{\nu}_I\|_1 = \|\nu_I\|_1 + \|\nu_{\overline{I}}\|_1 - \|\widehat{\nu}_I\|_1 \leqslant \|z_I\|_1 + \|\nu_{\overline{I}}\|_1,$$

and therefore

$$\|z_{\overline{I}}\|_1 \leqslant \|\widehat{\nu}_{\overline{I}}\|_1 + \|\nu_{\overline{I}}\|_1 \leqslant \|z_I\|_1 + 2\|\nu_{\overline{I}}\|_1.$$

[9] MATLAB code for this experiment is available at GitHub repository `https://github.com/ai1-fr/poly-robust`.

It follows that

$$\|z\|_1 = \|z_I\|_1 + \|z_{\overline{I}}\|_1 \leqslant 2\|z_I\|_1 + 2\|\nu_{\overline{I}}\|_1. \tag{A.3}$$

Besides this, by definition of ρ we have

$$\|H^T N z\|_\infty \leqslant \rho. \tag{A.4}$$

2°. Since H satisfies $\mathbf{Q}_\infty(s,\kappa)$, we have

$$\|z\|_{s,1} \leqslant s\|H^T N z\|_\infty + \kappa\|z\|_1$$

where $\|z\|_{s,1}$ is the ℓ_1-norm of the s-dimensional vector composed of the s largest in magnitude entries of z. By (A.4) it follows that $\|z\|_{s,1} \leqslant s\rho + \kappa\|z\|_1$ which combines with the evident inequality $\|z_I\| \leqslant \|z\|_{s,1}$ (recall that $\mathrm{Card}(I) = s$) and with (A.3) to imply that

$$\|z\|_1 \leqslant 2\|z_I\|_1 + 2\|\nu_{\overline{I}}\|_1 \leqslant 2s\rho + 2\kappa\|z\|_1 + 2\|\nu_{\overline{I}}\|_1,$$

hence (recall that $\kappa \leqslant \frac{1}{2}$)

$$\|z\|_1 \leqslant \frac{2s\rho + 2\|\nu_{\overline{I}}\|_1}{1-2\kappa}. \tag{A.5}$$

On the other hand, since H satisfies $\mathbf{Q}_\infty(s,\kappa)$, we also have

$$\|z\|_\infty \leqslant \|H^T N z\|_\infty + \frac{\kappa}{s}\|z\|_1,$$

which combines with (A.5) and (A.4) to imply that

$$\|z\|_\infty \leqslant \rho + \frac{\kappa}{s}\frac{2s\rho + 2\|\nu_{\overline{I}}\|_1}{1-2\kappa} = (1-2\kappa)^{-1}\left[\rho + \frac{\|\nu_{\overline{I}}\|_1}{s}\right]. \tag{A.6}$$

We conclude that for all $1 \leqslant q \leqslant \infty$,

$$\|z\|_p \leqslant \|z\|_\infty^{\frac{q-1}{q}}\|z\|_1^{\frac{1}{q}} \leqslant \frac{(2s)^{\frac{1}{q}}}{1-2\kappa}\left[\rho + \frac{\|\nu_{\overline{I}}\|_1}{s}\right]. \qquad \square$$

REFERENCES

1. Balakrishnan, S., Du S.S., Li, J., and Singh, A., Computationally efficient robust sparse estimation in high dimensions, in *Conference on Learning Theory*, pp. 169–212, PMLR, 2017.
2. Bickel, P.J., Ritov, Y., and Tsybakov F.B., Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics*, 2009, vol. 37(4), pp. 1705–1732.
3. Bruce, F.G., Donoho, D.L., Gao, H.-Y., and Martin, R.D., Denoising and robust nonlinear wavelet analysis, in *Wavelet Applications*, vol. 2242, pp. 325–336, SPIE, 1994.
4. Candes, E.J. and Tao, T., Decoding by linear programming, *IEEE transactions on information theory*, 2005, vol. 51(12), pp. 4203–4215.
5. Chen, Y., Caramanis, C., and Mannor, S., Robust sparse regression under adversarial corruption, in *International conference on machine learning*, pp. 774–782, PMLR, 2013.
6. Chernousko, F.L., *State estimation for dynamic systems*, CRC Press, 1993.
7. Dalalyan, A. and Thompson, P., Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's m-estimator, *Advances in neural information processing systems*, 2019, vol. 32.

8. Diakonikolas, I and Kane, D.M., *Algorithmic High-Dimensional Robust Statistics*, Cambridge University Press, 2023.

9. Diakonikolas, I., Kong, W., and Stewart, A., Efficient algorithms and lower bounds for robust linear regression, in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2745–2754, SIAM, 2019.

10. Donoho, D.L., Statistical estimation and optimal recovery, *The Annals of Statistics*, 1994, vol. 22(1), pp. 238–270.

11. Donoho, D.L. and Huo, X., Uncertainty principles and ideal atomic decomposition, *Information Theory, IEEE Transactions on*, 2001, vol. 47(7), pp. 2845–2862.

12. Fogel, E. and Huang, Y.-F., On the value of information in system identification-bounded noise case, *Automatica*, 1982, vol. 18(2), pp. 229–238.

13. Foygel, R. and Mackey, L., Corrupted sensing: Novel guarantees for separating structured signals, *IEEE Transactions on Information Theory*, 2014, vol. 60(2), pp. 1223–1247.

14. Granichin, O. and Polyak, P., *Randomized Algorithms of an Estimation and Optimization Under Almost Arbitrary Noises*, Nauka, 2003.

15. Grant, M. and Boyd, S., *The CVX Users' Guide. Release 2.1*, 2014. https://web.cvxr.com/cvx/doc/CVX.pdf

16. Huber, P., *Robust Statistics*, Wiley New York, 1981.

17. Juditsky, A. and Nemirovski, A., Near-optimality of linear recovery from indirect observations, *Mathematical Statistics and Learning*, 2018, vol. 1(2), pp. 171–225.

18. Juditsky, A. and Nemirovski, A., Near-optimality of linear recovery in Gaussian observation scheme under $\|\cdot\|_2^2$-loss, *The Annals of Statistics*, 2018, vol. 46(4), pp. 1603–1629.

19. Juditsky, A. and Nemirovski, A., On polyhedral estimation of signals via indirect observations, *Electronic Journal of Statistics*, 2020, vol. 14(1), pp. 458–502.

20. Juditsky, A. and Nemirovski, A., *Statistical Inference via Convex Optimization*, Princeton University Press, 2020.

21. Juditsky, A. and Nemirovski, A., On design of polyhedral estimates in linear inverse problems, *SIAM Journal on Mathematics of Data Science*, 2024, vol. 6(1), pp. 76–96.

22. Juditsky, A.B. and Nemirovski, A.S., Nonparametric estimation by convex programming, *The Annals of Statistics*, 2009, pp. 2278–2300.

23. Kurzhanski, A., *Identification–a theory of guaranteed estimates*, Springer, 1989.

24. Kurzhanski, A. and Vályi, I., *Ellipsoidal calculus for estimation and control*, Springer, 1997.

25. Liu, L., Shen, Y., Li, T., and Caramanis, C., High dimensional robust sparse regression, in *International Conference on Artificial Intelligence and Statistics*, pp. 411–421, PMLR, 2020.

26. Micchelli, C.A. and Rivlin, Y.J., *A Survey of Optimal Recovery*, pp. 1–54, Springer US, Boston, MA, 1977.

27. Micchelli, C.A. and Rivlin, T.J., Lectures on optimal recovery. In P.R. Turner, editor, *Numerical Analysis Lancaster 1984*, pp. 21–93, Springer Berlin Heidelberg, 1985.

28. Milanese, M. and Vicino, A., Optimal estimation theory for dynamic systems with set membership uncertainty: An overview, *Automatica*, 1991, vol. 27(6), pp. 997–1009.

29. Minsker, S., Ndaoud, M., and Wang, L., Robust and tuning-free sparse linear regression via square-root slope, *SIAM Journal on Mathematics of Data Science*, 2024, vol. 6(2), pp. 428–453.

30. Mosek, A., *The MOSEK optimization toolbox for MATLAB manual. Version 8.0*, 2015. http://docs.mosek.com/8.0/toolbox/

31. Polyak, B.T. and Tsypkin, Y.Z., Adaptive estimation algorithms: convergence, optimality, stability, *Avtomatika i telemekhanika*, 1979, no.3, pp. 71–84.

32. Polyak, B.T. and Tsypkin, Y.Z., Robust identification, *Automatica*, 1980, vol. 16(1), pp. 53–63.

33. Polyak, B.T. and Tsypkin, Y.Z., Robust pseudogradient adaptation algorithms, *Avtomatika i Telemekhanika*, 1980, no. 10, pp. 91–97.

34. Polyak, B.T. and Tsypkin, Y.Z., Optimal and robust methods for unconditional optimization, *IFAC Proceedings Volumes*, 1981, vol. 14(2), pp. 519–523.

35. Polyak, B.T. and Tsypkin, Y.Z., Criterial algorithms of stochastic optimization, *Avtomatika i Telemekhanika*, 1984, no. 6, pp. 95–104.

36. Polyak, B.T. and Tsypkin, Y.Z., Optimal recurrent algorithms for identification of nonstationary plants, *Computers & electrical engineering*, 1992, vol. 18(5), pp. 365–371.

37. Sardy, S., Tseng, P., and Bruce, A., Robust wavelet denoising, *IEEE transactions on signal processing*, 2001, vol. 49(6), pp 1146–1152.

38. Schweppe, F.C., *Uncertain dynamic systems*, Englewood Cliffs, NJ: Prentice-Hall, 1973.

39. Tukey, J.W., A survey of sampling from contaminated distributions, *Contributions to probability and statistics*, 1960, pp. 448–485.

40. Van de Geer, S., *Estimation and Testing under Sparsity*, Springer, 2016.

41. Yu, C. and Yao, W., Robust linear regression: A review and comparison, *Communications in Statistics-Simulation and Computation*, 2017, vol. 46(8), pp. 6261–6282.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 740–755.

SPECIAL ISSUE

Tight Approximations of Chance Constrained Sets Through Pack-Based Probabilistic Scaling

V. Mirasierra[*,a], **M. Mammarella**[**,b], **F. Dabbene**[**,c], **and T. Alamo**[*,d]

[*] *Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain*
[**] *Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), c/o Politecnico di Torino, Corso Duca Degli Abruzzi, 10129, Italy*
e-mail: [a] *vmirasierra@us.es,* [b] *martina.mammarella@cnr.it,* [c] *fabrizio.dabbene@cnr.it,* [d] *talamo@us.es*

Received March 3, 2025
Revised May 20, 2025
Accepted June 27, 2025

Abstract—The computation of probabilistic safe regions remains an evergreen problem in stochastic settings. Although the exact computation of safe regions may be possible for some specific problems, the results are generally overly complex (e.g., nonconvex, nonconnected) making them impractical for real-time applications. In this work, we present a sample-based procedure to obtain tight inner approximations of the safe region. The proposed approach does not require any assumption on the underlying probability distribution and the computation of the inner approximation set can be done offline. Unlike similar approaches, the proposed pack-based probabilistic scaling includes a tightening constraint, which tunes the level of conservativeness of the resulting approximation.

Keywords: randomized algorithms, probabilistic robustness, uncertain systems, statistical learning theory

DOI: 10.31857/S0005117925080049

1. INTRODUCTION

Real-world systems are often not deterministic and subject to uncertainty, necessitating the development of robust and stochastic control strategies. In robust control [1–3], the uncertainty is assumed to be unknown, but confined in a compact region, and the controller is designed to guarantee constraint satisfaction for all admissible values of the uncertainty. In contrast, stochastic control [4–6], incorporates probabilistic considerations introducing the concept of chance constraints [7]. Unlike hard constraints, chance constraints can be occasionally violated, provided that the probability of satisfaction remains above a specified threshold.

Relaxing the constraints and taking probabilities into account make stochastic schemes less conservative than their robust counterpart. Moreover, they make it possible to deal with infinite support uncertainties. In return, the resulting design process is much more intricate for two main reasons: First, it is highly difficult to check whether solutions of chance-constrained problems are feasible, and second, chance constraints usually involve nonconvexity (see, e.g., [8, Fig. 1; 9, Fig.1].

In the last decade, sampling-based schemes have emerged as a valid tool to deal with stochastic problems. Notably, Prof. Boris Polyak played a pivotal role in this field, being among the first scholars to recognize the potential of randomized methods in tackling optimization problems under stochastic uncertainty; for instance, see the works [10–12]. These works paved the way for subsequent results combining sampling and optimization, as the scenario approach proposed in [13]. For an overview of these techniques, the reader is referred to [14, 15].

The probabilistic safe region or *chance-constrained set* (CCS) is defined as the region that contains all the points satisfying the chance constraints. In a general setting, the exact computation of the CCS is cumbersome and requires the uncertainty to follow a certain distribution [16, 17]. Besides, the complexity of its geometry can make it ill-suited for real-time applications [18]. Because of these limitations, it is pertinent to address the problem of approximating the safe regions using sets of manageable complexity.

For the stochastic control problem, several relaxations have been proposed, which rely on computationally efficient approximations of the chance-constrained set. These relaxations can be either based on some concentration inequalities, e.g. exploiting previous knowledge about the structure of the uncertainty [19], or they can be constructed using random sampling methods [20, 21].

The present work stems from the results in [9, 22], where a sample-based methodology to inner approximate the CCS named *probabilistic scaling* is presented. This approach computes first a simple approximating set, which is then scaled to meet the required probabilistic guarantees. These operations are all performed *offline* and the trade-off between the number of samples required and the tightening of the approximation can be adjusted by the user.

In this paper, we discuss and extend the pack-based probabilistic scaling approach presented in the preliminary conference publication [22], by defining a novel measure of the tightening of the approximating set. Then, we show how to design the approximating set to meet the required probabilistic guarantees while incorporating the specified tightening constraint. In this way, the user is given the capability to control at the same time the complexity and the fitting of the resulting approximating set, balancing the trade-off between the required number of samples and the computational complexity of the approximation problem (which is computed offline).

The paper is structured as follows. In Section 2 we introduce the problem of approximating the chance constrained set and the numerical example used to compare the different approaches. In Section 3 we go through statistical learning theory solutions to the problem, first introducing the classical probabilistic scaling approach (Section 3.1) and later describing the extension to the pack-based framework (Section 3.2). Then, Section 4 is dedicated to the tight immersed pack-based probabilistic scaling, which is the main contribution of this work. Last, Section 5 includes the comparative analysis of the different approaches in terms of conservativeness.

Notation: $\mathbb{N}_{\geqslant 0}$ is the set of natural numbers including 0. The notation $\oplus$ refers to the Minkowski sum of sets. Given a set of N scalars $\{x_1, x_2, \dots, x_N\}$, we denote $x_{1:N}$ the smallest one, $x_{2:N}$ the second smallest one, and so on and so forth until $x_{N:N}$, which is the largest. By the definition of $x_{1+r:N}$, for a given $r \geqslant 0$, no more than r elements of $\{x_1, \dots, x_N\}$ are strictly smaller than $x_{1+r:N}$. We refer to the binomial distribution as

$$B(s; N, \varepsilon) = \sum_{i=0}^{s} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i}.$$

2. APPROXIMATING CHANCE-CONSTRAINED SETS

Let us consider a robustness problem, where the controller parameters and the auxiliary variables are parameterized by means of a decision variable vector $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$, which is denoted as *design parameter*. The uncertainty vector w represents one of the admissible uncertainty realizations of a random vector with given probability distribution $\Pr_{\mathcal{W}}$ with (possibly unbounded) support $\mathcal{W}$. Then, the generic uncertain constraint can be defined as

$$g(\theta, w) \leqslant 0, \tag{1}$$

where the function $g : \mathbb{R}^{\Theta \times \mathcal{W}} \to \mathbb{R}$ captures the requirement for θ given w. In particular, in a robust setting, one requires that the constraint (1) holds for all possible values of w. Clearly, there might

be situations where dealing with this kind of constraint in a fully robust manner is senseless, e.g., when the support of w is unbounded [23]. In that case, one may accept that the constraint (1) is violated by a fraction of the elements of $\mathcal{W}$. This concept is rigorously formalized in the definition of chance constraints.

Definition 1 [set of probability ε-CCS [9]]. Consider a probability measure $\Pr_{\mathcal{W}}$ over $\mathcal{W}$. Given the violation level $\varepsilon \in (0,1)$, we define the chance-constrained set of probability ε (ε-CCS) as follows

$$\Omega(\varepsilon) = \{\theta \in \Theta \mid \Pr_{\mathcal{W}}\{g(\theta, w) > 0\} \leqslant \varepsilon\}.$$

Recently, several approaches have been proposed to construct a probabilistically guaranteed approximation of the chance-constrained set. These approaches are based on sample-based results (see e.g., [21, 24, 25]). Given $\mathcal{W}$, consider a collection of N independent identically distributed (i.i.d.) samples $\mathbf{z} = \{w_1, \ldots, w_N\}$ drawn from $\mathcal{W}$. In this case, we say that $\mathbf{z}$ belongs to the Cartesian product $\mathcal{W}^N \doteq \mathcal{W} \times \cdots \times \mathcal{W}$ (N times) and, correspondingly, we say that $\mathbf{z}$ is drawn according to the product probability measure $\Pr_{\mathcal{W}^N}$. Let us introduce the concept of an indicator function, later used to redefine the chance-constrained set.

Definition 2 [indicator function of g]. Given $\theta \in \Theta$ and $w \in \mathcal{W}$, then the indicator function $I^g : \Theta \times \mathcal{W} \to \{0,1\}$ of constraint (1) is defined as

$$I^g(\theta, w) \doteq \begin{cases} 0 & \text{if } g(\theta, w) \leq 0 \\ 1 & \text{otherwise.} \end{cases}$$

In the context of statistical learning theory, we can compute approximations of the ε-CCS by means of a constraint on the *empirical mean* defined as

$$\frac{1}{N}\sum_{i=1}^{N} I^g(\theta, w_i).$$

That is, given $\mathbf{z} = \{w_1, \ldots, w_N\} \in \mathcal{W}^N$ and a discarding parameter $r \geqslant 0$, then the parameter $\rho = \frac{r}{N}$ bounds the empirical mean so that the set

$$\Phi_{\rho_N}(\mathbf{z}) \doteq \left\{\theta \in \Theta : \frac{1}{N}\sum_{i=1}^{N} I^g(\theta, w_i) \leqslant \rho\right\} \tag{2}$$

constitutes an approximation of $\Omega(\varepsilon)$. Note that the expression $\frac{1}{N}\sum_{i=1}^{N} I^g(\theta, w_i) \leqslant \frac{r}{N}$ means that the constraint $g(\theta, w_i) \leq 0$ is violated by no more than r elements of $\mathbf{z}$.

Remark 1. We note that, given ε, $\Omega(\varepsilon)$ is a fixed set. On the other hand, when the ε-CCS is approximated by means of sampling techniques (see e.g., [9, 26]), then the corresponding approximated set has a random nature, being generated from the random samples $\mathbf{z} \in \mathcal{W}^{\mathbf{N}}$.

Assuming that the indicator function I^g has finite Vapnik–Chervonenkis (VC) dimension [27] and that $\rho < \varepsilon$, then the probability of $\Phi_{\rho_N}(\mathbf{z})$ being an inner approximation of $\Omega(\varepsilon)$, i.e.,

$$\Pr_{\mathcal{W}^N}\{\Phi_{\rho_N}(\mathbf{z}) \subseteq \Omega(\varepsilon)\}$$

converges to 1 as the number of samples N converges to infinity. In [28], the sample complexity bounds for N are explicitly computed, which guarantee that $\Phi_{\rho_N}(\mathbf{z})$ is included in $\Omega(\varepsilon)$ with a given confidence $\delta \in (0,1)$, i.e., $\Pr_{\mathcal{W}^N}\{\Phi_{\rho_N}(\mathbf{z}) \subseteq \Omega(\varepsilon)\} \geqslant 1 - \delta$.

The resulting sample complexity grows linearly with the VC dimension of I^g multiplied by a factor larger than $\frac{1}{\varepsilon}$. However, as shown in [9], this approximation may be very conservative. Also, when the function g is not convex, the resulting approximation is generally non-convex and is often non-connected. This may hinder its practical application and makes it generally unsuitable for real-time problems.

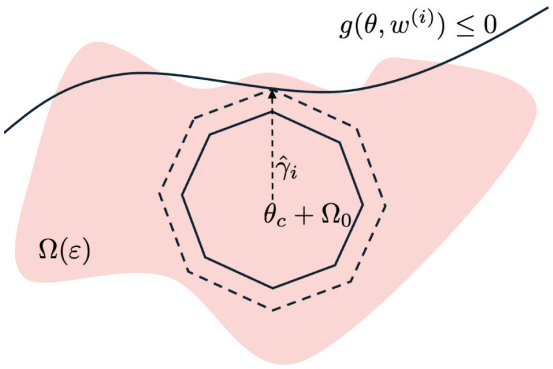

Fig. 1. Scheme of the probabilistic scaling approach.

Building on these notions, [9, 22] introduced the probabilistic scaling idea. At the basis of this approach is the introduction of an initial *simple approximating set* (SAS) $\theta_c \oplus \Omega_0$, which has to possess two main characteristics: i) be able to capture sufficiently well the "shape" of the probabilistic set $\Omega(\varepsilon)$, while at the same time being ii) sufficiently simple. This initial SAS does not need to offer any guarantee of probabilistic nature, but it should be able to capture the shape of the ε-CCS.

In [9] it was shown how to *scale* this set around its center θ_c to obtain a scalable SAS

$$\Omega(\gamma) = \theta_c \oplus \gamma\Omega_0,$$

and a sample-based procedure was introduced to construct a probabilistically meaningful approximation of the ε-CCS. Specifically, given a shape Ω_0 and a scaling center θ_c, the goal of probabilistic scaling is to find the largest scaling factor $\bar{\gamma}$ such that

$$\Pr_{\mathcal{W}}\{\theta_c \oplus \bar{\gamma}\Omega_0 \subseteq \Omega(\varepsilon)\} \geqslant 1 - \delta, \tag{3}$$

and therefore, also the chance constraint

$$\Pr_{\mathcal{W}}\{g(\theta, w) \leqslant 0\} \geq 1 - \varepsilon \tag{4}$$

is satisfied with a probability not lower than $1 - \delta$.

The procedure for constructing such an approximation is discussed in detail in [9], and recalled formally in Section 3.1.

In Fig. 1, we give a simple illustration of the approach, where we assume that the red area represents the ε-CCS, which as observed can be in general nonconvex. Then:

(1) Select "candidate" approximating set $\theta_c + \Omega_0$ (black polygon);
(2) To design the optimal scaling $\bar{\gamma}$, extract N samples $\mathbf{z} = \{w_1, \ldots, w_N\} \in \mathcal{W}^N$;
(3) For each random sample w_i, compute the maximum scaling γ_i so that the scaled set (dashed polygon) does not violate the constraint corresponding to w_i;
(4) Select the optimal scaling as $\bar{\gamma} = \gamma_{1+r:N}$, i.e., as the r smallest value of γ_i.

Then, (3) holds for $B(r; N, \varepsilon) \leqslant \delta$.

Despite the undeniable benefits of exploiting *probabilistic scaling*, especially in the extended version where the computational complexity is further reduced by employing the so-called simple-approximating sets (SAS) [9], the scaling solution may result to be conservative. This issue is illustrated by means of the following example from [22].

Example 1. We consider a problem involving individual chance constraints, where every constraint is tangent to the unit circle of a given dimension at a random point, drawn from a uniform distribution. In this case, clearly, the unit ball is the safe region with probability 1, whereas the

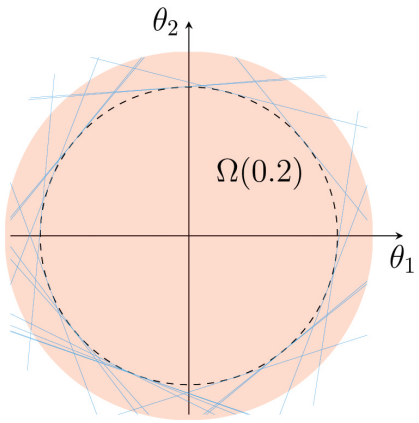

Fig. 2. The red circle represents $\Omega(0.2)$, the dashed black circle is the SAS (unit circle), and the cyan lines are the sampled constraints.

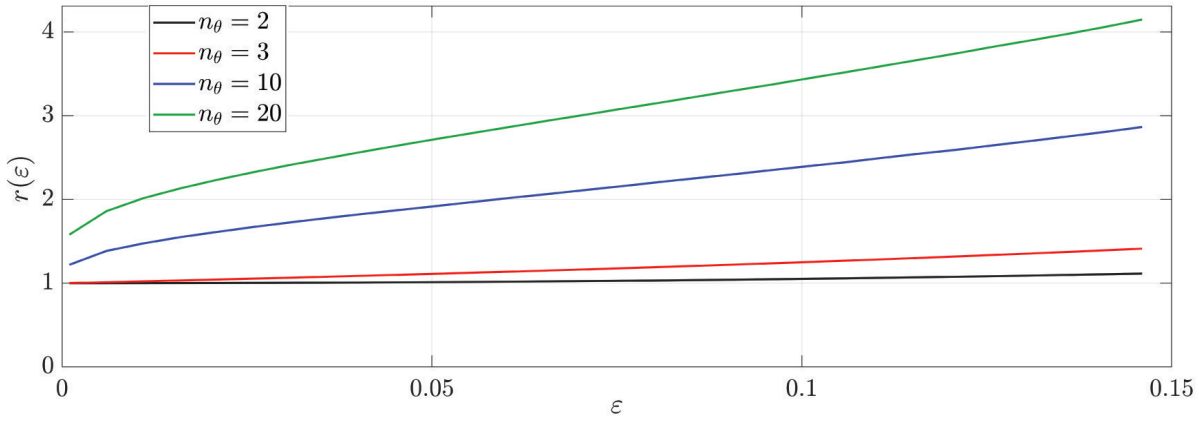

Fig. 3. Radius r of $\Omega(\varepsilon)$ as a function of ε for different problem dimensions n_θ.

ε-CCS is always a slightly larger scaled version of the unit ball as ε increases. In particular, it can be easily shown that the exact radius corresponding to the chance constrained region $\Omega(\varepsilon)$ can be computed using some transcendental functions. Figure 2 illustrates this example in $\mathbb{R}^2$: where the dashed line is the unit circle in $\mathbb{R}^2$ and the outer red circle represents the chance constrained set $\Omega(\varepsilon)$ for the specific value $\varepsilon = 0.20$.

Assume that we want to approximate the ε-CCS using the empirical mean approximation $\Phi_{\rho_N}(\mathbf{z})$ introduced in (2). To this end, we generate N random linear constraints tangent to points drawn from a uniform probability distribution on the surface of the unit hypersphere and construct the approximation as the intersection of them (possibly discarding the "worst" ones). It is clear that such an approximation will fail to capture the red circle.

Additionally, assume we want to use a probabilistic scaling approach, and we choose the unit ball as the initial approximation $\theta_c \oplus \Omega_0$ of the chance-constrained set $\Omega(\varepsilon)$. Then, applying the previously described procedure, it would be possible to scale this initial geometry around its center (the origin) to obtain an inner approximation of $\Omega(\varepsilon)$ with a given confidence level $\delta \in (0, 1)$. However, it is evident that the scaling scheme will always yield the unit hypersphere as a final result, as each sampled constraint is tangent to it, implying that all computed scaling factors will be equal to one. Hence, simple sampling-based procedures will fail to capture the radius of the true set $\Omega(\varepsilon)$. Note that this radius may be significantly larger than one, especially when the n_θ increases, as shown in Fig. 3.

On the other hand, for the given example, one may notice that larger scale factors can be obtained if one scales the unit-circle taking into consideration only the regions in which more than a given number of constraints are violated. In this paper, we resort to the *pack-based strategy*, successfully employed in the context of statistical learning theory [28] and convex scenario [29], to obtain less conservative sample complexities and to guarantee that the obtained scaled set is included into the chance constrained set with a given confidence level. Specifically, the goal is to extend the pack-based strategy first proposed in [22] to obtain sample-based approximations of $\Omega(\varepsilon)$ with tunable complexity, which do not require any previous knowledge of the problem, e.g., symmetry. The ability to reduce the conservativeness of the proposed approach will be later demonstrated against the illustrative Example 1.

3. PRELIMINARY NOTIONS

In this section, we first recall some notions from the pack-based strategy, which are propaedeutical to the main results of this paper. Then, in the next section we present the pack-based probabilistic scaling (PBPS) approach discussed in [22], which will be later extended to further reduce the conservativeness of the approximating set. First, we introduce the definition of *pack of samples*.

Definition 3 [pack of L samples]. Given an integer L, a collection of L samples $\mathbf{z} = \{w_1, \dots, w_L\} \in \mathcal{W}^L$ is said to be a pack of dimension L.

Then, we extend the definition of indicator function in Definition 2 to the pack-based framework.

Definition 4 [pack indicator function]. Given integers s and L such that $0 \leqslant s < L$ and a pack $\mathbf{z} \in \mathcal{W}^L$ of dimension L, the *pack indicator function* $I_s^g : \Theta \times \mathcal{W}^L \to \{0, 1\}$ is defined as

$$I_s^g(\theta, \mathbf{z}) \doteq \begin{cases} 0 & \text{if } \sum_{\ell=1}^{L} I^g(\theta, w_\ell) \leqslant s \\ 1 & \text{otherwise,} \end{cases} \tag{5}$$

where $I_s^g(\theta, \mathbf{z})$ indicates whether the point θ violates more than s of the constraints associated with the uncertainty realizations of the pack $\mathbf{z}$.

Definition 5 [pack safe region]. The pack safe region $\Phi_s(\mathbf{z})$ is defined as the set of points which violate no more than s of the constraints associated with the uncertainty realizations of $\mathbf{z}$, and can be expressed as

$$\Phi_s(\mathbf{z}) \doteq \{\theta \in \Theta \,|\, I_s^g(\theta, \mathbf{z}) = 0\}.$$

In the next section, we present a generalization of the results on probabilistic scaling applied in the framework of pack-based strategy. In detail, we show how to scale the set $\theta_c \oplus \Omega_0$ around its center θ_c to guarantee with confidence level $\delta \in (0, 1)$, the inclusion of the scaled, pack-based set into $\Omega(\varepsilon)$.

3.1. Generalized Probabilistic Scaling

First, we introduce the definition of scaling factor in the pack-based framework.

Definition 6 [pack scaling factor]. Given a scalable SAS $\Omega(\gamma)$ defined by a scaling center $\theta_c \in \Theta$ and a shape Ω_0, and a pack $\mathbf{z} \in \mathcal{W}^L$, we define the pack scaling factor of $\Omega(\gamma)$ relative to the random constraints $g(\theta, w_i) \leqslant 0, \forall w_i \in \mathbf{z}$ as

$$\gamma^s(\theta_c, \Omega_0, \mathbf{z}) \doteq \begin{cases} 0 & \text{if } \theta_c \notin \Phi_s(\mathbf{z}) \\ \max\limits_{\theta_c \oplus \gamma\Omega_0 \subseteq \Phi_s(\mathbf{z})} \gamma & \text{otherwise.} \end{cases} \tag{6}$$

Now, we formalize the *generalized probabilistic scaling problem*, considering M i.i.d. packs $\mathbf{z}_i$, each one of dimension L. Note that the problem generalizes the probabilistic scaling introduced in [9], which can be obtained by letting $M = N$ and $L = 1$ (i.e., considering N packs of dimension 1).

Property 1 [generalized probabilistic scaling]. Given the accuracy parameter $\varepsilon \in (0, 1)$ and the confidence level $\delta \in (0, 1)$, consider the discarding integer parameter $r \geqslant 0$ and suppose that M is chosen such that

$$B(r; M, \varepsilon) \leqslant \delta. \tag{7}$$

Draw M i.i.d. L-dimensional packs, $\mathbf{z}_i \in \mathcal{W}^L$, $i = 1, \dots M$. For each pack $\mathbf{z}$, compute the corresponding pack scaling factor γ_i as

$$\gamma_i \doteq \gamma^s(\theta_c, \Omega_0, \mathbf{z})$$

according to (6) and define $\bar{\gamma} \doteq \gamma_{1+r:M} > 0$. Then, with probability no smaller than $1 - \delta$,

$$\Pr_{\mathcal{W}^L}\{\theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Phi_s(\mathbf{z})\} \leqslant \varepsilon.$$

Proof. This property can be demonstrated by particularizing the results of convex scenario [21, 26] to the case of a scalar decision variable. Another possibility is to derive the results using the properties of the generalized max function [30, Property 3]. Consider the following optimization problem:

$$\begin{aligned} &\max_{\gamma} \quad \gamma \\ &s.t. \quad \theta_c \oplus \gamma\Omega_0 \subseteq \Phi_s(\mathbf{z}), \quad i = 1, \dots, M. \end{aligned} \tag{8}$$

If this problem has a feasible solution, then we can rewrite it using the definition of $\gamma^s(\cdot)$ as

$$\begin{aligned} &\max_{\gamma} \quad \gamma \\ &s.t. \quad \gamma \leqslant \gamma^s(\theta_c, \Omega_0, \mathbf{z}), \quad i = 1, \dots, M. \end{aligned} \tag{9}$$

It has been proved in [21, 26] that if one discards no more than r constraints on a convex problem with M random constraints, then the probability of violating the constraints with the solution obtained from the random convex problem is no larger than ε, with probability no smaller than $1 - \delta$, where

$$\delta = \begin{pmatrix} d + r - 1 \\ d - 1 \end{pmatrix} B(d + r - 1; M, \varepsilon),$$

and d is the number of decision variables. We first notice that (9) is convex and has a unique scalar decision variable γ, i.e., $d = 1$. Also, the assumptions required in the application of the results of [21, 26] can be easily checked. In particular, non-degeneracy is implied by the fact that the problem is scalar, while uniqueness can be enforced by introducing a tie-break rule. Hence, if we allow r violations in the above minimization problem, then with probability no smaller than $1 - \delta$, with $\delta = B(r; M, \varepsilon)$, the optimal solution $\bar{\gamma}$ of problem (9) with no more than r constraint removed satisfies

$$\Pr_{\mathcal{W}^L}\{\bar{\gamma} > \gamma^s(\theta_c, \Omega_0, \mathbf{z})\} \leqslant \varepsilon.$$

Hence, we can conclude that with probability no smaller than $1 - \delta$

$$\Pr_{\mathcal{W}^L}\{\theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Phi_s(\mathbf{z})\} \leqslant \varepsilon.$$

Note that problem (9) with constraint removal can be solved directly by ordering the values $\gamma_i = \gamma^s(\theta_c, \Omega_0, \mathbf{z}_i)$. It is clear that if $r \geqslant 0$ violations are allowed, then the optimal value for γ is $\bar{\gamma} = \gamma_{1+r:N}$. Smaller values of γ would meet the inclusion of constraint but will not be optimal, while larger values of γ would no longer meet the inclusion constraint □.

As discussed before, the result in [22, Property 1] can be particularized from Property 2 by setting $M = N$ and $L = 1$. This is summarized in the next corollary.

Corollary 1 [classic probabilistic scaling]. *Suppose that N is chosen such that*

$$B(r; N, \varepsilon) \leqslant \delta.$$

Let $\mathbf{z} \in \mathcal{W}^N$. For each constraint $i = 1, \ldots, N$, define $\gamma_i = \gamma^r(\theta_c, \Omega_0, \mathbf{z}_i)$. Suppose that $\bar{\gamma} = \gamma_{1+r:N} > 0$. Then, with probability no smaller than $1 - \delta$,

$$\Pr_{\mathcal{W}}\{\theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Omega(\varepsilon)\} \leqslant \varepsilon.$$

The proof is straightforward and follows directly from Definition 1. The above corollary shows that the probabilistic scaling approach in [9] can be viewed as a special case of a more general pack-based scheme.

Calculating approximations of the ε-CCS using classical probabilistic scaling is generally easy to compute, does not require any assumption on the underlying probabilities (such as finite VC dimension), provides probabilistic guarantees to the scaled region, and its effectiveness has been proven [9]. Despite all its advantages, classical probabilistic scaling may still lead to very conservative solutions, as shown in Example 1. In that case, having that $\gamma_i = 1$, for all $i = 1, \ldots, N$ and all the constraints are taken into account independently, the act of discarding some of them has no effect on the resulting scaled approximating set.

In the next section, we outline the so-called pack-based probabilistic scaling, first proposed in [22]. For the same initial SAS, this variant of the classical probabilistic scaling applied in the framework of pack-based strategy may lead to less conservative results at the expense of (possibly) more demanding computational cost.

It is important to highlight that the discarding parameter r is set by the user and it should be selected taking into account that large values of r make the resulting set less sensitive to extreme values, at the expense of a larger sample complexity N. On the other hand, the convexity of the approximating scaled set is independent of the discarding parameter r and only depends on the choice of the SAS geometry.

Remark 2. Property 1 can also be particularized for the case $r = 0$. Suppose that M is such that $(1 - \varepsilon)^M \leqslant \delta$. Draw M i.i.d. L-dimensional packs $\mathbf{z}_i \in \mathcal{W}^L$ and define $\gamma_i \doteq \gamma^s(\theta_c, \Omega_0, \mathbf{z}_i)$. Suppose that $\bar{\gamma} = \gamma_{1:M} > 0$. Then, with probability no smaller than $1 - \delta$, $\Pr_{\mathcal{W}^L}\{\theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Phi_s(\mathbf{z})\} \leqslant \varepsilon$.

3.2. Pack-Based Probabilistic Scaling

The main underlying idea of pack-based probabilistic scaling is to divide the uncertainty samples into packs and to allow some constraint violations inside each pack. As opposed to regular probabilistic scaling, where the scaling factor associated with each constraint is computed *independently*, in the pack-based approach the constraints inside each pack are taken into account *together*. Ultimately, this can lead to tighter approximations of the ε-CCS and reduced sample complexity.

Let the N sampled constraints be divided into M packs of L constraints each, i.e., $\mathbf{z} = \{\mathbf{z}_1, \ldots \mathbf{z}_M\} = \{w_1, \ldots, w_N\}$, with $\mathbf{z} \in \mathcal{W}^N$ and $\mathbf{z}_i \in \mathcal{W}^L$ for $i = 1, \ldots, M$. The following theorem shows how to determine the scaling factor using a pack-based approach so that the scaled SAS is fully contained in the ε-CCS with given confidence δ.

Theorem 1 [pack-based probabilistic scaling]. *Consider a shape Ω_0, a scaling center θ_c, accuracy parameter $\varepsilon \in (0, 1)$, confidence level $\delta \in (0, 1)$, and nonnegative integers M, L, s with $L > s$ so that*

$$B(s; L, \varepsilon)^M \leqslant \delta. \tag{10}$$

For each pack of constraints $i = 1, \ldots, M$, let $\mathbf{z}_i \in \mathcal{W}^L$ and define $\gamma_i \doteq \gamma^s(\theta_c, \Omega_0, \mathbf{z}^{(i)})$ as in (6). Suppose that $\bar{\gamma} = \gamma_{1:M} > 0$. Then, with probability no smaller than $1 - \delta$,

$$\theta_c \oplus \bar{\gamma}\Omega_0 \subseteq \Omega(\varepsilon).$$

Proof. Let $p = 1 - B(s; L, \varepsilon)$. From Remark 2, we know that if M is chosen such that $B(s; L, \varepsilon)^M \leqslant \delta$ and $\bar{\gamma} = \gamma_{1:M} > 0$, then with probability no smaller than $1 - \delta$ we have

$$\Pr_{\mathcal{W}^L}\{\theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Phi_s(\mathbf{z})\} \leqslant p.$$

Equivalently, $\Pr_{\mathcal{W}^L}\{I_s^g(\theta, \mathbf{z}) = 1, \ \ \forall \theta \in \theta_c \oplus \bar{\gamma}\Omega_0\} \leqslant p$. Moreover, from Property 3 in Appendix A we have that

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta, \mathbf{z}) = 1\} \leqslant p \iff \Pr_{\mathcal{W}}\{I^g(\theta, w) = 1\} \leqslant \varepsilon. \tag{11}$$

Thus, we conclude that $\Pr_{\mathcal{W}}^L\{I^g(\theta, w) = 1\} \leqslant \varepsilon$, $\forall \theta \in \theta_c \oplus \bar{\gamma}\Omega_0$, equivalent to $\theta_c \oplus \bar{\gamma}\Omega_0 \subseteq \Omega(\varepsilon)$. □

Unlike regular probabilistic scaling, the sample complexity in PBPS is given by two parameters, namely the number of packs M and the size of each pack L. The sample complexity is calculated as $N = ML$. Consequently, condition (10) is defined by three tunable parameters: M, L and s. Similar to the discarding parameter r of regular probabilistic scaling, large values of the discarding parameter of each pack s make the approximating set more insensitive to extreme values. As for M and L, one could choose them according to any criterion, e.g., minimize the sample complexity N. Further details can be found in [22].

In the next section, we extend the PBPS approach introducing a constraint tightening scheme, namely the *tight immersion*, to obtain a tighter approximation of the ε-CCS. Moreover, this extension will provide a clear way to select the tuning parameters, as discussed in Section 4.1.

4. TIGHT IMMERSION

First, we introduce the notion of tight immersion.

Definition 7. τ-tight immersed. The set $\mathcal{S}$ is τ-tight immersed in the ε-CCS $\Omega(\varepsilon)$ if

$$\mathcal{S} \subseteq \Omega(\varepsilon), \tag{12a}$$
$$\mathcal{S} \not\subseteq \Omega(\tau\varepsilon), \tag{12b}$$

where $\tau \in [0, 1)$ is a measure of tightening.

Remark 3. If the ε-CCS $\Omega(\varepsilon)$ is strictly increasing with respect to ε, i.e., $\forall \tau_1, \tau_2 \in [0, 1)$ with $\tau_1 < \tau_2$, it follows that $\Omega(\tau_1\varepsilon) \subset \Omega(\tau_2\varepsilon)$. Hence, the larger τ is, the larger $\Omega(\tau\varepsilon)$ will be.

Tight immersion guarantees not only that the approximation set is *inside* the ε-CCS (12a), but also that it will be *not inside* a conservative set characterized by τ (12b). Therefore, it imposes a more restrictive condition than the regular *inner approximation*. However, tight immersion should never be used to compare the goodness of two different geometries. Indeed, as illustrated in Fig. 4,

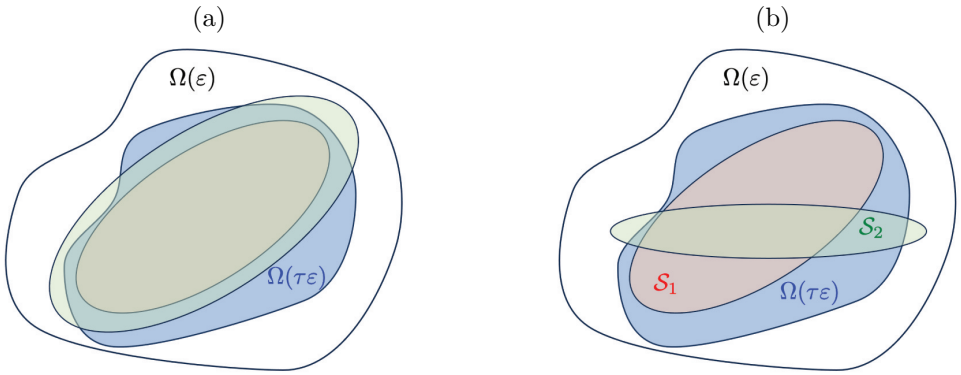

Fig. 4. Illustration of the concept of tight immersion.

we note that for the same geometry, the set with the largest value of τ fits the ε-CCS better (Fig. 4a). Instead, from Fig. 4b we note that for different geometries tight immersion by itself does not imply good approximation.

The following property is complementary to the Definition 7 of tight immersion.

Property 2. If the approximating set $\underline{\Omega}(\varepsilon)$ is τ-tight immersed in the set $\Omega(\varepsilon)$, then it is also $\tilde{\tau}$-tight immersed in it, with $\tilde{\tau} \in [0, \tau)$.

Proof. From Definition 7, we know that being $\underline{\Omega}(\varepsilon)$ τ-tight immersed in the set $\Omega(\varepsilon)$ implies $\underline{\Omega}(\varepsilon) \not\subseteq \Omega(\tau\varepsilon)$. Then, for any $\tilde{\tau} \in [0, \tau)$, we have $\Omega(\bar{\tau}\varepsilon) \subseteq \Omega(\tau\varepsilon)$. Consequently, the condition $\underline{\Omega}(\varepsilon) \not\subseteq \Omega(\bar{\tau}\varepsilon)$ (12b) holds for any $\tilde{\tau} \in [0, \tau)$ and this concludes the proof. □

Next, we finally demonstrate how to determine the pack parameters (M, L, s) so that, upon pack-based probabilistic scaling, the condition (12b) is met with confidence $1 - \bar{\delta}$, with $\bar{\delta} \in (0, 1)$. Hence, given a SAS Ω_0 centered in θ_c, we aim to determine the optimal scaling factor $\bar{\gamma}$ so that the scaled set $\mathcal{S} = \theta_c \oplus \bar{\gamma}\Omega_0$ is tight-immersed in $\Omega(\varepsilon)$.

Theorem 2 [tight-immersed pack-based probabilistic scaling]. *Consider the SAS with shape Ω_0 and scaling center θ_c, accuracy parameter $\varepsilon \in (0, 1)$, confidence level $\bar{\delta} \in (0, 1)$, tightening parameter $\tau \in [0, 1)$, and non negative integers M, L, s, with $L > s$ and such that the following condition holds*

$$B(s; L, \tau\varepsilon)^M \geqslant 1 - \bar{\delta}. \tag{13}$$

Draw M i.i.d. multisamples $\mathbf{z}_i \in \mathcal{W}^L$, with $i = 1, \ldots, M$, and define the pack scaling factor related to each i pack of random constraints as in (6), i.e., $\gamma_i = \gamma^s(\theta_c, \Omega_0, \mathbf{z}_i)$. Suppose that $\bar{\gamma} = \gamma_{1:M} > 0$. Then, with probability no smaller than $1 - \bar{\delta}$,

$$\theta_c \oplus \bar{\gamma}\Omega_0 \subseteq \Omega(\varepsilon), \quad \theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Omega(\tau\varepsilon).$$

Proof. Let $p = B(s; L, \tau\varepsilon)$. According to Property 5 in Appendix C, if we select the parameters (M, L, s) such that (13) holds, then we have that the optimal scaling factor is $\bar{\gamma} = \gamma_{1:M} > 0$ satisfies, with probability no smaller than $1 - \delta$, $\Pr_{\mathcal{W}^L}\{\theta_c \oplus \bar{\gamma}\Omega_0 \subseteq \Phi_s(\mathbf{z})\} \leqslant p$, equivalently rewritten as

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta, \mathbf{z}) = 0, \ \ \forall \theta \in \theta_c \oplus \bar{\gamma}\Omega_0\} \leqslant p.$$

Then, from Property 4 in Appendix B, we know that

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta, \mathbf{z}) = 0\} \leqslant p \iff \Pr_{\mathcal{W}}\{I^g(\theta, w) = 0\} \leqslant 1 - \tau\varepsilon. \tag{14}$$

Therefore, we can conclude that

$$\Pr_{\mathcal{W}}\{I^g(\theta, w) = 0\} \leqslant 1 - \tau\varepsilon, \quad \forall \theta \in \theta_c \oplus \bar{\gamma}\Omega_0,$$

i.e., $\theta_c \oplus \bar{\gamma}\Omega_0 \not\subseteq \Omega(\tau\varepsilon)$ with probability no smaller than $1 - \bar{\delta}$. □

Remark 4. We note that in Theorem 2 we use the tightening confidence $1 - \bar{\delta}$ instead of the original confidence $1 - \delta$. This tightening confidence is *user-defined* and can be set lower than the original confidence to limit the sample complexity.

4.1. Design of the Pack Parameters

In this section, we show how to design the parameters (M, L, s) of the pack-based approach to meet tight immersion with confidences δ and $\bar{\delta}$, respectively. From Property 1 and Theorem 2, we know that conditions (12a) and (12b) hold if the pack parameters (M, L, s) are selected such that

$$M \ln B(s; L, \varepsilon) \leqslant \ln \delta, \tag{15a}$$

$$M \ln B(s; L, \tau\varepsilon) \geqslant \ln(1 - \bar{\delta}). \tag{15b}$$

We note that (15a) embeds the probabilistic guarantees whereas (15b) is only used to tighten the solution. Since $B(s;L,\varepsilon)$ is a negative quantity, we can divide (15a) by $\ln B(s;L,\varepsilon)$ to obtain

$$M \geqslant \frac{\ln \delta}{\ln B(s;L,\varepsilon)}.$$

Hence, to satisfy (15a), it suffices to select M such that

$$M = \left\lceil \frac{\ln \delta}{\ln\, B(s;L,\varepsilon)} \right\rceil. \tag{16}$$

Analogously, for (15b) we have that M shall be selected so that the following condition holds

$$M \leqslant \frac{\ln(1-\bar{\delta})}{\ln B(s;L,\tau\varepsilon)}. \tag{17}$$

For a given set of probabilistic and tightening parameters $(\varepsilon, \delta, \tau, \bar{\delta})$, there exist multiple combinations of (M, L, s) that meet (16) and (17). In this paper, we propose two different criteria ζ: (i) minimize the number of possible combinations of $s+1$ constraints, i.e., $\zeta = M\binom{L}{s+1}$, or (ii) minimize the total sample complexity, i.e., $\zeta = ML$. Then, the pack parameters (M, L, s) are the solution of the following optimization problem

$$\begin{aligned} (M^o, L^o, s^o) = \underset{M,L,s\in\mathbb{N}_{\geqslant 0}}{\operatorname{argmin}} \quad & \zeta \\ \text{s.t.} \quad & M \leqslant \frac{\ln(1-\bar{\delta})}{\ln B(s;L,\tau\varepsilon)} \\ & M = \left\lceil \frac{\ln \delta}{\ln B(s;L,\varepsilon)} \right\rceil, \\ & L \geqslant s+1. \end{aligned} \tag{18}$$

To solve Problem (18), we exploit the *exhaustive search* approach [31] to find a proper combination of the pack parameters (M, L, s), as shown in the following example.

Example 2. Given $\varepsilon = 0.05$, $\delta = 0.001$, $\bar{\delta} = 0.1$, for each $s = [1, 30]$, we set M according to (16). Then, we test the values $L = [s+1, \cdots, s+300]$ and check if the pairs (L, s) satisfy (17). Last, among all the pairs that satisfy (17), we select the one that minimizes the ζ criterion (either $\zeta = M\binom{L}{s+1}$ or $\zeta = ML$). In Table 1, we report the pack parameters (M, L, s) obtained by solving Problem (18) with the proposed approach using both criteria ζ. Table 1 shows that, for either criterion, increasing the tightening parameter τ results in an increase in both the number of samples required (N) and the combinatorial complexity $(M\binom{L}{s+1})$. When the number of available samples

Table 1. Pack parameters, sample complexity and number of possible combinations of TI-PBPS for different values of $\tau = [0.2, 0.3, 0.4, 0.5]$ minimizing the two different criteria ζ

Criterion: minimize $M\binom{L}{s+1}$						Criterion: Minimize N				
τ	M	L	s	N	$M\binom{L}{s+1}$	M	L	s	N	$M\binom{L}{s+1}$
0.2	43	27	2	1.16e+03	1.25e+05	2	195	4	3.90e+02	4.46e+09
0.3	155	27	3	4.19e+03	2.72e+06	2	303	8	6.06e+02	1.05e+17
0.4	2681	20	4	5.36e+04	4.16e+07	4	278	10	1.11e+03	6.28e+19
0.5	15033	29	6	4.36e+05	2.35e+10	8	309	14	2.47e+03	9.68e+25

is limited, it is possible to achieve tight immersion with as little as 390 samples at the cost of a high combinatorial complexity. Moreover, it is possible to add any limitation on the number of samples $N = ML$ as a constraint in the optimization problem (18) and obtain the pack parameters that minimize the combinatorial complexity while satisfying the constraint on sample complexity. Notice that tight immersion is usually computed offline; therefore, the complexity of calculating the approximation of the CCS does not interfere with online control loops.

5. RESULTS

In this section, we use Example 1 to evaluate the approximations of the 0.05-CCS set by employing both regular PS and TI-PBPS for various problem dimensions n_θ. Hence, we choose the unit ball centered in the origin as the initial SAS approximation $\theta_c \oplus \Omega_0$. Therefore, the resulting approximating sets are balls centered in the origin and with radius α. Moreover, by means of a Monte Carlo simulation, we draw 10^7 random constraints from a uniform distribution of the constraints tangent to the unit circle of each studied dimension. Taking advantage of the symmetry of the problem, we calculate the points where the random constraints intersect a fixed axis and use them to compute the exact value of the radius for 0.05-CCS. Then, we compute the radii of the approximating sets obtained by employing regular probabilistic scaling and the novel TI-PBPS for different levels of tightening τ. To reduce variability, the TI-PBPS radii correspond to the median radius of three separated experiments, each containing different realizations of the constraints.

In Fig. 5 we can observe how, for this particular problem, the TI-PBPS is able to substantially improve the result from regular PS (dashed black line), providing approximating radii more similar to the real one (dashed red line). Moreover, as expected, the tightening of TI-PBPS improves as τ increases.

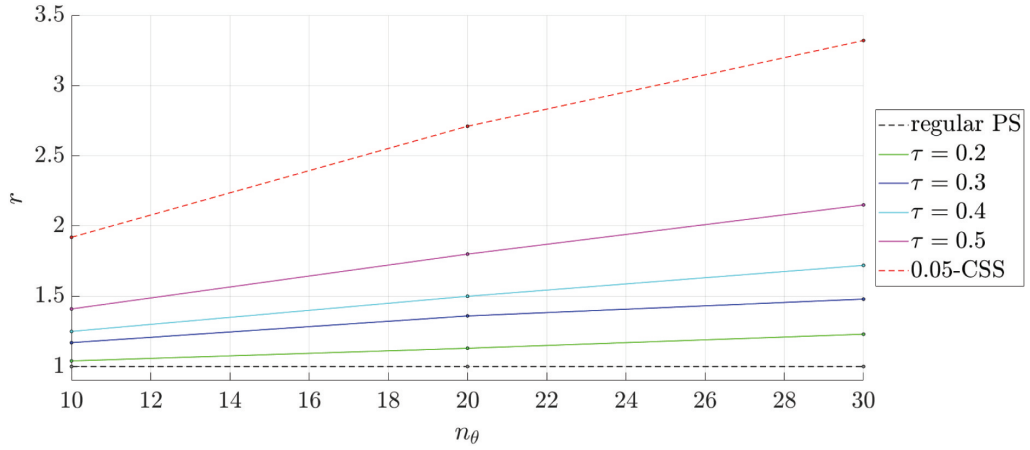

Fig. 5. Comparison of the radius (α) of the approximation set for different problem dimensions n_θ obtained by applying regular PS and TI-PBPS for $\varepsilon = 0.05$, $\delta = 0.001$, and $\bar{\delta} = 0.1$.

6. CONCLUSIONS

In this paper we have presented the probabilistic scaling approach to compute sample-based approximations of a chance constrained set. The proposed approach allows the user to first choose any set and then apply a linear transformation to approximate the safe region with the desired probabilistic guarantees. As a result, the complexity of the approximation is tuned a priori. A pack-based variant of probabilistic scaling with a tight-immersed approach is proposed, which prevents the solution from being conservative. The trade-off between the number of samples, problem complexity,

and the level of conservativeness of this approach can be tuned by the user. Future research directions point towards improving the proposed solution, e.g. applying importance sampling schemes, inspired by [32].

APPENDIX A

Property 3. Consider the integer parameters $L > s \geqslant 0$, the pack $\mathbf{z} \in \mathcal{W}^L$, and the probability parameter $\varepsilon \in (0,1)$. Then, for $w \in \mathcal{W}$, it holds

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=1\} \leqslant 1 - B(s;L,\varepsilon) \iff \Pr_{\mathcal{W}}\{I^g(\theta,w)=1\} \leqslant \varepsilon. \tag{A.1}$$

Proof. Define $E(\theta) = \Pr_{\mathcal{W}}\{I^g(\theta,w)=1\}$. Then, we have

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=0\} = \sum_{i=0}^{s}\binom{L}{i}E(\theta)^i(1-E(\theta))^{L-i} = B(s;L,E(\theta)). \tag{A.2}$$

Denote $p = 1-B(s;L,\varepsilon)$. Since $B(s;L,\varepsilon)$ is a strictly decreasing function of ε (see [23, Property 4]), we have

$$B(s;L,E(\theta)) \geqslant B(s;L,\varepsilon) = 1-p \iff E(\theta) \leqslant \varepsilon. \tag{A.3}$$

Therefore, we have

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=1\} \leqslant p \iff \Pr_{\mathcal{W}}\{I^g(\theta,w)=1\} \leqslant \varepsilon, \tag{A.4}$$

which concludes the proof. □

APPENDIX B

Property 4. Consider the integer parameters $L > s \geqslant 0$, the pack $\mathbf{z} \in \mathcal{W}^L$, the sample $w \in \mathcal{W}$, and probability parameter $\varepsilon \in (0,1)$. Then, we have

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=0\} \leqslant B(s;L,\tau\varepsilon) \iff \Pr_{\mathcal{W}}\{I^g(\theta,w)=0\} \leqslant 1-\tau\varepsilon.$$

Proof. Recalling the definition of $E(\theta) = \Pr_{\mathcal{W}}\{I^g(\theta,w)=1\}$, we have

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=0\} = \sum_{i=0}^{s}\binom{L}{i}E(\theta)^i(1-E(\theta))^{L-i} = B(s;L,E(\theta)).$$

Since $B(s;L,\tau\varepsilon)$ is strictly decreasing with respect of $\tau\varepsilon$ (Property 4 of [23]), we obtain

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=0\} = B(s;L,E(\theta)) \leqslant B(s;L,\tau\varepsilon) \iff \Pr_{\mathcal{W}}\{I^g(\theta,w)=1\} = E(\theta) \geqslant \tau\varepsilon.$$

Therefore,

$$\Pr_{\mathcal{W}^L}\{I_s^g(\theta,\mathbf{z})=0\} \leqslant B(s;L,\tau\varepsilon) \iff \Pr_{\mathcal{W}}\{I^g(\theta,w)=0\} \leqslant 1-\tau\varepsilon. \qquad \square$$

APPENDIX C

Property 5. Given the accuracy parameter $p \in (0,1)$ and the confidence level $\bar{\delta} \in (0,1)$, suppose that the number of packs M is chosen such that the following condition holds

$$1 - p^M \leqslant \bar{\delta},$$

Then, for each pack of constraints $i = 1, \dots, M$, draw the M i.i.d. multisamples $\mathbf{z} \sim \Pr_{\mathcal{W}^L}$ and define $\gamma_i = \gamma^s(c, \Omega_0, \mathbf{z}_i)$. Suppose that $\bar{\gamma} = \gamma_{1:M} > 0$. Then, with probability no smaller than $1 - \bar{\delta}$,

$$\Pr^M_{\mathcal{W}^L}\{c \oplus \bar{\gamma}\Omega_0 \subseteq \Phi^g_s(\mathbf{z})\} \leqslant p.$$

Proof. The proof follows the one of Property 1. Consider the following optimization problem

$$\begin{aligned} \min_{\gamma} \quad & \gamma \\ s.t. \quad & c \oplus \gamma\Omega_0 \not\subseteq \Phi^g_s(\mathbf{z}_)), \quad i = 1, \dots, M. \end{aligned} \tag{C.1}$$

If problem (C.1) has a feasible solution, according to (6) we can rewrite (C.1) as

$$\begin{aligned} \min_{\gamma} \quad & \gamma \\ s.t. \quad & \gamma > \gamma^s(c, \Omega_0, \mathbf{z}_)), \quad i = 1, \dots, M. \end{aligned} \tag{C.2}$$

According to the sampling-and-discard approach [21, 26], if one discards no more than $M - 1$ constraints and the number of decision variables d is $d = 1$, then the probability of violating the scaled approximating constraint set is no larger than $p \in (0, 1)$, with probability no smaller than $1 - \bar{\delta}$, where the confidence level $\bar{\delta}$ is defined as follows:

$$\begin{aligned} \bar{\delta} &= \begin{pmatrix} M-1 \\ M-1 \end{pmatrix} B(M-1; M, p) = \sum_{i=0}^{M-1} \begin{pmatrix} M \\ i \end{pmatrix} p^i (1-p)^{M-i} \\ &= 1 - \sum_{i=M}^{M} \begin{pmatrix} M \\ i \end{pmatrix} p^i (1-p)^{M-i} = 1 - p^M \end{aligned}$$

If we remove no more than $M - 1$ constraints, the optimal solution to Problem (C.2) is given by $\bar{\gamma} = \gamma_{1:M}$, with $\gamma_i = \gamma^s(c, \Omega_0, \mathbf{z})$. Correspondingly, we have $\Pr^M_{\mathcal{W}^L}\{\bar{\gamma} \leqslant \gamma(c, \Omega_0, \mathbf{z})\} \leqslant p$, from which we can conclude that, with probability no smaller than $1 - \bar{\delta}$,

$$\Pr^M_{\mathcal{W}^L}\{\theta_c \oplus \bar{\gamma}\Omega_0 \subseteq \Phi^g_s(\mathbf{z})\} \leqslant p. \qquad \square$$

REFERENCES

1. Bemporad, A. and Morari, M., Robust Model Predictive Control: A Survey, in *Robustness in Identification and Control*, Garulli, A. and Tesi, A., Eds., *Lect. Notes Control Inf. Sci.*, vol. 245, London: Springer, 1999, pp. 1–25. https://doi.org/10.1007/BFb0109870
2. Mayne, D., Seron, M. and Raković, S., Robust Model Predictive Control of Constrained Linear Systems with Bounded Disturbances, *Automatica*, 2005, vol. 41, no. 2, pp. 219–224.
3. Mayne, D., Raković, S., Findeisen, R., and Allgöwer, F., Robust Output Feedback Model Predictive Control of Constrained Linear Systems, *Automatica*, 2006, vol. 42, no. 7, pp. 1217–1222.
4. Lorenzen, M., Dabbene, F., Tempo, R., and Allgöwer, F., Constraint-Tightening and Stability in Stochastic Model Predictive Control, *IEEE Trans. Autom. Control*, 2016, vol. 62, no. 7, pp. 3165–3177.
5. Farina, M., Giulioni, L., and Scattolini, R., Stochastic Linear Model Predictive Control with Chance Constraints – a Review, *J. Process Control*, 2016, vol. 44, pp. 53–67.
6. Mesbah, A., Stochastic Model Predictive Control: An Overview and Perspectives for Future Research, *IEEE Control Syst. Mag.*, 2016, vol. 36, no. 6, pp. 30–44.
7. Charnes, A. and Cooper, W., Chance Constraints and Normal Deviates, *J. Amer. Statist. Assoc.*, 1962, vol. 57, no. 297, pp. 134–148.

8. Kücükyavuz, S. and Jiang, R., Chance-Constrained Optimization under Limited Distributional Information: A Review of Reformulations Based on Sampling and Distributional Robustness, *EURO J. Comput. Optim.*, 2022, vol. 10, art. no. 100030.

9. Mammarella, M., Mirasierra, V., Lorenzen, M., Alamo, T., and Dabbene, F., Chance-Constrained Sets Approximation: A Probabilistic Scaling Approach, *Automatica*, 2022, vol. 137, art. no. 110108.

10. Polyak, B. and Tempo, R., Probabilistic Robust Design with Linear Quadratic Regulators, *Syst. Control Lett.*, 2001, vol. 43, no. 5, pp. 343–353.

11. Calafiore, G. and Polyak, B., Stochastic Algorithms for Exact and Approximate Feasibility of Robust LMIs, *IEEE Trans. Autom. Control*, 2001, vol. 46, no. 11, pp. 1755–1759.

12. Dabbene, F., Gay, P., and Polyak, B., Recursive Algorithms for Inner Ellipsoidal Approximation of Convex Polytopes, *Automatica*, 2003, vol. 39, pp. 1773–1781.

13. Calafiore, G. and Campi, M., The Scenario Approach to Robust Control Design, *IEEE Trans. Autom. Control*, 2006, vol. 51, no. 5, pp. 742–753.

14. Tempo, R., Calafiore, G., and Dabbene, F., *Randomized Algorithms for Analysis and Control of Uncertain Systems, with Applications*, London: Springer, 2013.

15. Polyak, B. and Shcherbakov, P., Randomization in Robustness, Estimation, and Optimization, in *Uncertainty in Complex Networked Systems: In Honor of Roberto Tempo*, Basar, T., Ed., Cham: Springer, 2018, pp. 181–208.

16. Kataoka, S., A stochastic Programming Model, *Econometrica: J. Econom. Soc.*, 1963, pp. 181–196.

17. Prékopa, A., Logarithmic Concave Measures with Application to Stochastic Programming, *Acta Scientiarum Mathematicarum*, 1971, vol. 32, pp. 301–316.

18. Geng, X. and Xie, L., Data-Driven Decision Making in Power Systems with Probabilistic Guarantees: Theory and Applications of Chance-Constrained Optimization, *Annu. Rev. Control*, 2019, vol. 47, pp. 341–363.

19. Soudjani, S., and Majumdar, R., Concentration of Measure for Chance-Constrained Optimization, *IFAC-PapersOnLine*, 2018, vol. 51, no. 16, pp. 277–282.

20. Lejeune, M. and Prékopa, A., Relaxations for Probabilistically Constrained Stochastic Programming Problems: Review and Extensions, *Annals of Operations Research*, 2018, pp. 1–22.

21. Campi, M. and Garatti, S., A Sampling-and-Discarding Approach to Chance-Constrained Optimization: Feasibility and Optimality, *J. Optim. Theory Appl.*, 2011, vol. 148, no. 2, pp. 257–280.

22. Alamo, T., Mirasierra, V., Dabbene, F., and Lorenzen, M., Safe Approximations of Chance Constrained Sets by Probabilistic Scaling, *Proc. 18th Eur. Control Conf. (ECC)*, IEEE, 2019, pp. 1380–1385.

23. Alamo, T., Tempo, R., Luque, A., and Ramirez, D., Randomized Methods for Design of Uncertain Systems: Sample Complexity and Sequential Algorithms, *Automatica*, 2015, vol. 52, pp. 160–172.

24. Lorenzen, M., Dabbene, F., Tempo, R., and Allgöwer, F., Stochastic MPC with Offline Uncertainty Sampling, *Automatica*, 2017, vol. 81, pp. 176–183.

25. Mammarella, M., Lorenzen, M., Capello, E., Park, H., Dabbene, F., Guglieri, G., Romano, M., and Allgöwer, F., An Offline-Sampling SMPC Framework with Application to Autonomous Space Maneuvers, *IEEE Trans. Control Syst. Technol.*, 2018, vol. 28, no. 2, pp. 388–402.

26. Calafiore, G., Random Convex Programs, *SIAM J. Optim.*, 2010, vol. 20, no. 6, pp. 3427–3464.

27. Vapnik, V., *The Nature of Statistical Learning Theory*, New York: Springer Science & Business Media, 1999.

28. Alamo, T., Tempo, R., and Camacho, E., Randomized Strategies for Probabilistic Solutions of Uncertain Feasibility and Optimization Problems, *IEEE Trans. Autom. Control*, 2009, vol. 54, no. 11, pp. 2545–2559.

29. Alamo, T., Tempo, R., and Camacho, E., Improved Sample Size Bounds for Probabilistic Robust Control Design: A Pack-Based Strategy, *Proc. 46th IEEE Conf. Decis. Control*, IEEE, 2007, pp. 6178–6183.

30. Alamo, T., Manzano, J., and Camacho, E., Robust Design through Probabilistic Maximization, in *Uncertainty in Complex Networked Systems*, Springer, 2018, pp. 247–274.

31. Nievergelt, J., Exhaustive Search, Combinatorial Optimization and Enumeration: Exploring the Potential of Raw Computing Power, *Proc. Int. Conf. Curr. Trends Theory Pract. Comput. Sci.*, Springer, 2000, pp. 18–35.

32. Lukashevich, A., Gorchakov, V., Vorobev, P., Deka, D., and Maximov, Y., Importance Sampling Approach to Chance-Constrained DC Optimal Power Flow, *IEEE Trans. Control Netw. Syst.*, 2023, vol. 11, no. 2, pp. 928–937.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 756–768.

SPECIAL ISSUE

Optimal Robust Tracking of a Discrete Minimum-Phase Plant under the Unknown Bias and Norm of an External Disturbance and the Unknown Norm of Uncertainties

V. F. Sokolov

Komi Scientific Center, Ural Branch, Russian Academy of Sciences, Syktyvkar, Russia
e-mail: sokolov@ipm.komisc.ru

Received March 3, 2025
Revised May 20, 2025
Accepted June 27, 2025

Abstract—This paper addresses a problem of the optimal robust tracking of a given bounded reference signal for a discrete-time minimum-phase plant with a known approximate nominal model under a bounded and biased external disturbance and coprime factor perturbations. The bias and norm of the external disturbance and the gains of the perturbations are assumed to be unknown. The control criterion is the worst-case asymptotic tracking error in the class of the disturbances and perturbations under consideration, which depends on the above unknown parameters and the reference signal. A solution of the optimal tracking problem with a given accuracy is based on optimal errors quantification within the ℓ_1-theory of robust control, polyhedral estimation of the unknown parameters, and treating the control criterion as the identification criterion.

Keywords: robust control, optimal control, bounded disturbance, uncertainty, errors quantification, set-membership approach

DOI: 10.31857/S0005117925080053

1. INTRODUCTION

This paper addresses the optimal tracking problem of a linear discrete dynamic plant with a given and tested transfer function. By assumption, the plant is affected by a bounded external disturbance with an unknown bias and unknown bounds and by perturbations (uncertainties) for its output and control with unknown norms (gains). The problem is addressed within the ℓ_1-theory of robust control, laid down in [1, 2] and corresponding to the signal space ℓ_∞ of bounded real sequences. The problem has the following difficulty: to minimize a criterion in the form of the worst-case asymptotic tracking error in the class of admissible disturbances, it is necessary to compensate for the unknown bias and justify an optimal estimator for the criterion under the non-identifiability of all the unknown parameters mentioned above.

The solution of the optimal tracking problem described is based on optimal errors quantification using the set-membership approach and treating the control criterion as an ideal identification criterion. The set-membership approach in system identification, initially involving the assumption of known upper bounds on deterministic disturbances, gained wide popularity in the late 1980s and was reduced to the development of computable upper and lower approximations (ellipsoids, parallelotopes, etc.) of parameter sets consistent with measurement data. (Here, we refer to the first special issues of two leading journals on control theory [3, 4].) Applications of these approximations to control problems are rarely described and are accompanied by various additional assumptions, such as a priori known stabilizing control. This approach is criticized by supporters of stochastic

disturbance models for its conservatism caused by a priori assumptions on known upper bounds on disturbances. In parallel, active research in the field of identification for robust control and uncertainty quantification began in the early 1990s. A decade and a half later, it was noted in the review [5] that estimation of uncertainty sets was often mistakenly attributed to identification for control, as in most of the corresponding studies, the control objective was not considered during identification. The problems of model verification and uncertainty estimation remain topical to the present time [6, 7], but are still considered mainly beyond the context of control problems and with artificial criteria motivated by the objectives of identification itself.

In this paper, bias estimation and errors quantification are based on the set-membership approach and treating the control criterion as an ideal identification criterion. The potential applicability of such a combined framework arises from two circumstances. First, in the ℓ_1-theory of robust control, explicit representations are obtained for asymptotic performance indices in terms of induced norms of the transfer functions of a closed-loop control system and the norms of all disturbances and uncertainties [8–11]. Second, the bounded disturbance model allows for the direct use of current measurement data for online model verification [12]. In the general case, such an approach to control-oriented identification is computationally intractable due to the complexity of computing current optimal estimates. But it is computationally tractable in the case of linear or linear-fractional, with respect to the estimated parameters, performance indices [13]. In the problem under consideration, the performance index (control criterion) is a non-convex quadratic-fractional function of the unknown parameters (see the representation (9)). For a known bias, the control criterion becomes linear-fractional, and the problem of errors quantification for this case was solved in [14], where the idea of estimating the unknown bias using a grid of test values was also formulated. Below, we rigorously justify this idea and prove a rigorous result on the solution of the asymptotically optimal tracking problem with a given accuracy. Simulation results and related remarks illustrate the effectiveness of the solution proposed.

Notation:
$|\varphi|$ is the Euclidean norm of a vector $\varphi \in \mathbb{R}^n$;
$x_s^t = (x_s, x_{s+1}, \dots, x_t)$ for a real sequence $x = (\cdots, x_{-1}, x_0, x_1, \cdots)$;
$|x_s^t| = \max_{s \leqslant k \leqslant t} |x_k|$;
$\|x\|_{ss} = \limsup_{t \to +\infty} |x_t|$;
$\|x\|_\infty = \sup_t |x_t|$ is the norm in the space ℓ_∞ of bounded sequences;
$\|x\|_1 = \sum_{k=0}^{+\infty} |x_k|$ is the norm in the space ℓ_1 of absolutely summable sequences;
$\|G\| = \sum_{k=0}^{+\infty} |g_k| = \|g\|_1$ is the induced norm of a stable linear time-invariant causal system $G : \ell_\infty \to \ell_\infty$ with a transfer function $G(\lambda) = \sum_{k=0}^{+\infty} g_k \lambda^k$.

2. THE PLANT MODEL AND MEANINGFUL PROBLEM STATEMENT

The plant model is described by the equation

$$a(q^{-1})y_t = b(q^{-1})u_t + v_t, \quad t = 1, 2, 3, \dots, \tag{1}$$

where $y_t \in \mathbb{R}$ is the measured output of the plant at a time instant t, $u_t \in \mathbb{R}$ is the control input, $v_t \in \mathbb{R}$ is a total disturbance, and q^{-1} is the backward shift operator ($q^{-1}y_t = y_{t-1}$). The initial conditions $y_{1-n}^0 = (y_{1-n}, \dots, y_0)$ are arbitrary, and $u_t = 0$ for $t \leqslant 0$. The polynomials

$$a(\lambda) = 1 + a_1\lambda + \ldots + a_n\lambda^n, \quad b(\lambda) = b_1\lambda + \ldots + b_m\lambda^m$$

characterize the **nominal model** of the plant, i.e., the model without the disturbance v. The total disturbance v has the form

$$v_t = c^w + \delta^w w_t + \delta^y \Delta^1(y)_t + \delta^u \Delta^2(u)_t, \quad \|w\|_\infty \leqslant 1, \ \delta^w > 0, \ \delta^y > 0, \ \delta^u > 0. \tag{2}$$

The parameters c^w and δ^w in (2) characterize the bias of the external disturbance $c^w + \delta^w w_t$ and the upper bound on the unbiased disturbance $\delta^w w$, respectively. The numbers $\delta^y > 0$ and $\delta^u > 0$ are the gains (induced norms) of the perturbations affecting the output and control, respectively, and

$$|\Delta^1(y)_t| \leqslant p_t^y = \max_{t-\mu\leqslant k\leqslant t-1} |y_k|, \quad |\Delta^2(u)_t| \leqslant p_t^u = \max_{t-\mu\leqslant k\leqslant t-1} |u_k|. \tag{3}$$

In the ℓ_1-theory of robust control, these perturbations are called uncertainties with limited memory μ, which ensures the independence of the asymptotic dynamics of the closed-loop control system from the initial data. The uncertainty memory μ is chosen by the designer to be arbitrarily large, but not infinite, without compromising the guaranteed control performance. The description of disturbances in the form (2), (3) is equivalent [10, 11] to the inequalities

$$|v_t - c^w| \leqslant \delta^w + \delta^y p_t^y + \delta^u p_t^u \ \ \forall t. \tag{4}$$

A priori information about the plant is contained in the following assumptions.

A1. The polynomials $a(\lambda)$ and $b(\lambda)$ of the nominal plant are known, $b_1 \neq 0$.

A2. The roots of the polynomial $\frac{b(\lambda)}{\lambda}$ lie outside the closed unit circle of the complex plane.

A3. The parameter column vector $\delta = (\delta^w, \delta^y, \delta^u)^{\mathrm{T}}$ is unknown, and the bias $c^w \in [c^w_{\min}, c^w_{\max}]$ is unknown, albeit with given $c^w_{\min}$ and $c^w_{\max}$.

A4. The asymptotic upper bound $\|r\|_{ss}$ of the reference signal r or its upper bound is known.

Assumption A1 also covers the case when the "true" nominal model is unknown and its estimator, obtained by some identification method, is available for testing. Assumption A2 ensures the boundedness of the control input u if the plant output y is bounded. (Such a plant is called minimum-phase.) Assumption A4 will be commented upon after the rigorous formulation of the problem at the end of Section 3. Another mandatory assumption restricting the norms of the perturbations will be introduced in Section 3 after Theorem 1.

Meaningful problem statement: it is required to design a control law minimizing the worst asymptotic tracking error of a given bounded signal for a set of disturbances satisfying inequalities (4).

To solve the optimal problem with a given accuracy, one needs to quantify the errors online (i.e., determine their unknown parameters δ) to estimate the tracking performance and compensate for the unknown bias c^w.

3. THE TRACKING PERFORMANCE OF AN OPTIMAL CONTROLLER UNDER A KNOWN BIAS c^w. PROBLEM STATEMENT

Let $r = (r_1, r_2, r_3, \ldots)$ be a given bounded signal ($r \in \ell_\infty$). The control criterion of the tracking problem has the form

$$J_\mu(c^w, \delta) = \sup_{v\in V} \|y - r\|_{ss}, \quad \|y - r\|_{ss} := \limsup_{t\to+\infty} |y_t - r_t|, \tag{5}$$

where V is the set of all disturbances v satisfying inequalities (4).

Consider a controller described by the equation

$$b(q^{-1})u_t = (a(q^{-1}) - 1)y_t + r_t - c^w. \tag{6}$$

Note that (6) specifies the value of u_{t-1}, not u_t, which does not figure in this equation. For the output of the closed-loop control system (1), (6), we then obtain

$$y_t - r_t = v_t - c^w = \delta^w w_t + \delta^y \Delta^1(y)_t + \delta^u \Delta^2(u)_t. \tag{7}$$

Due to the arbitrariness and unpredictability of the right-hand side in (7), the controller (6) is **optimal** for the control criterion (5).

Definition 1. The closed-loop system (1), (6) is said to be robustly stable in the class of disturbances V if $J_\mu(c^w, \delta) < +\infty$.

To formulate a theorem on the performance of the optimal controller (6), we denote its transfer functions relating y and r to the control input u :

$$G_{uy}(\lambda) = \frac{a(\lambda) - 1}{b(\lambda)}, \quad G_{ur}(\lambda) = \frac{1}{b(\lambda)}.$$

Theorem 1. *Under Assumptions* A1 *and* A2*, the following assertions are true.*

1) *The closed-loop system* (1), (6) *is robustly stable in the class V with a disturbance memory $\mu = +\infty$ if and only if*

$$\delta^y + \delta^u \|G_{uy}\| < 1. \tag{8}$$

For the system with the zero initial conditions y^0_{1-n} and $\mu = +\infty$,

$$J(c^w, \delta) := J_{+\infty}(c^w, \delta) = \frac{\delta^w + \delta^y \|r\|_{ss} + \delta^u (|c^w| + \|r\|_{ss}) \|1/b(q^{-1})\|}{1 - \delta^y - \delta^u \|G_{uy}\|}. \tag{9}$$

2) *For the system with arbitrary initial conditions y^0_{1-n} and $\mu < +\infty$,*

$$J_\mu(c^w, \delta) \leqslant J(c^w, \delta) \quad \forall\, \mu > 0, \tag{10}$$

and if the sequence $|r|$ uniformly often falls into the neighborhood of the upper limit $\|r\|_{ss}$ (see the definition in [10]*), then for any initial conditions*

$$J_\mu(c^w, \delta) \nearrow J(c^w, \delta) \quad (\mu \to +\infty), \tag{11}$$

where the sign $\nearrow$ means monotonic convergence from below as $\mu \to +\infty$.

The proof of Theorem 1 was given in [14].

The final assumption (A5) restricting the norms of the perturbations follows from Theorem 1.

A5. A number $\bar\delta$ is known such that

$$\delta^y + \delta^u \|G_{uy}\| \leqslant \bar\delta < 1. \tag{12}$$

Assumption A5 is not restrictive compared to the robust stability condition (8). According to the meaning of the problem, the parameter $\bar\delta$ is assigned by the designer and can be chosen arbitrarily close to 1. But for values of $\delta^y + \delta^u \|G_{uy}\|$ close to 1, the control criterion $J_\mu(c^w, \delta)$ becomes too large and the nominal model with the given tested polynomials $a(\lambda)$ and $b(\lambda)$ or the plant with such perturbations can be considered unacceptable.

Problem statement. Under a priori information A1–A5 and a given tracking signal r, it is required to design a feedback control law $u_t = U_t(y_1^t, u_1^{t-1})$ (with finite memory) that ensures the inequality

$$\|y - r\|_{ss} \leqslant J(c^w, \delta) \tag{13}$$

with a given accuracy.

The main difficulty of the problem is to ensure inequality (13) under the non-identifiability of c^w and δ (see subsection 4.1).

The index (9), used below as an identification criterion, depends on $\|r\|_{ss}$. If this value is a priori unknown, the recursively computable non-decreasing estimators $R_t = \max_{1 \leqslant k \leqslant t} |r_k| \leqslant \|r\|_\infty$ can be used instead to obtain a fundamentally theoretically unimprovable tracking performance guarantee with $\|r\|_{ss}$ replaced by $\|r\|_\infty$.

4. OPTIMAL TRACKING

The solution of the problem is based on optimal errors quantification for the nominal model being tested.

4.1. Optimal Errors Quantification with a Known Bias c^w

Due to the plant equation (1) and inequalities (4), given a known bias c^w, complete information about the unknown δ at a time instant t is contained in the a priori assumption A5 and the inclusion

$$\delta \in D_t = \left\{ \hat{\delta} \geqslant 0 \mid |a(q^{-1})y_k - b(q^{-1})u_k - c^w| \leqslant \hat{\delta}^w + \hat{\delta}^y p_k^y + \hat{\delta}^u p_k^u \ \ \forall k \leqslant t \right\}, \tag{14}$$

where $\hat{\delta} = (\hat{\delta}^w, \hat{\delta}^y, \hat{\delta}^u)^{\mathrm{T}}$. The system of inequalities in (14) is equivalent to the description of system (1)–(4) on the interval $[1, t]$ for any control u_0^{t-1}. Then the best estimator of the parameter δ in terms of the control criterion J, consistent with the measurements y_0^t and u_0^{t-1}, has the form

$$\delta_t = \operatorname*{argmin}_{\hat{\delta} \in D_t} J(c^w, \hat{\delta}). \tag{15}$$

The optimal problem (15) is a linear-fractional programming problem with the unknown row vector $\hat{\delta}$. It is reduced to a linear programming problem in the standard way by introducing an additional real variable [15]. The number of linear inequalities with respect to $\hat{\delta}$ in the description of the sets D_t can infinitely increase as t grows. To ensure the boundedness of the number of inequalities and the convergence of the polyhedral and vector estimators of the unknown column vector δ in finite time, we choose the parameter $\varepsilon_1 > 0$, which specifies the dead zone size when updating the estimators. The initial polyhedral estimator of δ has the form

$$P_0 = \left\{ \hat{\delta} = (\hat{\delta}^w, \hat{\delta}^y, \hat{\delta}^u)^{\mathrm{T}} \mid \hat{\delta} \geqslant 0, \ \hat{\delta}^y + \hat{\delta}^u \|G_{uy}\| \leqslant \bar{\delta} < 1 \right\}, \quad \delta_0 = (0, 0, 0)^{\mathrm{T}}.$$

Denoting

$$\nu_{t+1} = |a(q^{-1})y_{t+1} - b(q^{-1})u_{t+1} - c^w|, \quad \phi_{t+1} = (1, p_{t+1}^y, p_{t+1}^u)^{\mathrm{T}}, \tag{16}$$

we write the new inequality in the description of D_{t+1} as

$$\delta \in \Omega_{t+1} = \left\{ \hat{\delta} \mid \nu_{t+1} \leqslant \hat{\delta}^{\mathrm{T}} \phi_{t+1} \right\}. \tag{17}$$

Let P_t and δ_t be the polyhedral and vector estimators of δ at a time instant t. We set

$$P_{t+1} = \begin{cases} P_t & \text{if } \nu_{t+1} \leqslant \delta_t^{\mathrm{T}} \phi_{t+1} + \varepsilon_1 |\phi_{t+1}| \\ P_t \cap \Omega_{t+1} & \text{otherwise,} \end{cases} \tag{18}$$

$$\delta_{t+1} = \operatorname*{argmin}_{\hat{\delta} \in P_{t+1}} J(c^w, \hat{\delta}). \tag{19}$$

According to (18), the polyhedral estimator P_{t+1} is updated by adding a new inequality only if the distance from δ_t to the half-space $\Omega_{t+1} \subset \mathbb{R}^3$ exceeds ε_1. Note that all estimators P_t are unbounded in the direction of growth of the variable $\hat{\delta}^w$.

4.2. Optimal Tracking under an Unknown Bias c^w

To compensate for the unknown bias $c^w \in [c^w_{\min}, c^w_{\max}]$, we will estimate it using a grid of the form

$$c_k^w = c^w_{\min} + k\varepsilon_2, \ \ k = 0, 1, \ldots, N, \qquad \varepsilon_2 = \frac{c^w_{\max} - c^w_{\min}}{N} > 0, \tag{20}$$

which yields a guaranteed estimator of the bias c^w with the desired accuracy $\varepsilon_2/2$ by choosing a sufficiently large N. For each bias c^w_k and each time instant t, we compute the polyhedral $P_{k,t}$ and vector $\delta_{k,t}$ estimators of the unknown vector δ. We define the best estimate number k_t of the vector δ at a time instant t by the formula

$$k_t = \operatorname*{argmin}_k J(c^w_k, \delta_{k,t}). \tag{21}$$

The control input u_t at a time instant t is determined by the *adaptive controller* corresponding to this estimate:

$$b(q^{-1})u_{t+1} = (a(q^{-1}) - 1)y_{t+1} + r_{t+1} - c^w_{k_t}. \tag{22}$$

After measuring the output y_{t+1}, the residuals

$$\nu_{k,t+1} = |a(q^{-1})y_{t+1} - b(q^{-1})u_{t+1} - c^w_k|$$

and the estimators $P_{k,t+1}$ and $\delta_{k,t+1}$ for $k = 0, 1, \ldots, N$ are computed according to (16)–(19). (The corresponding formulas, with the subscript k in each, are omitted here for brevity.)

Theorem 2. *Under Assumptions* A1–A5*, let the plant* (1) *be regulated by the adaptive controller* (22) *with the estimator* (16)–(19), (21) *and the dead zone parameter* ε_1 *in* (18) *such that*

$$0 < \varepsilon_1 < \frac{1 - \bar{\delta}}{1 + \|G_{uy}\|}. \tag{23}$$

Then the number of updates in the polyhedral $P_{k,t}$ *and vector* $\delta_{k,t}$ *estimators is finite for all* $k \in \{0, 1, \ldots, N\}$*, and the tracking error satisfies the inequality*

$$\|y - r\|_{ss} \leqslant J(c^w_{k_\infty}, \delta_\infty + \varepsilon_1(1,1,1)^{\mathrm{T}}) = J(c^w_{k_\infty}, \delta_\infty) + O(\varepsilon_1) \quad (\text{as } \varepsilon_1 \to 0), \tag{24}$$

where k_∞ *is the final value of the best estimate number* (21) *for the unknown* δ*,* δ_∞ *is the final value of* $\delta_{k_\infty,t}$*, and*

$$J(c^w_{k_\infty}, \delta_\infty) \leqslant J\left(c^w, \delta + \left(\frac{\varepsilon_2}{2} + \varepsilon_1, \varepsilon_1, \varepsilon_1\right)\right) = J(c^w, \delta) + O(\varepsilon_1 + \varepsilon_2) \quad (\varepsilon_1 + \varepsilon_2 \to 0). \tag{25}$$

Proof. For any control u_t and any $k \in \{0, 1, \ldots, N\}$, from the plant equation (1) and the representation (4) of the total disturbance v it follows that

$$|a(q^{-1})y_{t+1} - b(q^{-1})u_{t+1} - c^w_k| \leqslant |c^w - c^w_k| + \delta^w + \delta^y p^y_{t+1} + \delta^u p^u_{t+1} \quad \forall\, t. \tag{26}$$

By the representation (4), inequalities (26) allow treating the plant as a virtual object of the form (1), in which the virtual external disturbance has the bias c^w_k and the norm of the unbiased external disturbance does not exceed

$$\bar{\delta}^w_k = |c^w - c^w_k| + \delta^w. \tag{27}$$

We prove that the number of updates in the estimators $P_{k,t}$ and $\delta_{k,t}$ is finite for all k. For each update of the estimators, according to (18), we have

$$\varepsilon_1|\phi_{t+1}| < \nu_{t+1} - \delta_t^{\mathrm{T}}\phi_{t+1}.$$

Then, for any $\hat{\delta} \in \Omega_{t+1}$, (17) implies

$$\varepsilon_1|\phi_{t+1}| < |(\hat{\delta} - \delta_t)^{\mathrm{T}}\phi_{t+1}| \leqslant |\hat{\delta} - \delta_t||\phi_{t+1}|$$

and, consequently, $|\hat{\delta} - \delta_t| > \varepsilon_1$. Hence, for all k, the distance from the estimator $\delta_{k,t}$ to the half-space $\Omega_{k,t+1}$ is greater than ε_1. As $P_{k,t+1} \subset \Omega_{t+1}$, the distance from $\delta_{k,t}$ to $P_{k,t+1}$ is also greater than ε_1. The polyhedral estimators $P_{k,t}$ decrease monotonically in time due to the addition of new inequalities. Moreover, the balls of radius $\varepsilon_1/2$ centered at $\delta_{k,t}$ have empty intersection with similar balls centered at the future updated estimators $\delta_{k,s}$ (for $s > t$) and, consequently, with all balls $\delta_{k,s}$ for $s \neq t$. It follows that the number of possible updates in the estimators $\delta_{k,t}$ is finite for all k, since they all lie in the corresponding bounded sets $\{\hat{\delta}_k \mid J(c_k^w, \hat{\delta}_k) \leqslant J(c_k^w, (\bar{\delta}_k^w, \delta^y, \delta^u)^{\mathrm{T}})\}$, where $\bar{\delta}_k^w$ is given by (27).

We denote by $\delta_{k,\infty} = (\delta_{k,\infty}^w, \delta_{k,\infty}^y, \delta_{k,\infty}^u)^{\mathrm{T}}$ the final values, i.e., the limit values of the estimators $\delta_{k,t}$ achieved in a finite time $t_{k,\infty}$, and set $t_\infty = \max_k t_{k,\infty}$. Then $\delta_{k,t} = \delta_{k,\infty}$ for all $t \geqslant t_\infty$ and all k.

Let k_∞ be the steady-state number of the best estimate of the vector δ :

$$k_\infty = \operatorname*{argmin}_k J(c_k^w, \delta_{k,\infty}).$$

Due to (21), we have

$$J(c_{k_\infty}^w, \delta_{k_\infty}) \leqslant J(c_k^w, \delta_{k,\infty}) \quad \forall k. \tag{28}$$

For all $t \geqslant t_\infty$, in view of (18), the residuals (26) in the closed-loop adaptive system with the steady-state controller satisfy

$$\nu_{k_\infty,t} \leqslant \delta_{k,\infty}^{\mathrm{T}} \phi_t + \varepsilon_1 |\phi_t| \leqslant (\delta_{k_\infty}^{\mathrm{T}} + \varepsilon_1(1,1,1))\phi_t. \tag{29}$$

By Theorem 1, this inequality implies (24).

We denote by

$$k_* = \operatorname*{argmin}_k |c^w - c_{k_\infty}^w|$$

the number of the estimate c_k^w closest to c^w. Then $|c^w - c_{k_*}^w| \leqslant \varepsilon_2/2$ and, due to (18),

$$\begin{aligned} |a(q^{-1})y_{t+1} - b(q^{-1})u_{t+1} - c_{k_*,t}^w| &\leqslant \frac{\varepsilon_2}{2} + \delta^{\mathrm{T}}\phi_{t+1} + \varepsilon_1|\phi_t| \\ &\leqslant \delta + \left(\frac{\varepsilon_2}{2} + \varepsilon_1, \varepsilon_1, \varepsilon_1\right)\Big)^{\mathrm{T}} \phi_{t+1} \end{aligned} \tag{30}$$

for all $t \geqslant t_\infty$. According to Theorem 1, this inequality yields

$$J(c_{k_*}^w, \delta_{k_*,\infty}) \leqslant J\left(c^w, \delta + \left(\frac{\varepsilon_2}{2} + \varepsilon_1, \varepsilon_1, \varepsilon_1\right)\right). \tag{31}$$

Using (28) with $k = k_*$ and (31), we obtain inequality (25). Finally, the term $O(\varepsilon_1 + \varepsilon_2)$ in (25) follows from the fact that $J(c^w, \delta)$ is a fractional rational function of δ and its denominator is separated from 0 by Assumption A5. The proof of Theorem 2 is complete.

Remark 1. Inequalities (24) and (25) mean the suboptimality of the solution of the tracking problem (13). The estimate $O(\varepsilon_1 + \varepsilon_2)$ of the solution accuracy in the stated optimal problem, guaranteed by inequality (25), is only qualitative and cannot be used for computations since c^w and δ are unknown. For a particular realization of the disturbances, the best computable estimate of the solution accuracy in the optimal tracking problem is the current difference

$$(\Delta J)_t = J(c_{k_t}^w, \delta_{k_t,t} + \varepsilon_1(1,1,1)^{\mathrm{T}}) - J(c_{k_t}^w, \delta_{k_t,t}), \tag{32}$$

which is consistent with the measured data y_{1-n}^t, u_1^t and will be guaranteed as the estimators converge in a finite time. Although the convergence time of the estimators to the final value is

unknown, a long period of unchanged estimates actually confirms the validity of this estimate by Theorem 2. If the current estimate of the solution accuracy is unsatisfactory, one can (at any time) decrease the dead zone parameter ε_1 to improve the accuracy. In this case, the number of updates in the estimators $P_{k,t}$ and $\delta_{k,t}$ may increase. The grid step ε_2 has a more transparent impact on the optimization accuracy (the term $\varepsilon_2/2$ is added to the estimates $\delta^w_{k,t}$) and can be chosen a priori, while keeping in mind that a decrease in the grid step ε_2 will cause an increase in the number of polyhedral $P_{k,t}$ and vector $\delta_{k,t}$ estimators computed in parallel.

5. SIMULATION

Let the plant be described by equation (1) with the unknown parameters

$$\theta^* = [a_1^*, a_2^*, b_1^*, b_2^*, b_3^*] = [-2.7;\ 1.8;\ 2;\ -3.36;\ 1.4] \tag{33}$$

of the nominal model, and let a nominal model with poles 0.7 and 0.8 (the roots of $a(\lambda)$) and zeros 1.1 and 1.3 (the roots of $b(\lambda)$) and the coefficient $b_1 = 2$ be available for testing. This nominal model matches an unstable minimum-phase plant (1) with the parameters

$$\theta = [a_1, a_2, b_1, b_2, b_3] = [-2.6786;\ 1.7857;\ 2;\ 3.3566;\ 1.3986], \tag{34}$$

slightly differing from the parameters (33). Let the plant with the parameters (33) be regulated by the adaptive controller (22), which is optimal for the tested plant with the parameters (34). The characteristic polynomial of this closed-loop system has roots $0.7512 \pm 8.9242i, 1.3032$, and 1.0945 (with an accuracy of 10^{-4}), being greater than 1 by absolute value; therefore, the closed-loop system without the perturbations is stable. The dynamics of this closed-loop system can be treated as the dynamics of that with a plant with nominal parameters θ and additional relatively small perturbations

$$\Delta(y_{t-2}^{t-1}, u_{t-2}^{t-1}) = (a_1 - a_1^*)y_{t-1} + (a_2 - a_2^*)y_{t-2} + (b_1^* - b_1)u_{t-1} + (b_1^* - b_1)u_{t-2},$$

arising from the "inaccurate" coefficients of the nominal model tested. The disturbance v_t in the nominal model with the parameters θ is described by

$$v_t = c^w + \delta^w w_t + k_t^y\, \delta^y\, |y_{t-\mu}^{t-1}| + k_t^u\, \delta^u |\, u_{t-\mu}^{t-1}|,\ |k_t^y| \leqslant 1,\ |k_t^u| \leqslant 1,\ \mu = 20. \tag{35}$$

Let the tracking signal be $r_t = 10 \sin t$ for all t.

Example 1. Random disturbances. Let the unknown parameters in the description (35) have the values

$$c^w = 5,\ \delta^w = 1,\ \delta^y = \delta^u = 0.1, \tag{36}$$

and let w_t, k_t^y, k_t^u be independent pseudorandom variables uniformly distributed on $[-1, 1]$. The simulation was performed with the following adaptive control parameters: the dead zone parameter $\varepsilon_1 = 10^{-6}$, $c^w_{\min} = -10$, $c^w_{\max} = 10$, and the grid step $\varepsilon_2 = 0.5$.

Figure 1 shows the graphs of the tracking error $y - r$ on the left and the current optimal control criterion estimates $J(c^w_{k_t}, \delta_{k_t,t})$ on the right.

Next, the final values $J(c^w_k, \delta_{k,1000})$ for all k, consistent with measurements on the interval [1, 1000], are presented in Fig. 2 on the left. The switching of the estimates $c^w_{k_t,t}$ of the unknown bias $c^w = 5$ are provided in Fig. 2 on the right. Despite the symmetry of the distributions of the random variables w_t, k_t^y, k_t^u about zero, the steady-state bias estimate $c^w_{1000} = 4.5$ differs from $c^w = 5$.

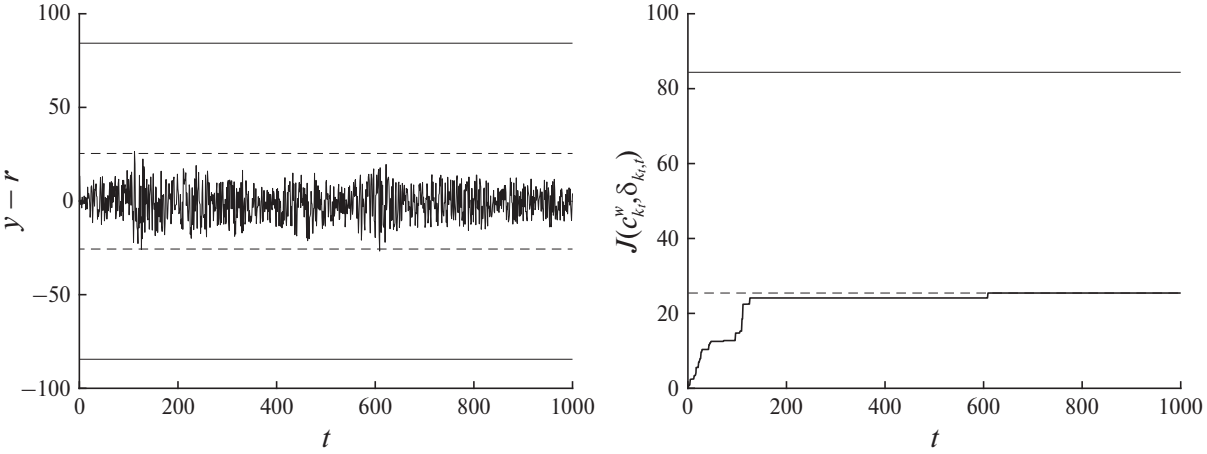

Fig. 1. The graphs of $y - r$ (left) and $J(c^w_{k_t}, \delta_{k_t,t})$ (right). Solid lines correspond to $\pm J(c^w, \delta)$ and dashed lines to $\pm J(c^w_{k_\infty}, \delta_{k_\infty})$.

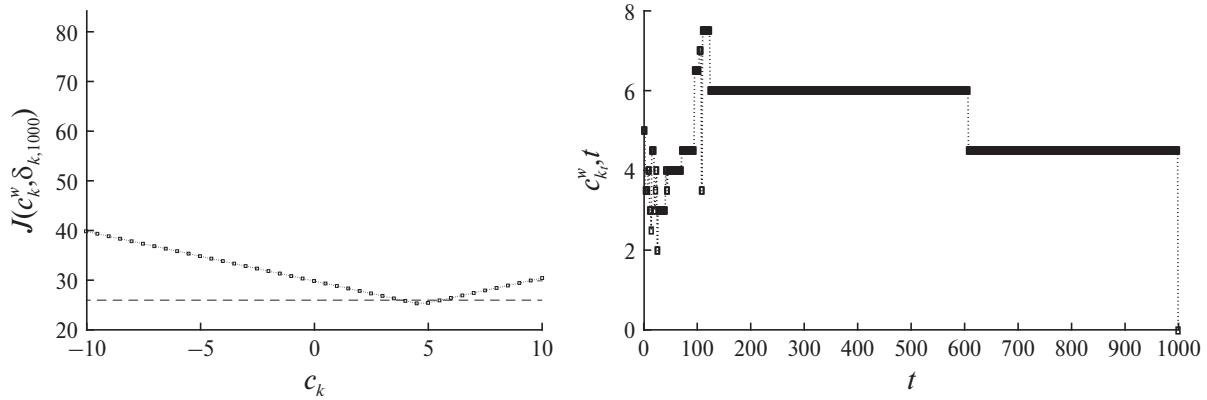

Fig. 2. The values $J(c^w_k, \delta_{k,1000})$ (left) and the switching of the estimates $c^w_{k_t,t}$ (right).

In all the numerical experiments with random perturbations and deterministic "oscillatory" disturbances, the steady-state upper bounds of the tracking error $J(c^w_{k_\infty}, \delta_{k_\infty})$, consistent with the measurements, are significantly (several times) smaller than the unknown optimal upper bound of $J(c^w, \delta)$. When quantifying the errors, the perturbations do not manifest themselves in any way since the current estimates δ_t of the unknown vector δ usually have the form $\delta_t = (c^w_t, 0, 0)$.

Remark 2. Proponents of stochastic disturbance models in system identification theory constantly criticize the set-membership approach for its seemingly inevitable conservatism due to the necessary a priori information about upper bounds on deterministic disturbances. (Here, only the conservatism of the set estimators of unknown parameters is implied.) As illustrated by Example 1, the use of set-membership estimation and the control criterion as the identification criterion makes the measurement-consistent performance guarantees non-conservative and, furthermore, improves performance guarantees compared to the optimal control criterion (5), since particular disturbance realizations are generally far from the disturbances maximizing the tracking error. This is analogous to the fact that in the stochastic case, average performance indices are better than the worst possible values on particular "bad" realizations. However, in problems with stochastic disturbances, disturbance model verification is usually not discussed. The optimality of tracking within the deterministic ℓ_1-theory is based on the verification of the disturbance model and the use of sufficiently complete information about unknown parameters obtained in the control process, and the price for optimality is a corresponding increase in the volume of necessary computations.

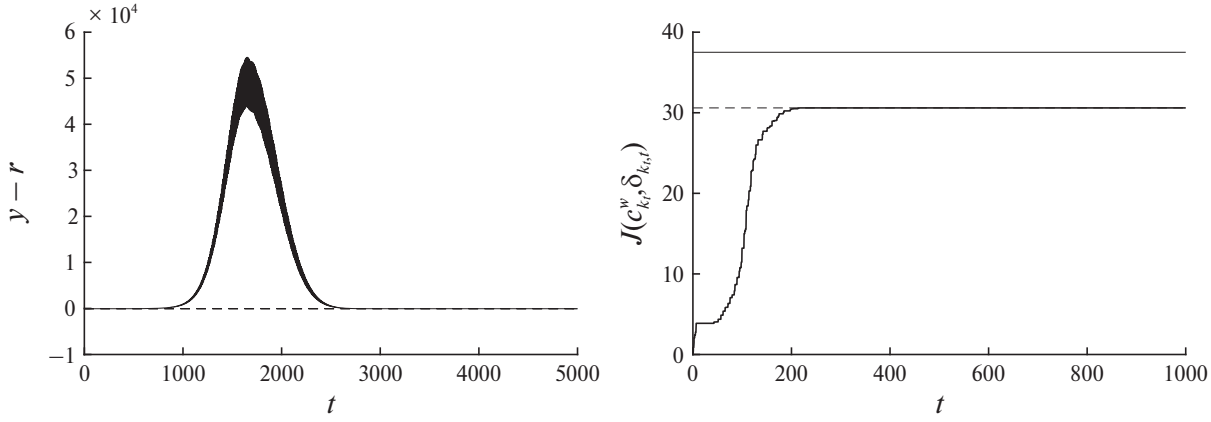

Fig. 3. The graphs of $y-r$ (left) and $J(c^w_{k_t}, \delta_{k_t,t})$ (right). Solid line corresponds to $J(c^w, \delta)$ and dashed line to $J(c^w_{k_\infty}, \delta_{k_\infty})$.

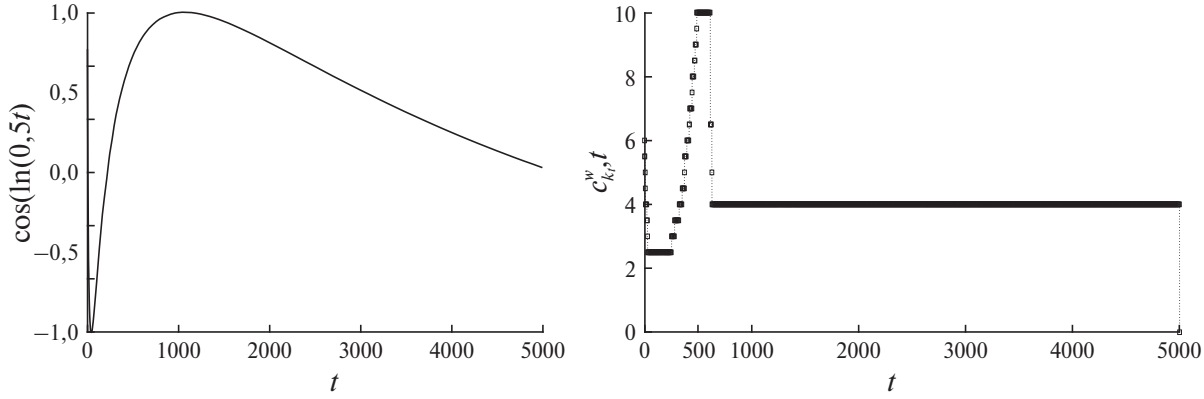

Fig. 4. The graphs of k^y_t and switching of the estimates of the unknown bias $c^w = 5$.

Example 2. "Bad" deterministic disturbances. This example is intended to demonstrate a "bad" total disturbance under which the presence of perturbations in the total disturbance v becomes evident.

Consider the plant (33) with the total disturbance (35) and the parameters (36) with the reduced value $\delta^u = 0.05$ and the deterministic sequences

$$w_t = \cos(50t),\ k^y_t = \sin(70t),\ k^u_t = \cos(\ln(0.5t)). \tag{37}$$

The left graph in Fig. 3 shows the tracking error of $y - r$. Under this disturbance, for all $t \geqslant 498$, the estimates are $\delta^u_{k_t,t} > 0$ and $\delta^y_{k_t,t} = 0$; the last estimates are $\delta_{k_{5000},5000} = (1.6993; 0; 0.0581)$ and $c^w_{5000} = 4$. Thus, starting from the time instant $t = 498$, the perturbation affecting the control manifests itself in the estimates $\delta_{k_t,t}$ of the disturbance norms.

As is known, stable linear time-invariant systems may have large deviations from zero due to nontrivial initial conditions or a bounded disturbance [16, 17]. In Example 2, a large deviation of the tracking error (with $\max_t |y_t - r_t| = 5.4573 \times 10^4$) in the nonlinear closed-loop system (1)–(6) can be caused both by the switching of the controllers corresponding to different estimates of the biases and by the "asymmetry" of the sequence $k^u_t = \cos(\ln(0.5t))$ about zero (see the left graph in Fig. 4). The current optimal estimates $c^w_{k_t,t}$ of the unknown bias c^w are shown in the right graph of Fig. 4, where the steady-state bias estimate is $c^w_{k_\infty} = 4 \neq c^w = 5$.

Remark 3. Despite the tracking error values in the transient mode having the order of magnitude 10^4 (unacceptable in applications), the asymptotic behavior of the tracking error under this disturbance is characterized by the numbers

$$\max_{t\in[4001,5000]}|y_t - r_t| = 4.4163, \qquad \max_{t\in[4901,5000]}|y_t - r_t| = 2.7566.$$

Thus, the factual steady-state tracking error is by an order of magnitude smaller than the final guaranteed tracking error estimate $J(c^w_{k_\infty}, \delta_{k_\infty}) = 30.4421$, consistent with the measurement data. In turn, this estimate is better than the optimal (but unknown!) value $J(c^w, \delta) = 37.2971$ guaranteed by Theorem 1, despite the low chosen "accuracy" of the bias estimates (the grid size $\varepsilon_2 = 0.5$). Finally, the optimal value $J(c^w, \delta)$ itself is less than the worst possible asymptotic tracking error since $J(c^w, \delta)$ ignores that the tested plant has the nominal parameters θ^* instead of θ. As a result, the adaptive compensation algorithm for the unknown bias c^w fulfills its purpose despite possible large deviations of the tracking error from zero and even "adapts" to particular realizations of the disturbance v, reducing excessive conservatism in the guaranteed performance estimates under non-"maximizing" disturbances.

Remark 4. According to the above graphs of the switching of the estimates $c^w_{k_t}$, the unknown bias c^w is non-identifiable in the description (4) even in the absence of perturbations since control always deals with particular realizations of the disturbance v_t for which the biases (for any reasonable definition) can be (more correctly, will be) different. That is, the term "bias" with respect to the constant c^w in the description (4) refers precisely to the concept of bias for the class of all disturbances satisfying (4). At the same time, the current estimates $c^w_{k_t}$ can (or rather should) be considered a correct definition (in the context of the control problem being solved) of the current estimates of the bias for a particular realization of the total disturbance v.

Remark 5. The volume and speed of computations in the above examples are characterized by the following indicators. The computation time on a laptop with 15.2 GB RAM and an Intel Core Ultra 5 125H processor is 2.99 s in Example 1 and 15.1169 s in Example 2. The number of inequalities in the polyhedral estimators $P_{k,t}$ is 12–15 in Example 1 and 64–81 in Example 2. The ratio of these limits approximately matches that of the interval lengths, equal to 5. The indicators of Example 2 on the time interval [1, 10 000] remain the same, meaning that the transient processes for the particular disturbance v under consideration have already been completed by the time instant $t = 5000$. Note that the computation time is determined mainly by the time to calculate the polyhedral estimators $P_{k,t}$ and the optimal vector estimators $\delta_{k,t}$ in $\mathbb{R}^3$ and is almost independent of the dimension of the nominal parameter vector θ.

Remark 6. The number of inequalities in the description of the polyhedral estimators $P_{k,t}$ and, as a consequence, the computation time of the optimal estimates in (21) can be reduced by eliminating possible redundant inequalities after adding the new inequalities (17); for details, see [18]. In the simulation results presented, this was not done in order to demonstrate the number of possible estimator updates even for a very small value of the dead zone parameter ε_1 (almost zero from the viewpoint of assessing the model quality).

6. CONCLUSIONS

This paper has considered a discrete minimum-phase plant with a known or specified nominal model (for testing), a biased and bounded external disturbance, and perturbations with unknown norms and an unknown bias. For this plant, the optimal tracking problem of a given bounded signal with a given accuracy has been addressed. The problem difficulty lies in the need to compensate for the bias based on reasonable optimal estimation of control performance under the

non-identifiability of all unknown parameters. The solution of this problem involves errors quantification, set-membership estimation of unknown parameters, and the use of the control criterion as an ideal identification criterion. Within such an approach, it becomes possible to more deeply understand, demonstrate, and implement the maximum capabilities of feedback control. The importance of feedback research was emphasized by L. Guo, a leading expert in adaptive control and identification of systems, in the abstract of his paper [19]:

> "The main purpose of adaptive feedback is to deal with dynamical systems with internal and/or external uncertainties, by using the on-line observed information. Thus, a fundamental problem in adaptive control is to understand the maximum capability and limits of adaptive feedback."

In this context, we also provide a quotation from the abstract of his another paper [20]:

> "Finally, we will consider more fundamental problems on the maximum capability and limitations of the feedback mechanism in dealing with uncertain nonlinear systems, where the feedback mechanism is defined as the class of all possible feedback laws."

The solution presented in this paper not only ensures, with a given accuracy, the same tracking performance estimate as under the known parameters of the nominal plant and disturbances, but also gives significantly better guaranteed performance estimates depending on particular realizations of deterministic disturbances. Thus, it is implicitly considered that particular realizations of disturbances are usually far from those maximizing the control criterion: in order to maximize the tracking error estimate, the total disturbance v_t must not only take maximum values on a long time interval but also have definite signs on this interval.

FUNDING

This work was supported by a state order of Federal Research Center Komi Scientific Center, the Ural Branch of the Russian Academy of Sciences, project no. 125031203621-2.

REFERENCES

1. Khammash, M. and Pearson, J.B., Performance Robustness of Discrete-Time Systems with Structured Uncertainty, *IEEE Trans. Automat. Control*, 1991, vol. AC-36, no. 4, pp. 398–412.
2. Khammash, M. and Pearson, J.B., Analysis and Design for Robust Performance with Structured Uncertainty, *Syst. Control Lett.*, 1993, vol. 20, no. 3, pp. 179–187.
3. Special Issue on System Identification for Robust Control Design, Kosut, R., Goodwin, G., and Polis, M., Eds., *IEEE Trans. Automat. Control*, 1992, vol. 37, no. 7.
4. Soderstrom, T. and Astrom, K., Eds., Trends in System Identification (Special Issue on System Identification), *Automatica*, 1995, vol. 31, no. 12, pp. 1689–1907.
5. Gevers, M., A Personal View of the Development of System Identification, *IEEE Control Syst. Magazine*, 2006, vol. 26, no. 6, pp. 93–105.
6. FLamnabhi-Lagarrigue, F., Annaswamy, A., Engell, S., et. al., Systems & Control for the Future of Humanity, Research Agenda: Current and Future Roles, Impact and Grand Challenges, *Annual Reviews in Control*, 2017, vol. 43, pp. 1–64.
7. Ljung, L., Revisiting Total Model Errors and Model Validation, *J. Syst. Sci. Complex*, 2021, vol. 34, pp. 1598–1603.
8. Khammash, M.H., Robust Steady-State Tracking, *IEEE Trans. Automat. Control*, 1995, vol. 40, no. 11, pp. 1872–1880.
9. Khammash, M.H., Robust Performance: Unknown Disturbances and Known Fixed Inputs, *IEEE Trans. Automat. Control*, 1997, vol. 42, no. 12, pp. 1730–1734.

10. Sokolov, V.F., Asymptotic Robust Performance of the Discrete Tracking System in the ℓ_1-metric, *Autom. Remote Control*, 1999, vol. 60, no. 1, part 2, pp. 82–91.

11. Sokolov, V.F., *Robastnoe upravlenie pri ogranichennykh vozmushcheniyakh* (Robust Control under Bounded Disturbances), Syktyvkar: Komi Research Center, the Ural Branch of the Russian Academy of Sciences, 2011.

12. Sokolov, V.F., Control-Oriented Model Validation and Errors Quantification in the ℓ_1 Setup, *IEEE Trans. Autom. Control*, 2005, vol. 50, no. 10, pp. 1501–1509.

13. Sokolov, V.F., Problems of Adaptive Optimal Control of Discrete-Time Systems under Bounded Disturbance and Linear Performance Indexes, *Autom. Remote Control*, 2018, vol. 79, no. 6, pp. 1086–1099.

14. Sokolov, V.F., Optimal Error Quantification and Robust Tracking under Unknown Upper Bounds on Uncertainties and Biased External Disturbance, *Mathematics*, 2024, vol. 12(2), art. no. 197.

15. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge: Cambridge University Press, 2003.

16. Polyak, B.T., Shcherbakov, P.S., and Smirnov, G., Peak Effects in Stable Linear Difference Equations, *J. Difference Equat. Appl.*, 2018, vol. 24, no. 9, pp. 1488–1502.

17. Polyak, B.T., Tremba, A.A., Khlebnikov, M.V., Shcherbakov, P.S., and Smirnov, G.V., Large Deviations in Linear Control Systems with Nonzero Initial Conditions, *Autom. Remote Control*, 2015, vol. 76, no. 6, pp. 957–976.

18. Walter, E., Piet-Lahanier, H., Exact Recursive Polyhedral Description of the Feasible Parameter Set for Bounded Error, *IEEE Trans. Automat. Control*, 1989, vol. 34, pp. 911–915.

19. Guo, L., Exploring the Maximum Capability of Adaptive Feedback, *Int. J. Adapt. Control Signal Process.*, 2002, vol. 16, no. 5, pp. 341–354.

20. Guo, L., Feedback and Uncertainty: Some Basic Problems and Results, *Annual Reviews in Control*, 2020, vol. 49, pp. 27–36.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 769–781.

SPECIAL ISSUE

On Optimal Control Problems with Control in a Disc

R. Hildebrand[*,a] **and T. Chikake Mapungwana**[*,b]

[*] *Moscow Institute of Physics and Technology, Dolgoprudny, Russia*
e-mail: [a] *khildebrand.r@mipt.ru,* [b] *tendaichikake@phystech.edu*

Received March 3, 2025
Revised May 20, 2025
Accepted June 27, 2025

Abstract—We consider a time-optimal control problem with Fuller-type symmetry and with control in the 2-dimensional unit disk. The problem can be solved analytically, with an implicit representation of the Bellman function. The optimal value of this problem serves as an upper bound on the optimal value of another optimal control problem with Fuller-type symmetry and with a second-order singular trajectory, which cannot be solved analytically.

Keywords: optimal control, Fuller-type symmetry, singular trajectories, time-optimal control

DOI: 10.31857/S0005117925080067

1. INTRODUCTION

Optimal control problems are usually solved by means of the Pontryagin Maximum Principle (PMP), which leads to a Hamiltonian system with discontinuous right-hand side [7]. Since the Lipschitz-condition, as a consequence of the discontinuity, does not hold, the theorem of existence and uniqueness of solutions to the Ordinary Differential Equation (ODE) does not hold, and the phase portrait can exhibit various kinds of singularities, which appear in the optimal synthesis as singular trajectories. Since such singular trajectories often arise as part of the optimal synthesis, it is necessary to have a good understanding of these singularities. To this end a model problem is considered which on the one hand, is easy enough to be solved, in the best case analytically, and on the other hand, exhibits the kind of singularity under study.

The phenomenon of singular trajectories in optimal control was first discovered in [1], where an example of a system was exhibited where the optimal control performs an infinite number of switchings in finite time, so-called *chattering*. Singular trajectories were studied systematically in, e.g., [2–4, 6]. The phenomenon of chattering was studied in detail in [5, 8]. A more complicated system exhibiting chattering combined with a fractal optimal control pattern was investigated in [9, 10].

In this work we study singular trajectories of second order, continuing the research program initiated in the seminal work [8]. The first optimal control problem with a second-order singular trajectory has been solved in [1]. It is given by the formulation

$$\min \int_0^\infty \frac{x^2}{2}\,dt : \qquad \dot{x} = y, \quad \dot{y} = u, \quad u \in [-1, 1]. \tag{1}$$

The second-order singular trajectory present in the phase portrait of the corresponding Hamiltonian system is the trajectory $x(t) = y(t) \equiv 0$. A junction of a generic optimal trajectory of problem (1) with the singular one is performed in finite time with an infinite number of switchings of the control u between the extremal values ± 1. The Bellman function of problem (1) has been explicitly

computed in [9] and is given by

$$\omega_{1D}(x,y)=\begin{cases} -\frac{1}{2}x^2y-\frac{1}{3}xy^3-\frac{1}{15}y^5-\gamma(y^2+2x)^{\frac{5}{2}}, & x\geqslant -\beta y|y|;\\ \frac{1}{2}x^2y-\frac{1}{3}xy^3+\frac{1}{15}y^5-\gamma(y^2-2x)^{\frac{5}{2}}, & x\leqslant -\beta y|y|, \end{cases} \tag{2}$$

where $\beta\approx 0.4446$ solves the equation

$$36\beta^4+3\beta^2-2=0$$

and $\gamma=\frac{-\beta^2+2\beta-\frac{2}{3}}{10(1-2\beta)^{\frac{3}{2}}}\approx 0.06753$. Recall that the expression $-\omega_{1D}(x_0,y_0)$ is the optimal value of the problem with initial data $x(0)=x_0$, $y(0)=y_0$.

The optimal synthesis of the problem exhibits a continuous symmetry. Namely, if $(x(t),y(t),u(t))$ is an optimal solution of problem (1), then for every $\lambda>0$ the trajectory

$$(x_\lambda(t),y_\lambda(t),u_\lambda(t))=(\lambda^2x(\lambda^{-1}t),\lambda y(\lambda^{-1}t),u(\lambda^{-1}t)) \tag{3}$$

is also optimal [8]. Similarly, for every $\lambda>0$ the Bellman function obeys the relation

$$\omega_{1D}(\lambda^2x,\lambda y)=\lambda^5\omega_{1D}(x,y). \tag{4}$$

In [8] an analog of problem (1) with two-dimensional control was considered, namely

$$\min\int_0^\infty \frac{\|x\|^2}{2}\,dt: \qquad \dot x=y,\quad \dot y=u,\quad u\in U=\mathbb{D}, \tag{5}$$

where $\mathbb{D}=\{u\in\mathbb{R}^2\mid \|u\|\leqslant 1\}$ is the unit disk. Here $x(t)$, $y(t)$ are vector-valued functions. This problem also features a second-order singular trajectory at the origin $(x,y)=(0,0)$ of the space $\mathbb{R}^2\times\mathbb{R}^2$. It exhibits the same symmetry (3), but also an additional rotational symmetry. Namely, for every optimal solution $(x(t),y(t),u(t))$ of problem (5) and every orthogonal matrix $O\in O(2)$ the trajectory

$$(Ox(t),Oy(t),Ou(t)) \tag{6}$$

is also optimal. The Bellman function ω_{2D} of problem (5) satisfies the symmetries

$$\omega_{2D}(\lambda^2x,\lambda y)=\lambda^5\omega_{2D}(x,y),\quad \omega_{2D}(Ox,Oy)=\omega_{2D}(x,y)$$

for every $\lambda>0$ and every $O\in O(2)$. The rotational symmetry implies that the dynamics of the optimal synthesis factors through to the Gramian $\begin{pmatrix}\langle x,x\rangle & \langle x,y\rangle\\ \langle x,y\rangle & \langle y,y\rangle\end{pmatrix}$ of the vectors x,y. The value at time t of this 2×2 matrix is determined solely by the initial value of the Gramian itself.

It was shown in [8, Proposition 7.8] that for linearly dependent initial vectors $x(0),y(0)$ problem (5) reduces to problem (1), and the vectors $x(t),y(t)$ stay in the 1-dimensional subspace spanned by the initial vectors for all time. In particular, for linearly dependent vectors $x=r_x\cdot(\cos\varphi,\sin\varphi)^T$, $y=r_y\cdot(\cos\varphi,\sin\varphi)^T$ the Bellman function satisfies

$$\omega_{2D}(x,y)=\omega_{1D}(r_x,r_y),$$

where r_x,r_y are allowed to take arbitrary real values.

Besides the optimal trajectories of problem (5) emanating from linearly dependent initial values, a family of *self-similar* optimal trajectories has been computed in [8, Proposition 7.9, Corollary 7.3].

For these trajectories, the tangent of the angle between the vectors x and y is constant and equals $-\sqrt{5}/2$ (the values $\sqrt{5}$ and $\sqrt{5/2}$ in [8, p.233] are both erroneous), and between the vectors y and u it is $-\sqrt{5}$. Moreover, $2\langle y, y\rangle = \sqrt{6\langle x, x\rangle}$. It follows that the Gramian of the initial values $x(0), y(0)$ is given by

$$\begin{pmatrix} \langle x(0), x(0)\rangle & \langle x(0), y(0)\rangle \\ \langle x(0), y(0)\rangle & \langle y(0), y(0)\rangle \end{pmatrix} = \begin{pmatrix} \frac{\lambda_0^4}{54} & -\frac{\lambda_0^3}{27} \\ -\frac{\lambda_0^3}{27} & \frac{\lambda_0^2}{6} \end{pmatrix} \tag{7}$$

for some $\lambda_0 > 0$. The Gramian of the corresponding trajectory evolves according to the formula

$$\begin{pmatrix} \langle x(t), x(t)\rangle & \langle x(t), y(t)\rangle \\ \langle x(t), y(t)\rangle & \langle y(t), y(t)\rangle \end{pmatrix} = \begin{pmatrix} \frac{\lambda(t)^4}{54} & -\frac{\lambda(t)^3}{27} \\ -\frac{\lambda(t)^3}{27} & \frac{\lambda(t)^2}{6} \end{pmatrix}, \quad \lambda(t) = \lambda_0 - t. \tag{8}$$

It follows that the parameter λ_0 is the arrival time at the singular trajectory, which is located at the origin $(x, y) = (0, 0)$.

While the angles between the vectors x, y, u remain constant, the vectors themselves revolve ever more rapidly around the origin, making an infinite number of revolutions in finite time. More concretely, the direction each vector is pointing is given by the time-varying angle [8, Proposition 7.9, Corollary 7.3]

$$\varphi(t) = \sigma\sqrt{5}\log(\lambda_0 - t) + \text{const}, \tag{9}$$

where $\sigma \in \{-1, +1\}$ determines the direction of revolution and the additive constant on the initial conditions.

The complete optimal synthesis of problem (5) is currently unknown. In this paper we compute an upper bound on the objective value by constructing a sub-optimal solution. This upper bound has to satisfy several, potentially conflicting criteria:

- the bound should be reasonable close to the true value
- the bound should be efficiently usable numerically, e.g., given by a global analytic expression

As we have seen from the analysis above, the optimal solutions for different initial values can be quite different from each other. Constructing a bound which is everywhere close to the optimal value and at the same time not given by a multitude of different expressions for different phase space regions is a challenging task.

We cope with this difficulty by solving the *time-optimal* control problem

$$\min (T - t_0): \quad \dot{x} = y, \quad \dot{y} = u, \quad u \in U = \mathbb{D}, \quad x(T) = y(T) = 0, \quad x(t_0) = x_0, \quad y(t_0) = y_0, \tag{10}$$

which up to a shift of the time variable t has the same feasible set of trajectories as problem (5) but another objective value. This is accomplished in Section 2. In Section 3 we substitute the obtained time-optimal solution in the objective value of the original problem (5) to obtain the upper bound. Finally, in Section 4 we compare the upper bound with the optimal value of problem (5) on those trajectories where the latter is known. It turns out that, on the one hand, the quality of the bound is reasonably good on all initial values for which the optimal solution of problem (5) is known, and on the other hand it is given by a single analytic expression.

2. TIME-OPTIMAL CONTROL PROBLEM

In this section we analytically solve the time-optimal control problem (10).

Let us apply the PMP. Introduce adjoint variables ϕ, ψ and assemble the Pontryagin function

$$\mathcal{H} = -1 + \langle \phi, y \rangle + \langle \psi, u \rangle. \tag{11}$$

The optimal control is then given by $\hat{u} = \arg\max_{u \in U} \mathcal{H} = \frac{\psi}{\|\psi\|}$ whenever $\psi \neq 0$. The dynamics of the adjoint variables is given by

$$\dot{\psi} = -\frac{\partial \mathcal{H}}{\partial y} = -\phi, \quad \dot{\phi} = -\frac{\partial \mathcal{H}}{\partial x} = 0.$$

Since the terminal time instant T is not fixed, we also have the transversality condition

$$\mathcal{H}(T) = \langle \psi, \hat{u} \rangle - 1 = \|\psi\| - 1 = 0. \tag{12}$$

Hence the function $\psi(t) = \phi t + \psi(0)$ is affine and at $t = T$ terminates on the unit circle.

By virtue of the rotational symmetry, without loss of generality we may assume that $\dot{\psi} = -\phi = (\alpha, 0)^T$ is collinear with the unit basis vector $e_1 = (1, 0)^T$ and $\alpha \geqslant 0$. In the case $\alpha > 0$ we shall choose the initial value t_0 of the time variable such that $\psi(0) = (0, \beta)^T$ is collinear with e_2, and $\beta \geqslant 0$.

Case $\alpha\beta = 0$: In this case the adjoint variable ψ evolves in a 1-dimensional linear subspace of $\mathbb{R}^2$. Hence also u, y, x have to evolve in this subspace, and the problem reduces to the well-known 1-dimensional time-optimal control problem with acceleration bounded by 1, which is given by (10) with all variables considered as scalars.

In this case the adjoint variable ϕ is a scalar constant. Let us shift the time variable such that the final time T is zero. Then the adjoint variable ψ has terminal value $\psi(T) = \psi(0) = \sigma \in \{-1, +1\}$ and is given by $\psi(t) = \sigma - \phi t$. The optimal control u is piece-wise constant, given by

$$\hat{u}(t) = \begin{cases} +1, & \text{if } \ \sigma > \phi t, \\ -1, & \text{if } \ \sigma < \phi t. \end{cases}$$

Equivalently, with $\tilde{\phi} = \sigma\phi$ we get

$$\hat{u}(t) = \begin{cases} +\sigma, & \text{if } \ 1 > \tilde{\phi} t, \\ -\sigma, & \text{if } \ 1 < \tilde{\phi} t. \end{cases}$$

This yields

$$y(t) = \int_0^t \hat{u}(s)\, ds = \begin{cases} \sigma t, & \text{if } \ 1 > \tilde{\phi} t, \\ -\sigma(t - \tilde{\phi}^{-1}) + \sigma\tilde{\phi}^{-1}, & \text{if } \ 1 < \tilde{\phi} t \end{cases}$$

and further

$$x(t) = \int_0^t y(s)\, ds = \begin{cases} \sigma \dfrac{t^2}{2}, & \text{if } \ 1 > \tilde{\phi} t, \\ -\sigma \dfrac{t^2 - \tilde{\phi}^{-2}}{2} + 2\sigma\tilde{\phi}^{-1}(t - \tilde{\phi}^{-1}) + \sigma \dfrac{\tilde{\phi}^{-2}}{2}, & \text{if } \ 1 < \tilde{\phi} t \end{cases}$$

The case $1 > \tilde{\phi} t$ is hence possible only if $x = \sigma \frac{y^2}{2} = -\frac{y|y|}{2}$, in which case $t_0 = \sigma y_0$. For any other point in the x, y plane we must have $1 < \tilde{\phi} t$. Hence the locus of the equation $x + \frac{y|y|}{2} = 0$ separates the x, y plane in two regions, where different values of the control must be optimal. The separating curve consists of the two trajectories which arrive directly at the origin (see Fig. 1).

In particular, for $1 < \tilde{\phi} t$ we have

$$\hat{u} = -\sigma = -\operatorname{sgn}\left(x_0 + \frac{y_0|y_0|}{2}\right).$$

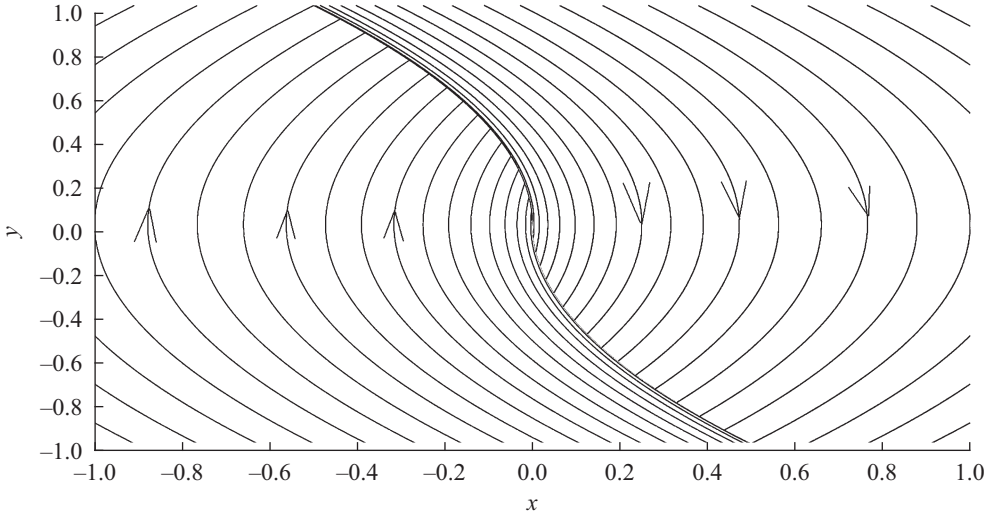

Fig. 1. Optimal synthesis for the 1-dimensional time-optimal control problem. To the left of the curve separating the two regions we have $\hat{u} = +1$ and $\sigma = -1$, to the right $\hat{u} = -1$ and $\sigma = 1$. The trajectories of the system switch to the opposite control when they arrive at the separating curve.

Further, $\sigma y_0 = -t_0 + 2\tilde{\phi}^{-1}$, $\sigma x_0 = -\frac{t_0^2}{2} + 2\tilde{\phi}^{-1}t_0 - \tilde{\phi}^{-2}$ and hence

$$\tilde{\phi}^{-1} = -\sqrt{\sigma x_0 + \frac{y_0^2}{2}}, \quad t_0 = -\sigma y_0 - 2\sqrt{\sigma x_0 + \frac{y_0^2}{2}}.$$

For $t_0 < t < -\sqrt{\sigma x_0 + \frac{y_0^2}{2}}$ we thus finally obtain

$$y(t) = -\sigma t - 2\sigma\sqrt{\sigma x_0 + \frac{y_0^2}{2}}, \quad x(t) = -\frac{\sigma t^2}{2} - 2\sigma\sqrt{\sigma x_0 + \frac{y_0^2}{2}}\,t - x_0 - \frac{\sigma y_0^2}{2}$$

with $\sigma = \operatorname{sgn}\left(x_0 + \frac{y_0|y_0|}{2}\right)$.

Case $\alpha > 0$, $\beta > 0$: In this case $\psi(t) = (\alpha t, \beta)^T$. Since $\|\psi(T)\| = 1$, we must have $\beta \leqslant 1$ and $\alpha^2 T^2 + \beta^2 = 1$.

Introduce the scaled time variable $\tau = \alpha t$ and the scaled starting point $\tau_0 = \alpha t_0$ and end-point $\bar{\tau} = \alpha T$. Then we get $\beta = \sqrt{1 - \bar{\tau}^2}$. Consequently, $\|\psi(t)\| = \sqrt{\tau^2 + 1 - \bar{\tau}^2}$. The optimal control is then given by

$$\hat{u} = \frac{\psi}{\|\psi\|} = \frac{(\tau, \beta)^T}{\sqrt{\tau^2 + \beta^2}}. \tag{13}$$

Since the system is autonomous, the Pontryagin function (11) is constant along the trajectory, and by the transversality condition (12) we get the energy integral $\mathcal{H} = -1 - \alpha y_1 + \sqrt{\tau^2 + \beta^2} \equiv 0$. Hence

$$y_1 = \frac{\sqrt{\tau^2 + \beta^2} - 1}{\alpha}. \tag{14}$$

For the other component of $y(t)$ we get the solution

$$y_2 = \int_T^t \frac{\beta}{\sqrt{\alpha^2 s^2 + \beta^2}}\, ds = \int_{\bar{\tau}}^{\tau} \frac{\beta}{\sqrt{s^2 + \beta^2}}\, \frac{ds}{\alpha} = \frac{\beta}{\alpha}\left(\operatorname{arsinh}\frac{\tau}{\beta} - \operatorname{artanh}\bar{\tau}\right). \tag{15}$$

Here we used that $\operatorname{arsinh}\frac{\bar{\tau}}{\beta} = \operatorname{arsinh}\frac{\bar{\tau}}{\sqrt{1-\bar{\tau}^2}} = \operatorname{artanh}\bar{\tau}$.

Integrating further, we obtain for the function $x(t) = \int_T^t y(s)\,ds = \int_\tau^{\bar\tau} y(s/\alpha)\,\frac{ds}{\alpha}$ that

$$x_1 = \frac{1}{2\alpha^2}\tau\sqrt{\tau^2+\beta^2} + \frac{\beta^2}{2\alpha^2}\operatorname{arsinh}\frac{\tau}{\beta} - \frac{\tau}{\alpha^2} - \frac{1}{2\alpha^2}(\beta^2\operatorname{artanh}\bar\tau - \bar\tau),$$
$$x_2 = \frac{\beta\tau}{\alpha^2}\operatorname{arsinh}\frac{\tau}{\beta} - \frac{\beta}{\alpha^2}\sqrt{\tau^2+\beta^2} - \frac{\beta\tau}{\alpha^2}\operatorname{artanh}\bar\tau + \frac{\beta}{\alpha^2}.$$

Let us compute the elements of the Gramian. After insertion into the scalar products and simplification we get

$$\begin{aligned}
\alpha^4\|x\|^2 &= \frac{1}{4}\tau^4 + \left(\beta^2\operatorname{artanh}^2\bar\tau + \frac{5\beta^2}{4} + 1\right)\tau^2 - \left(\beta^2\operatorname{artanh}\bar\tau + \bar\tau\right)\tau \\
&\quad * + \frac{1}{4}\left(4\beta^4 + 3\beta^2 + 1 + \beta^4\operatorname{artanh}^2\bar\tau - 2\beta^2\bar\tau\operatorname{artanh}\bar\tau\right) \\
&\quad * - \tau^2\sqrt{\tau^2+\beta^2} + \frac{1}{2}\left(3\beta^2\operatorname{artanh}\bar\tau + \bar\tau\right)\tau\sqrt{\tau^2+\beta^2} - 2\beta^2\sqrt{\tau^2+\beta^2} \\
&\quad * - 2\beta^2\tau^2\operatorname{artanh}\bar\tau\operatorname{arsinh}\frac{\tau}{\beta} + \beta^2\tau\operatorname{arsinh}\frac{\tau}{\beta} - \frac{\beta^2}{2}\left(\beta^2\operatorname{artanh}\bar\tau - \bar\tau\right)\operatorname{arsinh}\frac{\tau}{\beta} \\
&\quad * - \frac{3\beta^2}{2}\tau\sqrt{\tau^2+\beta^2}\operatorname{arsinh}\frac{\tau}{\beta} + \beta^2\tau^2\operatorname{arsinh}^2\frac{\tau}{\beta} + \frac{\beta^4}{4}\operatorname{arsinh}^2\frac{\tau}{\beta}, \qquad (16)
\end{aligned}$$

$$\begin{aligned}
*\alpha^3\langle x,y\rangle &= \frac{1}{2}\tau^3 + \left(\beta^2\operatorname{artanh}^2\bar\tau + \frac{\beta^2}{2} + 1\right)\tau - \frac{\beta^2}{2}\sqrt{\tau^2+\beta^2}\operatorname{arsinh}\frac{\tau}{\beta} \\
&\quad * + \frac{1}{2}\left(\beta^2\operatorname{artanh}\bar\tau + \bar\tau\right)\sqrt{\tau^2+\beta^2} - \frac{3}{2}\tau\sqrt{\tau^2+\beta^2} + \frac{\beta^2}{2}\operatorname{arsinh}\frac{\tau}{\beta} - \frac{1}{2}\left(\beta^2\operatorname{artanh}\bar\tau + \bar\tau\right) \\
&\quad * + \beta^2\tau\operatorname{arsinh}^2\frac{\tau}{\beta} - 2\beta^2\tau\operatorname{artanh}\bar\tau\operatorname{arsinh}\frac{\tau}{\beta}, \qquad (17)
\end{aligned}$$

$$*\alpha^2\|y\|^2 = \tau^2 - 2\sqrt{\tau^2+\beta^2} + \beta^2\operatorname{arsinh}^2\frac{\tau}{\beta} - 2\beta^2\operatorname{artanh}\bar\tau\operatorname{arsinh}\frac{\tau}{\beta} + \left(\beta^2\operatorname{artanh}^2\bar\tau + \beta^2 + 1\right). \qquad (18)$$

2.1. Computation of the Parameters $\alpha, \bar\tau, \tau_0$

In order to compute the time-optimal trajectory for a given initial value of the Gramian $\begin{pmatrix} \|x_0\|^2 & \langle x_0,y_0\rangle \\ \langle x_0,y_0\rangle & \|y_0\|^2 \end{pmatrix}$, we have to invert the above dependence to obtain the values $\alpha, \bar\tau, \tau_0$.

The dependence on α is algebraic. Multiplication of α by a constant λ multiplies the Gramian from the left and from the right by the diagonal matrix $\operatorname{diag}(\lambda^{-2}, \lambda^{-1})$. In the cone $\mathcal{S}_+^2$ of positive semi-definite 2×2 matrices this action defines 1-dimensional orbits of radial type, each of which intersects every affine compact non-zero section of the cone $\mathcal{S}_+^2$ in exactly one point. The orbit itself then depends only on the parameters $\tau_0, \bar\tau$. It can be represented, e.g., by the two ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$.

The dependence of the orbit on the parameters cannot be inverted in closed form. In order to shed light on it, let us compute the limit when the parameters $\tau_0, \bar\tau$ tend to the boundary of their domain of definition. Recall that $-1 < \bar\tau < 1$, $\tau_0 < \bar\tau$.

In the limit $\tau_0 \to \bar\tau$ we obtain $\operatorname{arsinh}\frac{\tau_0}{\beta} \to \operatorname{arsinh}\frac{\bar\tau}{\beta} = \operatorname{artanh}\bar\tau$, $\sqrt{\tau_0^2+\beta^2} \to 1$. Inserting with $\tau = \tau_0$ into (16), (17), (18) we obtain that the Gramian of the initial point (x_0, y_0) tends to 0. However, if at the same time $\alpha \to 0$ such that the ratio $\frac{\bar\tau - \tau_0}{\alpha} = T - t_0$ equals 1, then the control

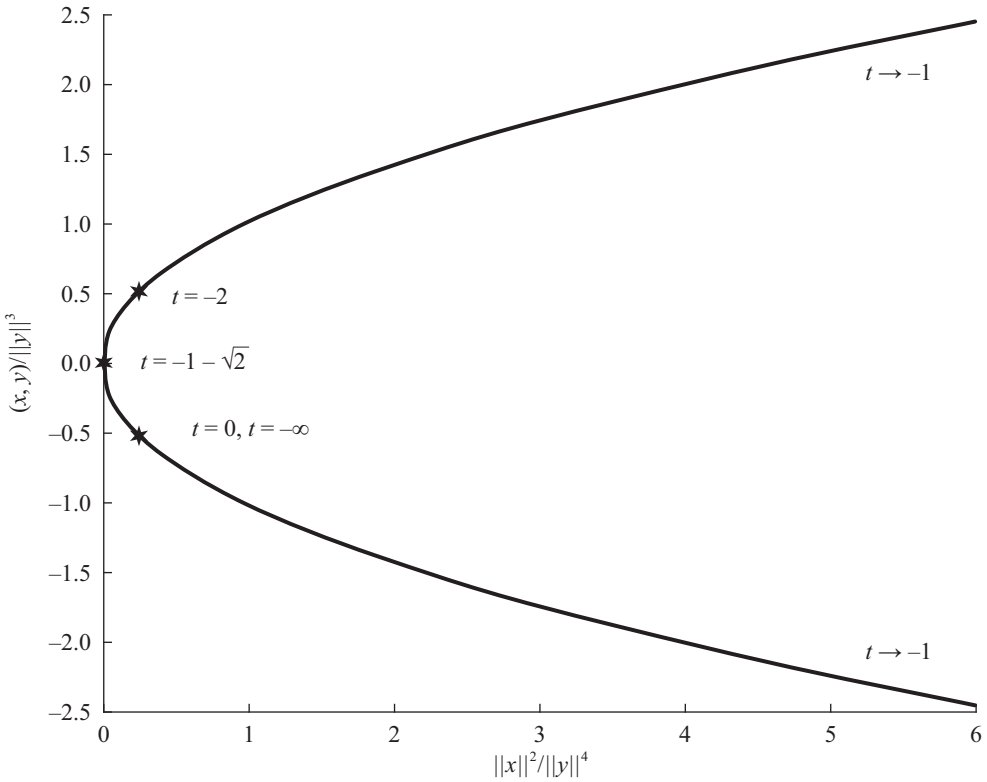

Fig. 2. Limits of the ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$ when $\bar\tau, \tau_0$ tend to the boundary of their domain of definition. The limit is different from $(\frac14, -\frac12)$ only for $\bar\tau \to 1$, $\tau_0 < 0$ and is located on a parabola.

tends to the constant function $\hat u \equiv (\bar\tau, \beta)^T$ and the trajectory tends to a segment of a parabola given by $x(t) = \frac12 (t-T)^2 \hat u$, $y(t) = (t-T)\hat u$. Hence $x_0 \to \frac12(\bar\tau, \beta)^T$, $y_0 \to -(\bar\tau, \beta)^T$, and the Gramian tends to $\begin{pmatrix} \frac14 & -\frac12 \\ -\frac12 & 1 \end{pmatrix}$. In particular, $\frac{\|x_0\|^2}{\|y_0\|^4} \to \frac14$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3} \to -\frac12$.

In the limit $\bar\tau \to \pm 1$ we get $\beta \to 0$, $\beta \operatorname{arsinh} \frac{\tau}{\beta} \to 0$, $\beta \operatorname{artanh} \bar\tau \to 0$, $\sqrt{\tau^2 + \beta^2} \to |\tau|$. Inserting into (16), (17), (18) we get

$$\alpha^4 \|x\|^2 \to \frac14 \tau^4 + \tau^2 + \frac14 - \tau^2|\tau| + \frac12 \bar\tau\tau|\tau| - \bar\tau\tau,$$

$$\alpha^3 \langle x, y\rangle \to \frac12 (|\tau| - 1)\left(\tau(|\tau| - 2) + \bar\tau\right),$$

$$\alpha^2 \|y\|^2 \to (|\tau| - 1)^2.$$

For $\bar\tau \to -1$ we have $\tau \leqslant -1$ and $|\tau| = -\tau$. Setting $\alpha = -1 - \tau_0$ such that again $T - t_0 \to 1$, this yields $\begin{pmatrix} \|x_0\|^2 & \langle x_0, y_0\rangle \\ \langle x_0, y_0\rangle & \|y_0\|^2 \end{pmatrix} \to \begin{pmatrix} \frac14 & -\frac12 \\ -\frac12 & 1 \end{pmatrix}$ and the ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$ tend to the same limits $\frac14, -\frac12$ as above.

For $\bar\tau \to 1$ we obtain

$$\begin{pmatrix} \|x_0\|^2 & \langle x_0, y_0\rangle \\ \langle x_0, y_0\rangle & \|y_0\|^2 \end{pmatrix} \to \begin{pmatrix} \dfrac14 \dfrac{(\tau_0(|\tau_0| - 2) + 1)^2}{\alpha^4} & -\dfrac12 \dfrac{(1 - |\tau_0|)\,(\tau_0(|\tau_0| - 2) + 1)}{\alpha^3} \\ -\dfrac12 \dfrac{(1 - |\tau_0|)\,(\tau_0(|\tau_0| - 2) + 1)}{\alpha^3} & \dfrac{(1 - |\tau_0|)^2}{\alpha^2} \end{pmatrix}$$

and $\frac{\|x_0\|^2}{\|y_0\|^4} \to \frac{1}{4}\frac{(\tau_0(|\tau_0|-2)+1)^2}{(1-|\tau_0|)^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3} \to -\frac{1}{2}\frac{\tau_0(|\tau_0|-2)+1}{|1-|\tau_0||(1-|\tau_0|)}$. If $\tau_0 \geqslant 0$, then these limits are again equal to $\frac{1}{4}, -\frac{1}{2}$. For $\tau_0 \leqslant 0$ they equal $\frac{\left((1+\tau_0)^2-2\right)^2}{4(1+\tau_0)^4}$, $\frac{(1+\tau_0)^2-2}{2|1+\tau_0|(1+\tau_0)}$, and $\lim_{\bar\tau\to 1}\frac{\|x_0\|^2}{\|y_0\|^4} = \left(\lim_{\bar\tau\to 1}\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}\right)^2$.

For $\bar\tau = 1$, $\tau_0 \in (-\infty,-1)$ the ratio $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$ rises monotonely from $-\frac{1}{2}$ to $+\infty$. For $\tau_0 \in (-1,0]$ it rises monotonely from $-\infty$ to $-\frac{1}{2}$. The boundary values of the ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$ are depicted on Fig. 2.

Finally, if $\tau_0 \to -\infty$, then the Gramian grows unbounded. However, if we simultaneously let $\alpha \to +\infty$ such that $\frac{-\tau_0}{\alpha} \to 1$, then the leading terms in τ_0 dominate and again $\begin{pmatrix} \|x_0\|^2 & \langle x_0,y_0\rangle \\ \langle x_0,y_0\rangle & \|y_0\|^2 \end{pmatrix} \to \begin{pmatrix} \frac{1}{4} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$.

Hence if the parameters $\bar\tau, \tau_0$ circumvent the boundary of their domain of definition, the pair $(\frac{\|x_0\|^2}{\|y_0\|^4}, \frac{\langle x_0,y_0\rangle}{\|y_0\|^3})$ moves along the parabola on Fig. 2, including the infinitely far point. Except the interval $(\bar\tau,\tau_0) \in \{1\}\times\mathbb{R}_{--}$ the ratio pair tends to the point $(\frac{1}{4},-\frac{1}{2})$. For $\bar\tau \in (-1,+1)$, $\tau_0 < \bar\tau$ the pair takes values right of the parabola.

The values of $\alpha, \bar\tau, \tau_0$ producing a given Gramian of x_0, y_0 can then be obtained as follows. Compute the ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$. Determine the values of $\tau_0, \bar\tau$ yielding these ratios. This can be done, e.g., by tracing the level lines of the ratios as a function of $\bar\tau, \tau_0$, finding their intersection, and refining the values with the Newton method. In a final step determine α, e.g., from (18).

3. UPPER BOUND ON THE OBJECTIVE VALUE

In this section we compute the objective value of problem (5) on the time-optimal trajectory computed in Section 2. We again consider the two cases from the previous section.

Case $\alpha > 0$, $\beta > 0$: The objective value of the time-optimal trajectory for the original cost function is given by

$$\frac{1}{2}\int_{t_0}^{T}\|x(s)\|^2\,ds = \frac{1}{2}\int_{\tau_0}^{\bar\tau}\|x(\tau/\alpha)\|^2\,\frac{d\tau}{\alpha} = \frac{1}{2\alpha^5}\int_{\tau_0}^{\bar\tau}\alpha^4\|x\|^2\,d\tau.$$

Integrating expression (16) with respect to τ we get

$$\begin{aligned}
&\frac{\tau^5}{20} + \frac{1}{3}\left(\frac{71}{36}\beta^2 + \beta^2\operatorname{artanh}^2\bar\tau + 1\right)\tau^3 - \frac{1}{2}\left(\beta^2\operatorname{artanh}\bar\tau + \bar\tau\right)\tau^2 \\
&+ \frac{1}{36}\left(9\beta^4\operatorname{artanh}^2\bar\tau - 18\beta^2\bar\tau\operatorname{artanh}\bar\tau + 56\beta^4 + 27\beta^2 + 9\right)\tau \\
&- \frac{1}{9}\left(3\beta^2\bar\tau - 5\beta^4\operatorname{artanh}\bar\tau\right)\sqrt{\tau^2+\beta^2} - \frac{11\beta^2}{8}\tau\sqrt{\tau^2+\beta^2} \\
&+ \frac{1}{18}\left(13\beta^2\operatorname{artanh}\bar\tau + 3\bar\tau\right)\tau^2\sqrt{\tau^2+\beta^2} - \frac{1}{4}\tau^3\sqrt{\tau^2+\beta^2} \\
&- \frac{5\beta^4}{8}\operatorname{arsinh}\frac{\tau}{\beta} - \frac{\beta^2}{2}\left(\beta^2\operatorname{artanh}\bar\tau - \bar\tau\right)\tau\operatorname{arsinh}\frac{\tau}{\beta} + \frac{\beta^2}{2}\tau^2\operatorname{arsinh}\frac{\tau}{\beta} - \frac{2\beta^2}{3}\tau^3\operatorname{artanh}\bar\tau\operatorname{arsinh}\frac{\tau}{\beta} \\
&- \frac{5\beta^4}{9}\sqrt{\tau^2+\beta^2}\operatorname{arsinh}\frac{\tau}{\beta} - \frac{13\beta^2}{18}\tau^2\sqrt{\tau^2+\beta^2}\operatorname{arsinh}\frac{\tau}{\beta} + \frac{\beta^4}{4}\tau\operatorname{arsinh}^2\frac{\tau}{\beta} + \frac{\beta^2}{3}\tau^3\operatorname{arsinh}^2\frac{\tau}{\beta}.
\end{aligned}$$

For $\tau = \bar\tau$ this expression evaluates to

$$\frac{1}{1080}\left((1024\beta^4 - 163\beta^2 + 54)\bar\tau - 675\beta^4\operatorname{artanh}\bar\tau\right).$$

Hence the objective value $-\omega^{TO}$ of the time-optimal trajectory obeys

$$\begin{aligned}
-\alpha^5\omega^{TO} = &-\frac{\tau_0^5}{40} - \frac{1}{6}\left(\frac{71}{36}\beta^2 + \beta^2\operatorname{artanh}^2\bar\tau + 1\right)\tau_0^3 + \frac{1}{4}(\beta^2\operatorname{artanh}\bar\tau + \bar\tau)\tau_0^2\\
&- \frac{1}{72}(9\beta^4\operatorname{artanh}^2\bar\tau + 27\beta^2 + 56\beta^4 - 18\beta^2\bar\tau\operatorname{artanh}\bar\tau + 9)\tau_0\\
&+ \frac{1}{2160}\left((1024\beta^4 - 163\beta^2 + 54)\bar\tau - 675\beta^4\operatorname{artanh}\bar\tau\right)\\
&+ \frac{1}{18}(3\beta^2\bar\tau - 5\beta^4\operatorname{artanh}\bar\tau)\sqrt{\tau_0^2+\beta^2} + \frac{11\beta^2}{16}\tau_0\sqrt{\tau_0^2+\beta^2}\\
&- \frac{1}{36}(13\beta^2\operatorname{artanh}\bar\tau + 3\bar\tau)\tau_0^2\sqrt{\tau_0^2+\beta^2} + \frac{1}{8}\tau_0^3\sqrt{\tau_0^2+\beta^2}\\
&+ \frac{5\beta^4}{16}\operatorname{arsinh}\frac{\tau_0}{\beta} + \frac{\beta^2}{4}(\beta^2\operatorname{artanh}\bar\tau - \bar\tau)\tau_0\operatorname{arsinh}\frac{\tau_0}{\beta} - \frac{\beta^2}{4}\tau_0^2\operatorname{arsinh}\frac{\tau_0}{\beta}\\
&+ \frac{\beta^2}{3}\tau_0^3\operatorname{artanh}\bar\tau\operatorname{arsinh}\frac{\tau_0}{\beta} + \frac{5\beta^4}{18}\sqrt{\tau_0^2+\beta^2}\operatorname{arsinh}\frac{\tau_0}{\beta} + \frac{13\beta^2}{36}\tau_0^2\sqrt{\tau_0^2+\beta^2}\operatorname{arsinh}\frac{\tau_0}{\beta}\\
&- \frac{\beta^4}{8}\tau_0\operatorname{arsinh}^2\frac{\tau_0}{\beta} - \frac{\beta^2}{6}\tau_0^3\operatorname{arsinh}^2\frac{\tau_0}{\beta}.
\end{aligned}$$

Case $\alpha\beta = 0$: For the 1-dimensional control problem we have

$$\frac{x(t)^2}{2} = \begin{cases} \dfrac{t^4}{8}, & \text{if } \ t > -\sqrt{\sigma x_0 + \dfrac{y_0^2}{2}},\\[2ex] \dfrac{1}{2}\left(\dfrac{t^2}{2} + 2\sqrt{\sigma x_0 + \dfrac{y_0^2}{2}}\,t + \sigma x_0 + \dfrac{y_0^2}{2}\right)^2, & t < -\sqrt{\sigma x_0 + \dfrac{y_0^2}{2}}. \end{cases}$$

Hence with $\tilde\phi^{-1} = -\sqrt{\sigma x_0 + \frac{y_0^2}{2}}$, $\sigma = \operatorname{sgn}\left(x_0 + \frac{y_0|y_0|}{2}\right)$, and $t_0 = -\sigma y_0 + 2\tilde\phi^{-1}$ the objective value of problem (1) on the time-optimal trajectory is given by

$$\begin{aligned}
\frac{1}{2}\int_{t_0}^{0} x(t)^2\,dt &= \frac{1}{2}\int_{t_0}^{\tilde\phi^{-1}}\left(\frac{t^2}{2} - 2\tilde\phi^{-1}t + \tilde\phi^{-2}\right)^2 dt + \int_{\tilde\phi^{-1}}^{0}\frac{t^4}{8}\,dt\\
&= -\frac{t_0^5}{40} + \frac{t_0^4\tilde\phi^{-1}}{4} - \frac{5}{6}t_0^3\tilde\phi^{-2} + t_0^2\tilde\phi^{-3} - \frac{t_0\tilde\phi^{-4}}{2} + \frac{\tilde\phi^{-5}}{12}\\
&= -\frac{23}{60}\tilde\phi^{-5} + \frac{\tilde\phi^{-4}\sigma y_0}{2} - \frac{\tilde\phi^{-2}\sigma y_0^3}{6} + \frac{\sigma y_0^5}{40}\\
&= \frac{\sigma y_0 x_0^2}{2} + \frac{x_0 y_0^3}{3} + \frac{\sigma y_0^5}{15} + \frac{23}{60}\left(\sigma x_0 + \frac{y_0^2}{2}\right)^{5/2}. \qquad (19)
\end{aligned}$$

4. QUALITY OF APPROXIMATION

In this section we compare the objective value of the constructed sub-optimal solution with the optimal objective value on those trajectories where the latter is known.

Let us first consider the 1-dimensional problem (1). Since both the optimal value (2) (multiplied by -1) and the objective value (19) of the time-optimal trajectory satisfy the symmetry (4), the relative gap between the two values depends only on the ratio $\frac{x}{|y|y}$. This gap is depicted on Fig. 3. It varies between approximately 5.4×10^{-5} and 5.6×10^{-2}.

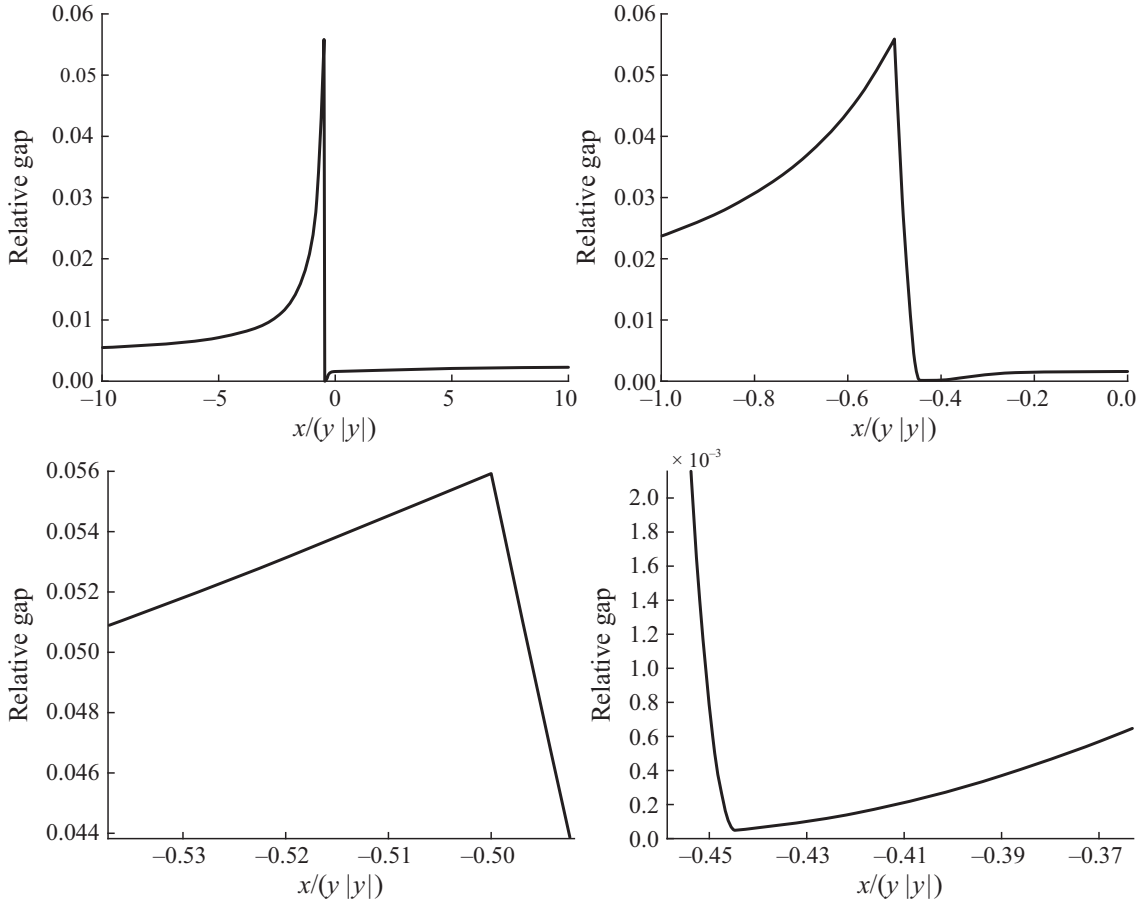

Fig. 3. Relative gap between the value of objective (1) on the time-optimal trajectory and on the optimal solution. The time-optimal trajectory switches control on the curve $x = -\frac{y|y|}{2}$, whereas the optimal trajectory switches control on the curve $x = -\beta y|y|$ with $\beta \approx 0.4446$. The figures on the right and bottom are zooms of the upper left figure.

Since both the time-optimal problem (10) and problem (5) reduce to their 1-dimensional versions Fig. 1 and (1) if the initial values of the vectors x, y are collinear, the same gap is achieved for the 2-dimensional problems for these initial values.

Let us now consider the self-similar trajectories found in [8]. First we compute the optimal value of problem (5) on these trajectories. By virtue of (4) the Bellman function on the self-similar trajectories satisfying (8) obeys

$$\omega_{2D}(x(t), y(t)) = \frac{\lambda(t)^5}{\lambda_0^5}\omega_{2D}(x(0), y(0)).$$

Differentiating with respect to t and using $\frac{d\omega(x(t),y(t))}{dt} = \frac{1}{2}\|x(t)\|^2$ yields

$$\omega_{2D}(x(0), y(0)) = -\frac{\lambda_0^5}{540}. \tag{20}$$

We now consider the value of the objective on the time-optimal trajectory with the same initial values (7). To this end we have to invert relations (16), (17), (18), i.e., determine the values of $\alpha, \tau, \bar{\tau}$ yielding these initial values.

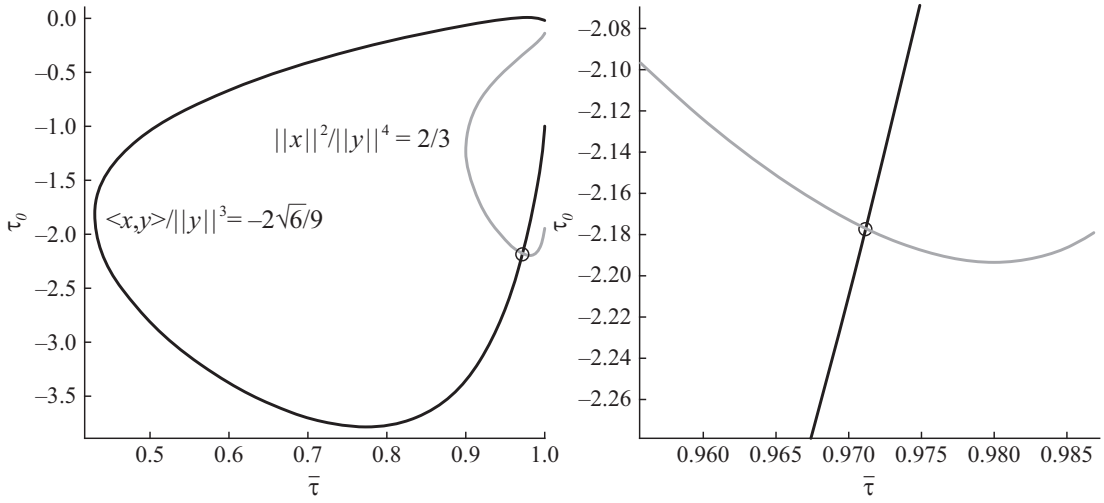

Fig. 4. Level curves of the ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$ corresponding to the self-similar trajectory in the $(\bar\tau,\tau_0)$ plane. The values of $\bar\tau,\tau_0$ producing an initial point on this trajectory are given by the unique intersection point of the curves (circle).

Following the scheme outlined in Section 2, we first consider the level curves of the ratios $\frac{\|x_0\|^2}{\|y_0\|^4}$, $\frac{\langle x_0,y_0\rangle}{\|y_0\|^3}$ in the $(\bar\tau,\tau_0)$ plane. From (7) we get the values

$$\frac{\|x_0\|^2}{\|y_0\|^4}=\frac{2}{3},\quad \frac{\langle x_0,y_0\rangle}{\|y_0\|^3}=\frac{2\sqrt{6}}{9}.$$

The corresponding level curves are depicted on Fig. 4. Refining the values obtained graphically by a Newton method we get

$$\bar\tau\approx 0.97116420999,\quad \tau_0\approx -2.17695799429.$$

Inserting into (18) and setting $\|y_0\|^2=\frac{\lambda_0^2}{6}$ by virtue of (7), we further obtain the value $\alpha\approx 4.13415835032\lambda_0^{-1}$.

Inserting the parameter values into the expression for the objective value on the time-optimal trajectory, we obtain the sub-optimal cost $\approx 0.0019779902706\lambda_0^5$. Compared with the optimal cost (20) computed above, this yields a relative gap of $\approx 6.8\times 10^{-2}$.

Let us now compare the optimal control on the self-similar trajectory with the time-optimal control on this trajectory. We choose the initial point which corresponds to the value $\lambda_0=1$, thus the optimal trajectory needs unit time to arrive at the origin.

The time-optimal control is given by (13), where τ runs from τ_0 to $\bar\tau$ and the time variable correspondingly from 0 to $T^{TO}=\alpha^{-1}(\bar\tau-\tau_0)\approx 0.7614904746$. Note that the control evolves clockwise around the origin. Note also that the arrival time at the origin is smaller than the arrival time $T=1$ for the optimal trajectory, because the time-optimal control minimizes precisely the arrival time. On Fig. 5 the polar angle of the time-optimal control is depicted as a function of time.

By (9) the optimal control on the self-similar trajectory evolves according to the formula

$$\hat u=(\cos\varphi(t),\sin\varphi(t))^T,\quad \varphi(t)=\sqrt{5}\log(T-t)+\text{const},$$

where $T=\lambda_0=1$ is the arrival time of the trajectory at the origin and the constant is the polar angle of the control at the initial time instant $t=0$. In order to determine this constant we first compute the initial value of y by formulas (14),(15). It amounts to

$$y(0)\approx\begin{pmatrix}0.2878394861\\-0.2895083711\end{pmatrix}.$$

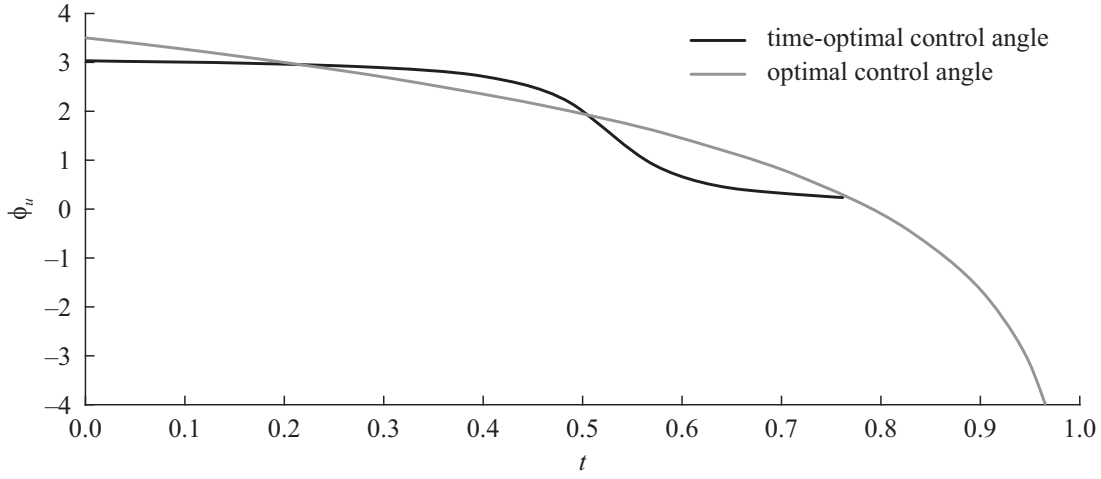

Fig. 5. Evolution of the polar angle of the optimal control and the time-optimal control for the same initial point, located on a self-similar trajectory with arrival time $T = 1$. For $t \to 1$ the optimal control angle tends to infinity.

As mentioned in Section 1, the optimal control is directed at an angle $\pi - \operatorname{atan}\sqrt{5}$ relative to y, leading to an initial control angle of ≈ 3.5035658841. For $t \to T = 1$ the polar angle of the control decreases logarithmically, and the control revolves an infinite number of times around the origin. The evolution of the angle as a function of time is depicted on Fig. 5.

5. CONCLUSIONS

In this paper we considered two optimal control problems, which share the feasible set of trajectories but have different objective values. While the time-optimal problem (10) can be solved analytically, for problem (5), which exhibits a singular trajectory of second order, only a limited number of optimal trajectories are known explicitly.

We describe the solution of the time-optimal control problem and use its solution to construct an upper bound on the objective value of problem (5). Comparison of the value of the sub-optimal (time-optimal) solution with the value of known optimal trajectories shows that the relative gap in objective value ranges from several thousandth of a per cent to several per cent. The difference in the polar angle of the two controls can, however, be quite substantial (up to 45 degrees).

The upper bound can be used to constrain the locus of the optimal trajectories of problem (5) in extended phase space (i.e., jointly with the adjoint variables) and thus simplify the analysis of the optimal synthesis of this problem. More concretely, it was shown in [9] that the Fuller symmetry (3) implies that the Bellman function is given by $\omega(x, y) = \frac{1}{5}(\langle\psi, y\rangle + 2\langle\phi, x\rangle)$, where ϕ, ψ are the adjoint variables to x, y. Hence an upper bound on the objective value at a given point (x, y) implies a linear inequality on the optimal values of the adjoint variables at this point.

Numerical experiments show that the self-similar trajectory is repulsive in the factor (orbit) space with respect to the action of the symmetry groups, while the trajectories corresponding to the 1D analog (1) are attractive. A rigorous proof of this property and a qualitative description of the complete optimal synthesis of the problem remain open and will be subject of future investigations.

FUNDING

The research is supported by the Ministry of Science and Higher Education of the Russian Federation, project no. FSMG-2024-0011.

REFERENCES

1. Fuller, A.T., Relay Control Systems Optimized for Various Performance Criteria, *Proceedings of the First World Congress IFAC*, Butterworth, 1960, pp. 510–519.
2. Kelley, H.J., Kopp, R.E. and Moyer, M.G., Singular Extremals, in *Topics in Optimization*, New York: Academic Press, 1967, pp. 63–101.
3. Kupka, I., Generic Properties of Extremals in Optimal Control Problems, in *Differential Geometric Control Theory*, Boston: Birkhäuser, 1983, vol. 27, pp. 310–315.
4. Lewis, R.M., Definitions of Order and Junction Condition in Singular Control Problems, *SIAM J. Contr. Optim.*, 1980, vol. 18, no. 1, pp. 21–32.
5. Lokutsievskiy, L.V., Generic Structure of the Lagrangian Manifold in Chattering Problems, *Sbornik Math.*, 2014, vol. 205, no. 3, pp. 432–458.
6. Marchal, C., Chattering Arcs and Chattering Controls, *J. Optimiz. Theory App.*, 1973, vol. 11, no. 5, pp. 441–468.
7. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., and Mischchenko, E.F., *The Mathematical Theory of Optimal Processes*, New York–London: Wiley, 1962.
8. Zelikin, M.I. and Borisov, V.F., *Theory of Chattering Control with Applications to Astronautics, Robotics, Economics, and Engineering*, Boston: Birkhäuser, 1994.
9. Zelikin, M.I., Melnikov, N.B., and Hildebrand, R., Topological Structure of a Typical Fibre of Optimal Synthesis for Chattering Problems, *Proc. Steklov Inst. Math.*, 2001, vol. 233, pp. 116–142.
10. Zelikin, M.I., Lokutsievskiy, L.V., and Hildebrand, R., Typicality of Chaotic Fractal Behaviour of Integral Vortices in Hamiltonian Systems with Discontinuous Right Hand Side, *Journal of Mathematical Sciences*, 2017, vol. 221, no. 1, pp. 1–136.

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board

ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 8, pp. 782–795.

SPECIAL ISSUE

Solving Large Multicommodity Network Flow Problems on GPUs

F. Zhang[*,a] **and S. Boyd**[*,b]

[*] *Stanford University, Stanford, USA*
e-mail: [a] *zfzhao@stanford.edu,* [b] *boyd@stanford.edu*

Received March 3, 2025
Revised May 20, 2025
Accepted June 27, 2025

Abstract—We consider the all-pairs multicommodity network flow problem on a network with capacitated edges. The usual treatment keeps track of a separate flow for each source-destination pair on each edge; we rely on a more efficient formulation in which flows with the same destination are aggregated, reducing the number of variables by a factor equal to the size of the network. Problems with hundreds of nodes, with a total number of variables on the order of a million, can be solved using standard generic interior-point methods on CPUs; we focus on GPU-compatible algorithms that can solve such problems much faster, and in addition scale to much larger problems, with up to a billion variables. Our method relies on the primal-dual hybrid gradient algorithm, and exploits several specific features of the problem for efficient GPU computation. Numerical experiments show that our primal-dual multicommodity network flow method accelerates state-of-the-art generic commercial solvers by 100 to 1000 times, and scales to problems that are much larger. We provide an open-source implementation of our method.

Keywords: multicommodity flows, primal-dual method, GPU-based optimizer

DOI: 10.31857/S0005117925080072

1. MULTICOMMODITY NETWORK FLOW OPTIMIZATION

1.1. Multicommodity Network Flow Problem

Our formulation of the multicommodity network flow (MCF) problem, given below, follows [1].

Network. We consider a directed network with n nodes and m edges which is completely connected, i.e., there is a directed path between each pair of nodes. Let $A \in \mathbf{R}^{n\times m}$ denote its incidence matrix, i.e.,

$$A_{i\ell} = \begin{cases} +1 & \text{edge } \ell \text{ enters node } i \\ -1 & \text{edge } \ell \text{ leaves node } i \\ 0 & \text{otherwise.} \end{cases}$$

Edge ℓ has a positive capacity c_ℓ. The total flow on edge ℓ (to be defined below) cannot exceed c_ℓ.

Traffic matrix. We consider the all-pairs multicommodity flow setting, i.e., there is traffic that originates at every node, destined for every other node. We characterize the traffic between all source-destination pairs via the traffic matrix $T \in \mathbf{R}^{n\times n}$. For any pair of distinct nodes i, j, $T_{ij} \geqslant 0$ is the traffic from (source) node j to (destination) node i. There is no traffic from a node to itself; for mathematical convenience we define the diagonal traffic matrix entries as $T_{ii} = -\sum_{j\neq i} T_{ij}$, the negative of the total traffic with destination node i. With this definition of the diagonal entries, we have $T\mathbf{1} = 0$, where $\mathbf{1}$ is the vector with all entries one.

Network utility. Let u_{ij} denote the strictly concave increasing utility function for traffic from node j to node i, for $j \neq i$. We will assume utility functions are differentiable with domains $\mathbf{R}_{++}$,

the set of positive numbers. (The methods we describe are readily extended to nondifferentiable utilities using subgradients instead of gradients.) The total utility, which we wish to maximize, is $\sum_{i\neq j} u_{ij}(T_{ij})$. For simplicity we take $u_{ii} = 0$, so we can write the total utility as

$$U(T) = \sum_{i,j} u_{ij}(T_{ij}).$$

The domain of U is $\mathcal{T} = \{T \mid T_{ij} > 0 \text{ for } i \neq j\}$, i.e., the traffic matrix must have positive off-diagonal entries.

Common examples of utility functions include the weighted log utility $u(s) = w \log s$, and the weighted power utility $u(s) = ws^{\gamma}$, with $\gamma \in (0, 1)$, where $w > 0$ is the weight.

Destination-based flow matrix. Following [1] we aggregate all flows with the same destination, considering it to be one commodity that is conserved at all nodes except the source and destination, but can be split and combined. The commodity flows are given by the (destination-based) flow matrix $F \in \mathbf{R}^{n\times m}$, where $F_{i\ell} \geqslant 0$ denotes the flow on edge ℓ that is destined to node i. The edge capacity constraint can be expressed as $F^T\mathbf{1} \leqslant c$, where the inequality is elementwise. A similar flow aggregation formulation, though source-based, was considered in [2].

Flow conservation. The flow destined for node i is conserved at all nodes $j \neq i$, including the additional injection of traffic T_{ij} that originates at node j and is destined for node i. This means that

$$T_{ij} + \sum_{\ell} A_{j\ell}F_{i\ell} = 0, \quad i, j = 1, \ldots, n, \quad j \neq i.$$

At the destination node, all traffic exits and we have (using our definition of T_{ii})

$$T_{ii} + \sum_{\ell} A_{i\ell}F_{i\ell} = 0, \quad i = 1, \ldots, n.$$

Combining these two, and using our specific definition of T_{ii}, flow conservation can be compactly written in matrix notation as

$$T + FA^T = 0.$$

Multicommodity flow problem. In the MCF problem, we seek a flow matrix that maximizes total network utility, subject to the edge capacity and flow conservation constraints. This can be expressed as the problem

$$\begin{array}{ll} \text{maximize} & U(T) \\ \text{subject to} & F \geqslant 0, \quad F^T\mathbf{1} \leqslant c, \quad T + FA^T = 0, \end{array} \tag{1}$$

with variables F and T, and implicit constraint $T \in \mathcal{T}$. The problem data are the network topology A, edge capacities c, and the traffic utility functions u_{ij}.

We can eliminate the traffic matrix T using $T = -FA^T$ and state the MCF problem in terms of the variable F alone as

$$\begin{array}{ll} \text{maximize} & U(-FA^T) \\ \text{subject to} & F \geqslant 0, \quad F^T\mathbf{1} \leqslant c, \end{array} \tag{2}$$

with variable F, and implicit constraint $-FA^T \in \mathcal{T}$. The number of scalar variables in this problem is nm. For future use, we define the feasible flow set as

$$\mathcal{F} = \{F \mid F \geqslant 0,\ F^T\mathbf{1} \leqslant c\}.$$

Existence and uniqueness of solution. First, let us show the MCF problem (1) is always feasible. Consider a unit flow from each source to each destination, over the shortest path, i.e., the smallest

number of edges, which exists since the graph is completely connected. We denote this flow matrix as F^{sp}. Now take $F = \alpha F^{\mathrm{sp}}$, where $\alpha = 1/\max_\ell((F^{\mathrm{sp}T}\mathbf{1})_\ell/c_\ell) > 0$, so we have $F^T\mathbf{1} \leqslant c$. Evidently F is feasible, and we have $T_{ij} = \alpha > 0$ for $i \neq j$, so $T = -FA^T \in \mathcal{T}$. This shows that the problem is always feasible. Let U^{sp} denote the corresponding objective function.

We can add the constraint $U(T) \geqslant U^{\mathrm{sp}}$ to the problem, without changing the solution set. With this addition, the feasible set is compact. It follows that the MCF problem (1) always has a solution. The solution need not be unique. The optimal T, however, is unique. We also note that the argument above tells us that the implicit constraint $T = -FA^T \in \mathcal{T}$ is redundant.

Solving MCF. The multicommodity flow problem (2) is convex [3], and so can be efficiently solved in principle. In [1] the authors use standard generic interior-point solvers such as commercial solver MOSEK [4], together with CVXPY [5], to solve instances of the problem with tens of nodes, and thousands of variables, in a few seconds on a CPU. In this paper, we introduce an algorithm for solving the MCF problem that fully exploits GPUs. For small and medium size problems, our method gives a substantial speedup over generic methods; in addition, it scales to much larger problems that cannot be solved by generic methods.

1.2. Optimality Condition and Residual

Optimality condition. Let $\tilde{\mathcal{F}}$ denote the closure of the feasible set, including the implicit constraint $T = -FA^T \in \mathcal{T}$,

$$\tilde{\mathcal{F}} = \mathcal{F} \cap \{F \mid -FA^T \in \mathbf{cl}(\mathcal{T})\},$$

where $\mathbf{cl}(\mathcal{T})$ denotes the closure of $\mathcal{T}$.

Then F is optimal for (2) if and only if $F \in \mathcal{F}$, $-FA^T \in \mathcal{T}$, and

$$\mathbf{Tr}(Z - F)^T G \geqslant 0$$

holds for all $Z \in \tilde{\mathcal{F}}$, where $G = \nabla_F(-U)(-FA^T)$ (see, e.g., [3, §4.2.3]). We have $G = U'A$, where $U'_{ij} = u'_{ij}((-FA^T)_{ij})$.

Optimality condition via projection onto $\mathcal{F}$. For future use, we express the above optimality condition in terms of projection of a matrix Q onto $\mathcal{F}$. Let Π denote Euclidean projection onto $\mathcal{F}$. Suppose $Q \in \mathbf{R}^{n\times m}$, and set $F = \Pi(Q)$, so $F \in \mathcal{F}$. Suppose in addition that $-FA^T \in \mathcal{T}$, so that $G = \nabla_F((-U)(-FA^T)$ exists. Then F is also Euclidean projection of Q onto $\tilde{\mathcal{F}}$. It follows that $\mathbf{Tr}(Z - F)^T G \geqslant 0$ for all $Z \in \tilde{\mathcal{F}}$, so the optimality condition above holds, and F is optimal. Evidently, it would hold if the weaker condition

$$G = \gamma(F - Q) \text{ for some } \gamma \geqslant 0$$

holds.

Summarizing: F is optimal if $F = \Pi(Q)$ for some Q, $-FA^T \in \mathcal{T}$, and $G = \gamma(F - Q)$ for some $\gamma \geqslant 0$. The converse is also true: If F is optimal then $F = \Pi(Q)$ for some Q with $-FA^T \in \mathcal{T}$ and $G = \gamma(F - Q)$ for some $\gamma \geqslant 0$. (Indeed, this holds with $\gamma = 1$ and $Q = F - G$.) This optimality condition is readily interpreted: It states that F is a fixed point of a projected gradient step with step size γ.

Optimality residual. For any $Q \in \mathbf{R}^{n\times m}$ with $F = \Pi(Q)$, we define the (optimality) residual as

$$r(Q) = \begin{cases} \min_{\gamma\geqslant 0} \|G - \gamma(F - Q)\|_F^2 & -FA^T \in \mathcal{T} \\ \infty & \text{otherwise,} \end{cases}$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm of a matrix, i.e., the sum of squares of its entries. When $-FA^T \in \mathcal{T}$, the righthand side is a quadratic function of γ, so the minimum is

easily expressed explicitly as

$$r(Q) = \begin{cases} \|G\|_F^2 - \dfrac{\mathbf{Tr}^2\, G^T(F-Q)}{\|F-Q\|_F^2} & -FA^T \in \mathcal{T},\ F \neq Q,\ \ \mathbf{Tr}\, G^T(F-Q) \geqslant 0 \\ \|G\|_F^2 & -FA^T \in \mathcal{T},\ F = Q \text{ or } \mathbf{Tr}\, G^T(F-Q) < 0 \\ \infty & \text{otherwise.} \end{cases} \tag{3}$$

Evidently $F = \Pi(Q)$ is optimal if and only if $r(Q) = 0$.

1.3. Related Work

Multicommodity network flow. Historically, different forms of MCF problems have been formulated and studied. Starting from [6, 7] which studied a version with linear utility functions, which can be formulated as a linear program, later works develop nonlinear convex program formulations [8, 9] and (nonconvex) mixed integer program formulations [10–12] of MCF problems for different application purposes. These various forms of MCF have been widely used in transportation management [13–15], energy and economic sectors [8, 10, 16], and network communication [11, 17, 18]. [19] surveys over two hundred studies on MCF problems between 2000 and 2019. In this work, we focus on nonlinear convex formulation of MCF problems and develop GPU-compatible algorithms for solving large problem instances. See [20] for a survey on nonlinear convex MCF problems. MCF models have very recently been exploited to design multi-GPU communication schedules for deep learning tasks [21, 22], but the underlying MCF problems are solved with CPU-based solvers.

First-order methods for convex optimization. First-order methods such as gradient descent algorithm, proximal point algorithm, primal-dual hybrid gradient algorithm, and their accelerated versions have been exploited to tackle different forms of convex optimization problems. Compared to second-order methods which exploit Hessian information, first-order methods are known for their low computational complexity and are thus attractive for solving large-scale optimization problems. Recently, primal-dual hybrid gradient algorithm has been explored for solving large linear programs [23–25] and optimal transport problems [26] on GPUs. Other first-order methods such as ADMM have been exploited for designing GPU-accelerated optimizers for optimal power flow problems [27, 28].

GPU-accelerated network flow optimization. Specialized to GPU-based optimizers for network flow optimization, [29] considers implementing a parallel routing algorithm on GPUs for SDN networks, which solves the Lagrangian relaxation of a mixed integer linear program. [30] implements a genetic method on GPUs for solving an integer linear program formulation of the routing problem. [31] considers a linear program formulation of multicommodity network flow problems and constructs a deep learning model for generating new columns in delayed column generation method. [32] implements an asynchronous push-relabel algorithm for single commodity maximum network flow problem, which is CPU-GPU hybrid. [33] exploits exactly the same flow aggregation formulation of MCF following [1] as we do and trains a neural network model for minimizing unconstrained Lagrangian relaxation objective, and feeds the result as warm start to Gurobi [34] to get the final answer. [35] integrates a source-based flow aggregation formulation of the multicommodity flow problem into solving the combined transportation model and exploits an accelerated variant of proximal alternating predictor-corrector algorithm. The authors claim that the proposed algorithm is GPU-friendly, but the numerical experiments are CPU-based, and involve small size networks. [36] adopts a primal-dual gradient method for solving combined traffic models, which however is not GPU-oriented.

1.4. Contribution

Motivated by the recent advancement of GPU optimizers, in this work we seek to accelerate large-scale nonlinear convex MCF problem solving with GPUs. Specifically, we adopt the MCF problem formulation in [1] (also described above) which is compactly matrix-represented and requires fewer optimization variables by exploiting flow aggregation. We show that this specific problem formulation can be efficiently solved with first-order primal-dual hybrid gradient method when run on GPUs.

To the best of our knowledge, our work is the first to tackle exactly solving convex MCF problems on GPUs. Classic works for solving such large-scale MCF problems usually adopt Lagrangian relaxation for the coupling constraint and solve the resulting subproblems with smaller sizes in parallel (see, e.g., [20]). In our work, we do not exploit any explicit problem decomposition strategy and our algorithmic acceleration is mainly empirical and depends on highly-optimized CUDA kernels for matrix operations. Moreover, we achieve problem size reduction via flow aggregation. Therefore, our method has a simpler form that does not involve massive subproblem solving and synchronizing, and is also exact.

1.5. Outline

We describe our algorithm in §2. Experimental results, using our PyTorch implementation, are presented and discussed in §3; very similar results obtained with our JAX implementation are given in Appendix B. We conclude our work in §4. The code, and all data needed to reproduce the results reported in this paper, can be accessed at `https://github.com/cvxgrp/pdmcf`.

2. PRIMAL-DUAL HYBRID GRADIENT

2.1. Primal-Dual Saddle Point Formulation

We first derive a primal-dual saddle point formulation of the MCF problem (1). Let $\mathcal{I}$ denote the indicator function of $\mathcal{F}$, i.e., $\mathcal{I}(F) = 0$ for $F \in \mathcal{F}$ and $\mathcal{I}(F) = \infty$ otherwise. We switch to minimizing $-U$ in (1) to obtain the equivalent problem

$$\begin{array}{ll} \text{minimize} & -U(T) + \mathcal{I}(F) \\ \text{subject to} & T = -FA^T, \end{array} \tag{4}$$

with variables T and F. We introduce a dual variable $Y \in \mathbf{R}^{n\times n}$ associated with the (matrix) equality constraint. Then the Lagrangian is

$$\mathcal{L}(T, F; Y) = -U(T) + \mathcal{I}(F) - \mathbf{Tr}\, Y^T(T + FA^T)$$

(see [3, Chap. 5]). The Lagrangian $\mathcal{L}$ is convex in the primal variables (T, F) and affine (and therefore concave) in the dual variable Y. If $(T, F; Y)$ is a saddle point of $\mathcal{L}$, then (T, F) is a solution to problem (4) (and F is a solution to the MCF problem (2)); the converse also holds.

We can analytically minimize $\mathcal{L}$ over T to obtain the reduced Lagrangian

$$\hat{\mathcal{L}}(F; Y) = \inf_T \mathcal{L}(T, F; Y) = -(-U)^*(Y) + \mathcal{I}(F) - \mathbf{Tr}\, Y^T FA^T, \tag{5}$$

where U^* is the conjugate function of U [3, §3.3]. This reduced Lagrangian is convex in the primal variable F and concave in the dual variable Y. If $(F; Y)$ is a saddle point of $\hat{\mathcal{L}}$, then F is a solution to the MCF problem (2) (see [37, § 1]). We observe that $\hat{\mathcal{L}}$ is convex-concave, with a bilinear coupling term.

2.2. Basic PDHG Method

The primal-dual hybrid gradient (PDHG) algorithm, as first introduced in [38] and later studied in [39, 40], is a first-order method for finding a saddle point of a convex-concave function with bilinear coupling term. The algorithm was extended to include over-relaxation in [40, §4.1], which has been observed to improve convergence in practice. For (5), PDHG has the form

$$\begin{aligned}
\hat{F}^{k+1/2} &= \mathbf{prox}_{\alpha\mathcal{I}}(F^{k-1/2} + \alpha Y^k A) \\
F^{k+1} &= 2\hat{F}^{k+1/2} - F^{k-1/2} \\
\hat{Y}^{k+1} &= \mathbf{prox}_{\beta(-U)^*}(Y^k - \beta F^{k+1} A^T) \\
F^{k+1/2} &= \rho\hat{F}^{k+1/2} + (1-\rho)F^{k-1/2} \\
Y^{k+1} &= \rho\hat{Y}^{k+1} + (1-\rho)Y^k
\end{aligned} \tag{6}$$

where $\mathbf{prox}_f(v) = \operatorname{argmin}_x(f(x) + (1/2)\|x - v\|_2^2)$ denotes the proximal operator of f [41], $\alpha, \beta > 0$ are positive step sizes satisfying $\alpha\beta \leqslant 1/\|A\|_2^2$, and $\rho \in (0, 2)$ is the over-relaxation parameter.

Reasonable choices for the parameters are

$$\alpha = \beta = 1/\|A\|_2, \qquad \rho = 1.9.$$

(An upper bound on $\|A\|_2$ can be used in place of $\|A\|_2$.)

Convergence. In [40] it has been shown that when there exists a saddle point of $\hat{\mathcal{L}}$, $(F^k; Y^k)$ converges to a saddle point of $\hat{\mathcal{L}}$ as $k \to \infty$. For MCF the existence of an optimal flow matrix and dual variable is known, so F^k converges to an optimal flow matrix. It follows that $r(F^{k-1/2} + \alpha Y^k A) \to 0$ as $k \to \infty$. We note that $-FA^T \in \mathcal{T}$ only holds eventually.

2.3. Proximal Operators

Here we take a closer look at the two proximal operators appearing in PDHG.

First proximal operator. We note that $\mathbf{prox}_{\alpha\mathcal{I}}$ appearing in the $\hat{F}^{k+1/2}$ update of (6) is projection onto $\mathcal{F}$,

$$\mathbf{prox}_{\alpha\mathcal{I}}(F) = \Pi(F).$$

Since the constraints that define $\mathcal{F}$ separate across the columns of F, we can compute $\Pi(F)$ by projecting each column f_ℓ of F onto the scaled simplex $\mathcal{S}_\ell = \{f \mid f \geqslant 0,\ \mathbf{1}^T f \leqslant c_\ell\}$. This projection has the form

$$\Pi_{\mathcal{S}_\ell}(f_\ell) = (f_\ell - \mu_\ell \mathbf{1})_+,$$

where μ_ℓ is the optimal Lagrange multiplier and $(a)_+ = \max\{a, 0\}$, which is applied elementwise to a vector. The optimal μ_ℓ is the smallest nonnegative value for which $(f_\ell - \mu_\ell \mathbf{1})_+^T \mathbf{1} \leqslant c_\ell$. This is readily found by a bisection algorithm; see §2.6.

Second proximal operator. The proximal operator appearing in the $\hat{Y}^{k+1}$ update step in (6) can be decomposed entrywise, since $\beta(-U)^*$ is a sum of functions of different variables. (The diagonal entries $-u_{ii}$ are zero, so $(-\beta u_{ii})^*$ is the indicator function of $\{0\}$, and its proximal operator is the zero function.) For each off-diagonal entry $i \neq j$ we need to evaluate

$$\mathbf{prox}_{\beta(-u_{ij})^*}(y).$$

These one-dimensional proximal operators are readily computed in the general case. For the weighted log utility $u(s) = w \log s$, we have

$$\mathbf{prox}_{\beta(-u)^*}(y) = \frac{y - \sqrt{y^2 + 4\beta w}}{2}.$$

For the weighted power utility $u(s) = ws^\gamma$, $\mathbf{prox}_{\beta(-u)^*}(y)$ is the unique negative number z for which

$$(-z)^{c_1+2} + y(-z)^{c_1+1} - c_1 c_2 = 0,$$

where

$$c_1 = \frac{\gamma}{1-\gamma} > 0, \quad c_2 = \beta\left(\frac{1}{\gamma} - 1\right)(w\gamma)^{\frac{1}{1-\gamma}} > 0.$$

2.4. Adaptive Step Sizes

In the basic PDHG algorithm (6), the step sizes α and β are fixed. It has been observed that varying them adaptively as the algorithm runs can improve practical convergence substantially [23]. We describe our implementation of adaptive step sizes here.

We express the step sizes as

$$\alpha^k = \eta/\omega^k, \qquad \beta^k = \eta\omega^k,$$

where $\eta \leqslant 1/\|A\|_2$ and $\omega^k > 0$ gives the primal weight. With $\omega^k = 1$ we obtain basic PDHG (6).

The primal weight ω^k is initialized as $\omega^0 = 1$ and adapted following [23, §3.3] as

$$\omega^{k+1} = \left(\frac{\Delta_Y^{k+1}}{\Delta_F^{k+1}}\right)^\theta \left(\omega^k\right)^{1-\theta}, \tag{7}$$

where $\Delta_F^{k+1} = \|F^{k+1/2} - F^{k-1/2}\|_F$, $\Delta_Y^{k+1} = \|Y^{k+1} - Y^k\|_F$ and θ is a parameter fixed as 0.5 in our implementation. The intuition behind the primal weight update (7) is to balance the primal and dual residuals; see [23, §3.3] for details. In [23] the authors update ω each restart. We do not use restarts, and have found that updating ω^k every k^{adapt} iterations, when both $\Delta_F^k > 10^{-5}$ and $\Delta_Y^k > 10^{-5}$ hold, works well in practice for MCF. In our experiments, we use $k^{\text{adapt}} = 100$. We can also stop adapting ω^k after some number of iterations, keeping it constant in future iterations. At least technically this implies that the convergence proof for constant ω holds for the adaptive algorithm.

A simple bound on $\|A\|_2$. We can readily compute a simple upper bound on

$$\|A\|_2 = \sqrt{\lambda_{\max}(AA^T)},$$

where $\lambda_{\max}$ denotes the maximum eigenvalue. We observe that AA^T is the Laplacian matrix associated with the network, for which the well-known bound

$$\lambda_{\max}(AA^T) \leqslant 2d_{\max}$$

holds, where $d_{\max}$ is the largest node degree in the graph. (For completeness we derive this in Appendix A.) Thus we can take

$$\eta = 1/\sqrt{2d_{\max}}. \tag{8}$$

2.5. Algorithm

We summarize our final algorithm, which we call PDMCF. We set $r^0 = +\infty$, $\alpha^0 = \eta/\omega^0$, and $\beta^0 = \eta\omega^0$, where η is given in (8) and $\omega^0 = 1$.

Algorithm 2.1. PDMCF

given $F^{-1/2}, Y^0$, parameter $\epsilon > 0$.

for $k = 0, 1, \ldots$

1. *Check stopping criterion.* Quit and return $\hat{F}^{k-1/2}$ if $r^k < nm\epsilon$ holds.
2. *Basic PDHG updates* (6).
   $\hat{F}^{k+1/2} = \Pi(F^{k-1/2} + \alpha^k Y^k A)$.
   $F^{k+1} = 2\hat{F}^{k+1/2} - F^{k-1/2}$.
   $\hat{Y}_{ij}^{k+1} = \begin{cases} \mathbf{prox}_{\beta^k(-u_{ij})^*}(Y_{ij}^k - \beta^k (F^{k+1}A^T)_{ij}) & j \neq i \\ 0 & j = i. \end{cases}$
   $F^{k+1/2} = \rho\hat{F}^{k+1/2} + (1-\rho)F^{k-1/2}$.
   $Y^{k+1} = \rho\hat{Y}^{k+1} + (1-\rho)Y^k$.
3. *Adaptive step size updates* (7) (if k is multiple of k^{adapt} and $\Delta_F^{k+1}, \Delta_Y^{k+1} > \tau$).
   $\omega^{k+1} = \left(\Delta_Y^{k+1}/\Delta_F^{k+1}\right)^{\theta} \left(\omega^k\right)^{1-\theta}$.
   $\alpha^{k+1} = \eta/\omega^{k+1}, \qquad \beta^{k+1} = \eta\omega^{k+1}$.

Initialization. We always take $F^{-1/2} = 0$ and $Y^0 = I - \mathbf{1}\mathbf{1}^T$. We can alternatively use a better guess of $F^{-1/2}$ and Y, for example in a warm start, when we have already solved a problem with similar data. We illustrate more on this in §3.1.

Stopping criterion. Since $\hat{F}^{k+1/2}$ is result of projection onto $\mathcal{F}$, our optimality residual (3) has the form

$$r^{k+1} = r(F^{k-1/2} + \alpha^k Y^k A).$$

We consider the stopping criterion $r^k < nm\epsilon$, i.e., the entrywise normalized residual r^k/nm is smaller than a user-specified threshold ϵ.

2.6. Implementation Details

Incidence matrix indexing. We only store the indices of the non-zero entries of A. Matrix multiplication with A and A^T can be efficiently computed by exploiting scatter and gather functions, which are highly optimized CUDA kernels and are available in most major GPU computing languages.

Projection onto scaled simplex. To compute μ_ℓ when $(f_\ell)_+^T\mathbf{1} > c_\ell$, we follow [42] and first sort f_ℓ from largest entry to smallest entry to form f'_ℓ. We then find the largest index t such that $f'_{\ell t} - ((\sum_{i=1}^t f'_{\ell i} - c_\ell)/t) > 0$. Finally we take $\mu_\ell = (\sum_{i=1}^t f'_{\ell i} - c_\ell)/t$.

Some recent work develops a more efficient method to compute the projection onto the simplex set ([43, 44] for example); we adopt the simpler algorithm described above for implementation simplicity.

3. EXPERIMENTS

We run all our experiments on a single H100 GPU with 80 Gb of memory supported by 26 virtual CPU cores and 241 Gb of RAM. The results given below are for our PyTorch implementation; similar results, reported in Appendix B, are obtained with our JAX implementation.

3.1. Examples

Data and parameters. We consider weighted log utilities of form $u_{ij}(T_{ij}) = w_{ij}\log T_{ij}$. We take $\log w_{ij}$ to be uniform on $[\log 0.3, \log 3]$. For network topology, we first create n two-dimensional

data points $\xi_i \in \mathbf{R}^2$, each denoted by (ξ_{ix}, ξ_{iy}) for $i = 1, \ldots, n$. We take ξ_{ix} and ξ_{iy} uniform on $[0, 1]$. Then we add both edges (ξ_i, ξ_j) and (ξ_j, ξ_i) when either ξ_i is among the q-nearest neighbors of ξ_j or vice versa. For each edge ℓ, we impose edge capacity c_ℓ where we take $\log c_\ell$ to be uniform on $[\log 0.5, \log 5]$.

We use the stopping criterion threshold $\epsilon = 0.01/(n(n-1))$ for small to medium size problems and $\epsilon = 0.03/(n(n-1))$ for large size problems. We compare to CPU-based commercial solver MOSEK, with default settings. MOSEK is able to solve the problems with high accuracy; we have checked that for all problem instances, the normalized utility differences between the results of PDMCF and MOSEK are no more than around 0.01. The pairwise normalized (optimal) utilities range between around 1 and 10, which means that PDMCF finds flows that are between 0.1% and 1% suboptimal compared to the flows found by MOSEK.

Small to medium size problems. Table 1 shows runtime for both MOSEK and PDMCF required to solve problem instances of various sizes. The column titled nm gives the number of scalar optimization variables in the problem instance. We see that our implementation of PDMCF on a GPU gives a speedup over MOSEK of 10 to 1000 times, with more significant speedup for larger problem instances. We also report runtime for PDMCF when run on CPU, which is still quicker than MOSEK but with a significantly lower speedup. Similar performance is also observed for our JAX implementation, reported in Appendix B.

Table 1. Runtime table for small and medium size problems

problem sizes				timing (s)			iterations
n	q	m	nm	MOSEK	PDMCF (CPU)	PDMCF (GPU)	
100	10	1178	1×10^5	5	1	0.5	490
200	10	2316	5×10^5	23	2	0.7	690
300	10	3472	1×10^6	95	6	0.8	840
500	10	5738	3×10^6	340	18	1.1	950
500	20	11 176	6×10^6	1977	34	1.4	790
1000	10	11 424	1×10^7	2889	1382	19.5	7220
1000	20	22 286	2×10^7	16 765	349	5.1	1040

Large size problems. Table 2 shows runtime for several large problem instances. MOSEK fails to solve all these problems due to memory limitations. PDMCF handles all these problem instances, with the largest one involving 10^9 variables.

Table 2. Runtime table for large size problems

problem sizes				timing (s)			iterations
n	q	m	nm	MOSEK	PDMCF (CPU)	PDMCF (GPU)	
3000	10	34 424	1×10^8	OOM	7056	96	4140
5000	10	57 338	3×10^8	OOM	19 152	395	3970
10 000	10	114 054	1×10^9	OOM	87 490	1908	4380

Scaling. We scatter plot the runtime data for small and medium problem instances in Fig. 1. Here we take 5 problem instances generated by iterating over random seeds $\{0, 1, 2, 3, 4\}$ for the different n, q values listed in Table 1. The x-axis represents optimization variable size nm and the y-axis represents runtime in seconds. We plot on a log-log scale. The lines show the affine function fits these data, with a slope around 1.5 for MOSEK and around 0.5 for PDMCF.

Convergence plot. Figure 2 shows the convergence for three problem instances with variable sizes $10^5, 10^6$, and 10^7 with PDMCF, where the x-axis represents iteration numbers. Especially in the initial iterations, we have infinite residual r^k since $-FA^T \notin \mathcal{T}$. For those iterations, we plot the

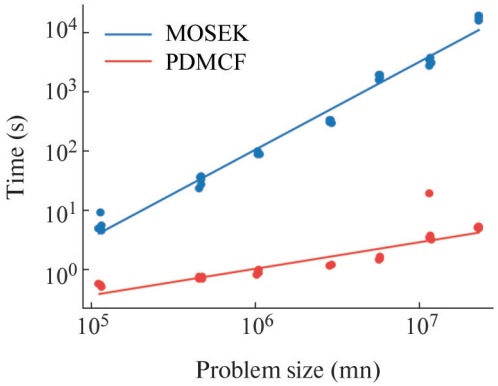

Fig. 1. Runtime plot for small and medium size problems.

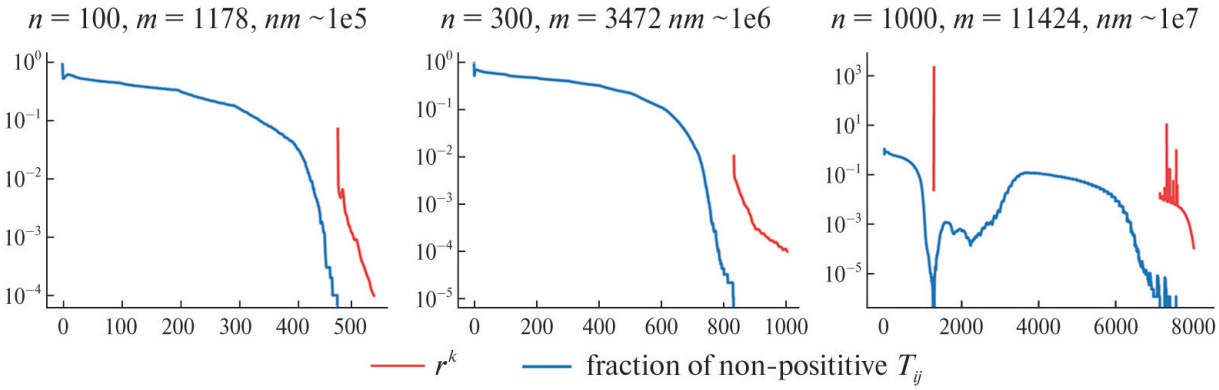

Fig. 2. Convergence plot for small and medium size problems.

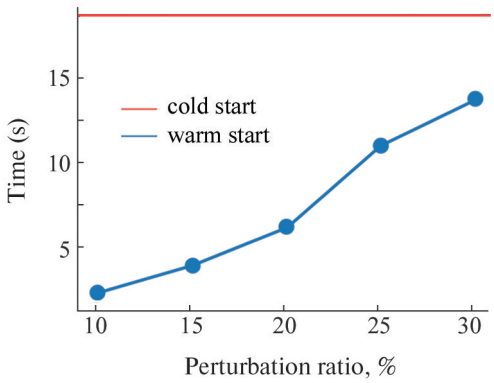

Fig. 3. Warm start plot for medium size problem.

fraction of nonpositive off-diagonal entries of T in blue. For feasible iterations, we plot the (finite) residual, in red.

Warm start. In §2.5 we start with some simple initial $F^{-1/2}$ and Y^0. We also test the performance of PDMCF with warm starts. In Fig. 3 we present how runtime changes under different warm starts. To form these warm starts, for some perturbation ratio ν, we randomly perturb entries of our utility weight matrix to derive $\tilde{w}_{ij} = (1 \pm \nu)w_{ij}$, each with probability a half. We solve the multicommodity network flow problem with perturbed utility weight $\tilde{w}$ with PDMCF until we land at a feasible point $(F^{\text{feas}}, Y^{\text{feas}})$ satisfying $(-F^{\text{feas}}A^T)_{ij} > 0$ for all distinct i, j. We record the primal weight at this point as ω^{feas}. We then solve the desired multicommodity network flow

problem with original utility weight w with $F^{-1/2} = F^{\text{feas}}$, $Y^0 = Y^{\text{feas}}$ and $\omega^0 = \omega^{\text{feas}}$. We note that setting $\omega^0 = \omega^{\text{feas}}$ is important for accelerated convergence, otherwise it usually requires a similar number of iterations to converge as a cold start if we simply set $\omega^0 = 1$. In Fig. 3, we take problem instance with $n = 1000$, $q = 10$. x-axis stands for perturbation ratio ν and y-axis represents runtime in seconds. As can be observed, with perturbation ratio $\nu = 10\%$, we harness $> 80\%$ saving of runtime. Such savings keep decreasing to around 30% when $\nu = 30\%$, which makes sense given that larger perturbation indicates more different utility weights between the original and perturbed problems, thus our warm start is expected to stay further from an optimal solution to the original problem instance.

4. CONCLUSION

In this work, we present the PDMCF algorithm which accelerates solving multicommodity network flow problems on GPUs. Our method starts with a destination-based formulation of multicommodity network flow problems which reduces optimization variable amount compared to classic problem formulation. We then apply the PDHG algorithm to solve this destination-based problem formulation. Empirical results verify that our algorithm is GPU-friendly and brings up to three orders of magnitude of runtime acceleration compared to classic CPU-based commercial solvers. Moreover, our algorithm is able to solve ten times larger problems than those that can be solved by commercial CPU-based solvers.

APPENDIX A

Upper bound on $\lambda_{\max}(AA^T)$. For a directed graph with incidence matrix A, $d_i = (AA^T)_{ii}$ is the degree of node i and for $i \neq j$, $-(AA^T)_{ij}$ is the number of edges connecting node i and node j, i.e., 2 if both edges (i, j) and (j, i) exist. Note that $\lambda_{\max}(AA^T) = \max_{\|x\|_2=1} x^T(AA^T)x = \max_{x\neq 0} \frac{x^T(AA^T)x}{x^Tx}$. We have

$$\begin{aligned} x^T(AA^T)x &= \sum_i (AA^T)_{ii}x_i^2 + \sum_{i\neq j}(AA^T)_{ij}x_ix_j \\ &= \sum_i d_ix_i^2 + \sum_{i\neq j}(AA^T)_{ij}x_ix_j \\ &\leqslant \sum_i d_ix_i^2 + \sum_{i\neq j}|(AA^T)_{ij}|(x_i^2/2 + x_j^2/2) \\ &= \sum_i x_i^2(d_i + \sum_{j\neq i}|(AA^T)_{ij}|) \\ &= \sum_i 2d_ix_i^2 \\ &\leqslant 2d_{\max}x^Tx. \end{aligned}$$

Therefore

$$\lambda_{\max}(AA^T) = \max_{x\neq 0}\frac{x^T(AA^T)x}{x^Tx} \leqslant 2d_{\max}.$$

APPENDIX B

JAX results. The results shown in §3.1 are for our PyTorch implementation. Here we provide the same results for our JAX implementation. Tables 3 and 4 show the runtimes on the same problem instances as reported in Tables 1 and 2. We note that JAX's just-in-time (JIT) compilation adds

Table 3. Runtime table for small and medium size problems (JAX)

problem sizes				timing (s)			iterations
n	q	m	nm	MOSEK	PDMCF (CPU)	PDMCF (GPU)	
100	10	1178	1×10^5	5	12	5	490
200	10	2316	5×10^5	23	57	6	690
300	10	3472	1×10^6	95	164	6	840
500	10	5738	3×10^6	340	548	7	950
500	20	11 176	6×10^6	1977	890	8	790
1000	10	11 424	1×10^7	2889	18 554	26	7150
1000	20	22 286	2×10^7	16 765	5143	15	1040

Table 4. Runtime table for large size problems (JAX)

problem sizes				timing (s)			iterations
n	q	m	nm	MOSEK	PDMCF (CPU)	PDMCF (GPU)	
3000	10	34 424	1×10^8	OOM	106 274	139	4140
5000	10	57 338	3×10^8	OOM	382 400	421	3970
10 000	10	114 054	1×10^9	OOM	1 809 517	2078	4380

runtime overhead for first-time function compilation and thus it does worse than its PyTorch counterpart on small size problems. The runtimes of these two versions are close for medium and large size problems, with JAX slightly slower.

ACKNOWLEDGEMENTS

We thank Anthony Degleris and Parth Nobel for valuable discussions on implementation details. We also thank Demyan Yarmoshik for very useful feedback on the revision of our original manuscript.

REFERENCES

1. Yin, P., Diamond, S., Lin, B., and Boyd, S., Network Optimization for Unified Packet and Circuit Switched Networks, ArXiv Preprint ArXiv:1808.00586, 2019. https://arxiv.org/abs/1808.00586
2. Bar-Gera, H. and Boyce, D., Origin-Based Algorithms for Combined Travel Forecasting Models, *Transportation Research Part B: Methodological*, 2003, vol. 37, pp. 405–422.
3. Boyd, S. and Vandenberghe, L., *Convex Optimization*, Cambridge: Cambridge University Press, 2004.
4. Mosek ApS. The MOSEK Optimization Toolbox for MATLAB Manual, Version 9.0, 2019. http://docs.mosek.com/9.0/toolbox/index.html
5. Diamond, S. and Boyd, S., CVXPY: A Python-Embedded Modeling Language for Convex Optimization, *Journal of Machine Learning Research*, 2016, vol. 17, pp. 1–5.
6. Ford, Jr. L. and Fulkerson, D., A Suggested Computation for Maximal Multi-Commodity Network Flows, *Management Science*, 1958, vol. 5, pp. 97–101.
7. Hu, T., Multi-Commodity Network Flows, *Operations Research*, 1963, vol. 11, pp. 344–360.
8. Gautier, A. and Granot, F., Forest Management: A Multicommodity Flow Formulation and Sensitivity Analysis, *Management Science*, 1995, vol. 41, pp. 1654–1668.
9. Ouorou, A. and Mahey, P., Minimum Mean Cycle Cancelling Method for Nonlinear Multicommodity Flow Problems, *European Journal of Operational Research*, 2000, vol. 121, pp. 532–548.
10. Manfren, M., Multi-Commodity Network Flow Models for Dynamic Energy Management–Mathematical Formulation, *Energy Procedia*, 2012, vol. 14, pp. 1380–1385.

11. Kabadurmus, O. and Smith, A., Multi-Commodity k-Splittable Survivable Network Design Problems with Relays, *Telecommunication Systems*, 2016, vol. 62, pp. 123–133.

12. Zantuti, A., The Capacity and Non-Simultaneously Multicommodity Flow Problem in Wide Area Network and Data Flow Management, *Proceedings of the 18th International Conference on Systems Engineering*, 2005, pp. 76–80. https://doi.org/10.1109/ICSENG.2005.81

13. Erera, A., Morales, J., and Savelsbergh, M., Global Intermodal Tank Container Management for the Chemical Industry, *Transportation Research Part E: Logistics and Transportation Review*, 2005, vol. 41, pp. 551–566.

14. Mesquita, M., Moz, M., Paias, A., and Pato, M., A Decompose-and-Fix Heuristic Based on Multi-Commodity Flow Models for Driver Rostering with Days-Off Pattern, *European Journal of Operational Research*, 2015, vol. 17, no. 9.

15. Rudi, A., Frohling, M., Zimmer, K., and Schultmann, F., Freight Transportation Planning Considering Carbon Emissions and In-Transit Holding Costs: a Capacitated Multi-Commodity Network Flow Model, *EURO Journal on Transportation and Logistics*, 2016, vol. 5, pp. 123–160. https://api.semanticscholar.org/CorpusID:53229695

16. Singh, I., A Dynamic Multi-Commodity Model of the Agricultural Sector: A Regional Application in Brazil, *European Economic Review*, 1978, vol. 11, pp. 155–179. https://api.semanticscholar.org/CorpusID:152973282

17. Wagner, D., Raidl, G., Pferschy, U., Mutzel, P., and Bachhiesl, P., A Multi-Commodity Flow Approach for the Design of the Last Mile in Real-World Fiber Optic Networks, *Operation Research Proceedings*, 2006, https://api.semanticscholar.org/CorpusID:17037020

18. Layeb, S., Heni, R., and Balma, A., Compact MILP Models for the Discrete Cost Multicommodity Network Design Problem, *Proceedings of the 2017 International Conference on Engineering & MIS*, 2017, pp. 1–7.

19. Salimifard, K. and Bigharaz, S., The Multicommodity Network Flow Problem: State of the Art Classification, Applications, and Solution Methods, *Operational Research*, 2022, vol. 22, pp. 1–47.

20. Ouorou, A., Mahey, P., and Vial, J., A Survey of Algorithms for Convex Multicommodity Flow Problems, *Management Science*, 2000, vol. 46, pp. 126–147.

21. Liu, X., Arzani, B., Kakarla, S., Zhao, L., Liu, V., Castro, M., Kandula, S., and Marshall, L., Rethinking Machine Learning Collective Communication as a Multi-Commodity Flow Problem, *Proceedings of the ACM SIGCOMM 2024 Conference*, 2024, pp. 16–37.

22. Basu, P., Zhao, L., Fantl, J., Pal, S., Krishnamurthy, A., and Khoury, J., Efficient All-To-All Collective Communication Schedules for Direct-Connect Topologies, *Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing*, 2024, pp. 28–41. https://doi.org/10.1145/3625549.3658656

23. Applegate, D., Díaz, M., Hinder, O., Lu, H., Lubin, M., O'Donoghue, B., and Schudy, W., Practical Large-Scale Linear Programming Using Primal-Dual Hybrid Gradient, *Neural Information Processing Systems*, 2021. https://api.semanticscholar.org/CorpusID:235376806

24. Lu, H. and Yang, J., cuPDLP.jl: A GPU Implementation of Restarted Primal-Dual Hybrid Gradient for Linear Programming in Julia, ArXiv Preprint ArXiv:2311.12180, 2024. https://arxiv.org/abs/2311.12180

25. Lu, H., Peng, Z., and Yang, J., MPAX: Mathematical Programming in JAX, ArXiv Preprint ArXiv:2412.09734, 2024. https://arxiv.org/abs/2412.09734

26. Ryu, E., Chen, Y., Li, W., and Osher, S., Vector and Matrix Optimal Mass Transport: Theory, Algorithm, and Applications, *SIAM Journal on Scientific Computing*, 2018, vol. 40, pp. A3675–A3698.

27. Degleris, A., Gamal, A., and Rajagopal, R., GPU Accelerated Security Constrained Optimal Power Flow, ArXiv Preprint ArXiv:2410.17203, 2024. https://arxiv.org/abs/2410.17203

28. Ryu, M., Byeon, G., and Kim, K., A GPU-Accelerated Distributed Algorithm for Optimal Power Flow in Distribution Systems, ArXiv Preprint ArXiv:2501.08293, 2025.

29. Wang, X., Zhang, Q., Ren, J., Xu, S., Wang, S., and Yu, S., Toward Efficient Parallel Routing Optimization for Large-Scale SDN Networks Using GPGPU, *Journal of Network and Computer Applications*, 2018, vol. 113, pp. 1–13.

30. Kikuta, K., Oki, E., Yamanaka, N., Togawa, N., and Nakazato, H., Effective Parallel Algorithm for GPGPU-Accelerated Explicit Routing Optimization, *2015 IEEE Global Communications Conference*, 2015, pp. 1–6.

31. Zhang, S., Ajayi, O., and Cheng, Y., A Self-Supervised Learning Approach for Accelerating Wireless Network Optimization, *IEEE Transactions on Vehicular Technology*, 2023, vol. 72, pp. 8074–8087.

32. Wu, J., He, Z., and Hong, B., Chapter 5 – Efficient CUDA Algorithms for the Maximum Network Flow Problem, *GPU Computing Gems Jade Edition*, 2012, pp. 55–66. https://www.sciencedirect.com/science/article/pii/B9780123859631000058

33. Liu, H., Huang, S., Qin, S., Yang, T., Yang, T., Xiang, Q., and Liu, X., Keep Your Paths Free: Toward Scalable Learning-Based Traffic Engineering, *Proceedings of the 8th Asia-Pacific Workshop on Networking*, 2024, pp. 189–191. https://doi.org/10.1145/3663408.3665813

34. Gurobi Optimization. LLC Gurobi Optimizer Reference Manual, 2024. https://www.gurobi.com

35. Yarmoshik, D. and Persiianov, M., On the Application of Saddle-Point Methods for Combined Equilibrium Transportation Models, *Proceedings of the 23rd International Conference on Mathematical Optimization Theory and Operations Research (MOTOR 2024)*, 2024, pp. 432–448. https://doi.org/10.1007/978-3-031-62792-7_29

36. Kubentayeva, M., Yarmoshik, D., Persiianov, M., Kroshnin, A., Kotliarova, E., Tupitsa, N., Pasechnyuk, D., Gasnikov, A., Shvetsov, V., Baryshev, L., and Shurupov, A., Primal-Dual Gradient Methods for Searching Network Equilibria in Combined Models with Nested Choice Structure and Capacity Constraints, ArXiv Preprint ArXiv:2307.00427, 2023. https://arxiv.org/abs/2307.00427

37. Malitsky, Y. and Pock, T., A First-Order Primal-Dual Algorithm with Linesearch, *SIAM Journal on Optimization*, 2018, vol. 28, pp. 411–432.

38. Zhu, M. and Chan, T., An Efficient Primal-Dual Hybrid Gradient Algorithm for Total Variation Image Restoration, *UCLA CAM Report*, 2008, vol. 34.

39. Chambolle, A. and Pock, T., A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging, *Journal of Mathematical Imaging and Vision*, 2011, vol. 40, pp. 120–145.

40. Chambolle, A. and Pock, T., On the Ergodic Convergence Rates of a First-Order Primal–Dual Algorithm, *Mathematical Programming*, 2015, vol. 159, pp. 253–287.

41. Parikh, N. and Boyd, S., Proximal Algorithms, *Foundations and Trends in Optimization*, 2014, vol. 1, pp. 127–239.

42. Held, M., Wolfe, P., and Crowder, H., Validation of Subgradient Optimization, *Mathematical Programming*, 1974, vol. 6, pp. 62–88.

43. Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T., Efficient Projections onto the l_1-Ball for Learning in High Dimensions, *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 272–279. https://doi.org/10.1145/1390156.1390191

44. Condat, L., Fast Projection onto the Simplex and the l_1 Ball, *Mathematical Programming*, 2016, vol. 158, pp. 575–585. https://doi.org/10.1007/s10107-015-0946-6

This paper was recommended for publication by P.S. Shcherbakov, a member of the Editorial Board