

Open Journal of Mathematical Optimization

Yannis Bekri, Anatoli Juditsky & Arkadi Nemirovski

First order algorithms for computing linear and polyhedral estimates

Volume 5 (2024), article no. 7 (15 pages)

<https://doi.org/10.5802/ojmo.35>

Article submitted on November 22, 2023, revised on May 7, 2024,
accepted on August 7, 2024.

© The author(s), 2024.



This article is licensed under the

CREATIVE COMMONS ATTRIBUTION 4.0 INTERNATIONAL LICENSE.

<http://creativecommons.org/licenses/by/4.0/>



First order algorithms for computing linear and polyhedral estimates

Yannis Bekri

LJK, Université Grenoble Alpes, Campus de Saint-Martin-d'Hères, 38401 France
yannis.bekri@univ-grenoble-alpes.fr

Anatoli Juditsky

LJK, Université Grenoble Alpes, Campus de Saint-Martin-d'Hères, 38401 France
anatoli.juditsky@univ-grenoble-alpes.fr

Arkadi Nemirovski

Arkadi Nemirovski, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
nemirovs@isye.gatech.edu

Abstract

It was recently shown [6, 8] that “properly built” linear and polyhedral estimates nearly attain minimax accuracy bounds in the problem of recovery of unknown signal from noisy observations of linear images of the signal when the signal set is an *elliptope*. However, design of nearly optimal estimates relies upon solving semidefinite optimization problems with matrix variables, what puts the synthesis of such estimates beyond the reach of the standard Interior Point algorithms of semidefinite optimization even for moderate size recovery problems. Our goal is to develop First Order Optimization algorithms for the computationally efficient design of linear and polyhedral estimates. In this paper we (a) explain how to eliminate matrix variables, thus reducing dramatically the design dimension when passing from Interior Point to First Order optimization algorithms and (b) develop and analyse a dedicated algorithm of the latter type — Composite Truncated Level method.

Digital Object Identifier 10.5802/ojmo.35

Acknowledgments This work was supported by Multidisciplinary Institute in Artificial intelligence MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

1 Introduction

In this paper we discuss numerical algorithms for construction of “presumably good” estimates in linear inverse problems. Specifically, consider the estimation problem as follows. We are given an observation $\omega \in \mathbf{R}^m$,

$$\omega = Ax + \xi \tag{1}$$

where $\xi \in \mathbf{R}^m$ is a zero mean random noise, $A \in \mathbf{R}^{m \times n}$ is the observation matrix, and x is an unknown signal known to belong to a given convex set $\mathcal{X} \subset \mathbf{R}^n$. Our objective is to recover the linear image $w = Bx$, $B \in \mathbf{R}^{\nu \times n}$, of x .

Our focus is on *linear* and *polyhedral estimates* for solving the problem in question.

When applied to the estimation problem above, *linear estimate* $\hat{w}_{\text{lin}}^H(\omega)$ of w is of the form $\hat{w}_{\text{lin}}^H(\omega) = H^T \omega$ where *contrast matrix* $H \in \mathbf{R}^{m \times \nu}$ is the estimate’s parameter. A *polyhedral estimate* $\hat{w}_{\text{poly}}^H(\omega)$ is specified by a *contrast matrix* $H \in \mathbf{R}^{m \times M}$ according to

$$\omega \mapsto \hat{x}^H(\omega) \in \text{Argmin}_{x \in \mathcal{X}} \{ \|H^T(\omega - Ax)\|_{\infty} \}, \quad \hat{w}_{\text{poly}}^H(\omega) := B\hat{x}^H(\omega).$$

Our interest in these two types of estimates stems from the fact that, as it was shown in [6, 8, 9], in the Gaussian case ($\xi \sim \mathcal{N}(0, \sigma^2 I_m)$), linear and polyhedral estimates with properly designed efficiently computable contrast matrices are near-minimax optimal in terms of their statistical risks over a rather general class of loss functions and signal sets which we call *ellitopes*.¹ In this paper, our goal is to investigate numerical algorithms for design of near-optimal linear and polyhedral estimates. Specifically, we aim at developing numerical routines for efficient computation of contrast matrices H , the principal parameters of the estimates of both types.



© Yannis Bekri & Anatoli Juditsky & Arkadi Nemirovski;
licensed under Creative Commons License Attribution 4.0 International

OJMO Volume 5 (2024), article no. 7

¹ Exact definitions of these sets are reproduced in the main body of the paper. For the time being, it suffices to point out an instructive example: a bounded intersections of finitely many sets of the form $\{x : \|Px\|_p \leq 1\}$, $p \geq 2$, is an *ellitope*.

As it was shown in [6, 8, 9], given the problem data, matrices A, B , the signal ellitope \mathcal{X} , and the *co-ellitopic norm* $\|\cdot\|$ in which estimation error is measured, computing the contrast matrices of linear and polyhedral estimates amounts to solving a well-structured convex optimization problem with linear objective and linear matrix inequality constraints. State-of-the-art optimization software, e.g., CVX [4] which relies upon Interior Point Semidefinite Programming (SDP) algorithms may be used to compute high-accuracy solutions to these problems. However, the structure of the optimization problems in question (presence of “dense” matrix arguments) results in prohibitively long processing times by IPM algorithms already for rather moderate problem dimensions (signal dimension n in the range of few dozens). In this paper we discuss an alternative approach to solving the problem of designing linear and polyhedral estimates which relies upon a first order algorithm, namely, the Composite Truncated Level method (CTL) of the bundle family. In particular, we show how matrix arguments can be eliminated from the contrast optimization problem and how the latter can be cast in the form amenable for first order algorithms.

The paper is organized as follows. In Section 2 we present the precise setting of the estimation problem and define optimization problems underlying the contrast design for linear and polyhedral estimates. To set up the first order optimization algorithm, we demonstrate how the problem of contrast computation for linear estimate can be reduced to that for the polyhedral one, and then explain in Section 3 how the latter problem can be rewritten in the form not involving matrix arguments and convenient for solving using first order algorithms. A small simulation study presented in Section 3.2 illustrates numerical performance of the proposed algorithms. We present the details of the Composite Truncated Level algorithm in Section A of the appendix. Proofs of technical statements are put to Section B.

2 Linear and polyhedral estimates

2.1 Estimation problem

Consider the problem of recovering linear image $w = Bx$ of unknown signal $x \in \mathbf{R}^n$ from noisy observation

$$\omega = Ax + \sigma\xi \quad (2)$$

where $B \in \mathbf{R}^{\nu \times n}$ and $A \in \mathbf{R}^{m \times n}$ are given matrices, $\sigma > 0$ is known, and x is known to belong to a given signal set \mathcal{X} . Throughout the paper ξ is $(0, I_m)$ -sub-Gaussian, denoted $\xi \sim \mathcal{SG}(0, I_m)$, i.e., for all $t \in \mathbf{R}^m$,

$$\mathbf{E} \left\{ e^{t^T \xi} \right\} \leq \exp \left(\frac{1}{2} \|t\|_2^2 \right). \quad (3)$$

Given a norm $\|\cdot\|$ on \mathbf{R}^ν and reliability tolerance $\epsilon \in (0, 1)$, we quantify the performance of a candidate estimate $\hat{w}(\cdot)$ by its ϵ -risk

$$\text{Risk}_\epsilon[\hat{w}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \inf_{\rho} \{ \rho : \text{Prob}_\xi \{ \|\hat{w}(Ax + \xi) - Bx\| > \rho \} \leq \epsilon \}. \quad (4)$$

We assume from now on that the signal set \mathcal{X} and the *polar* \mathcal{B}_* of the unit ball of $\|\cdot\|$ are basic ellitopes (see, e.g., [7] and [9, Section 4.2]). Specifically, we set

$$\begin{aligned} \mathcal{X} &= \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^T T_k x \leq t_k, k \leq K\}, \\ \mathcal{B}_* &= \{y \in \mathbf{R}^\nu : \exists s \in \mathcal{S} : y^T S_\ell y \leq s_\ell, \ell \leq L\}. \end{aligned} \quad (5)$$

Here $T_k \geq 0$ with $\sum_k T_k \succ 0$ (respectively, $S_\ell \geq 0$ with $\sum_\ell S_\ell \succ 0$, and $\mathcal{T} \subset \mathbf{R}_+^K$ (respectively, $\mathcal{S} \subset \mathbf{R}_+^L$) is a convex compact set which is monotone (i.e., $0 \leq t \leq t' \in \mathcal{T} \Rightarrow t' \in \mathcal{T}$)² and possesses a nonempty interior. We refer to K (respectively, to L) as *ellitopic dimension* of \mathcal{X} (respectively, of \mathcal{B}_*).

Every basic ellitope is a convex compact set with nonempty interior which is symmetric w.r.t. the origin. “Standard” examples of basic ellitopes are:

- A bounded intersection \mathcal{X} of K centered at the origin ellipsoids/elliptic cylinders $\{x \in \mathbf{R}^n : x^T T_k x \leq 1\}$ [$T_k \geq 0$]:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} := [0, 1]^K : x^T T_k x \leq t_k, k \leq K\}$$

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

² Here and in the sequel, relationships $t \leq s$ (or $t < s$) between $t, s \in \mathbf{R}^K$ are understood entrywise, i.e., as $t_i \leq s_i$ (respectively, as $t_i < s_i$), $i = 1, \dots, K$.

■ A $\|\cdot\|_p$ -ball in \mathbf{R}^n with $p \in [2, \infty)$:

$$\{x \in \mathbf{R}^n : \|x\|_p \leq 1\} = \left\{ x : \exists t \in \mathcal{T} = \{t \in \mathbf{R}_+^n, \|t\|_{p/2} \leq 1\} : \underbrace{x_k^2}_{x^T T_k x} \leq t_k, k \leq n \right\}.$$

2.2 The estimates

The interest in ellitopes in the present context is motivated by the results of [7, 8, 9] which state that, in the situation in question, “near-optimal” (with risks within logarithmic in K and L factors from the minimax risk) estimates can be found among *linear* and *polyhedral* ones.

2.2.1 Linear estimate

Linear estimate is specified by an $m \times \nu$ contrast matrix H according to

$$\hat{w}_H(\omega) = H^T \omega.$$

Let

$$\mathfrak{r}_\varkappa[H] := \min_{\lambda, \mu, \Theta} \left\{ \phi_{\mathcal{T}}(\mu) + \phi_{\mathcal{S}}(\lambda) + \sigma^2 \varkappa^2 \text{Tr}(\Theta) : \begin{array}{c|c|c} \lambda \geq 0, \mu \geq 0, & & \\ \hline \frac{\sum_{\ell} \lambda_{\ell} S_{\ell}}{\frac{1}{2}(B - H^T A)^T} & \frac{\frac{1}{2}(B - H^T A)}{\sum_k \mu_k T_k} & \frac{\frac{1}{2}H^T}{\Theta} \\ \hline \frac{1}{2}H & & \end{array} \succeq 0 \right\} \quad (6)$$

where for $\mathcal{G} \subset \mathbf{R}^p$

$$\phi_{\mathcal{G}}(z) = \sup_{g \in \mathcal{G}} z^T g : \mathbf{R}^p \rightarrow \mathbf{R} \cup \{+\infty\}$$

is the support function of \mathcal{G} .

Proposition 1 (cf. [9, Proposition 4.14]). *Let $\hat{w}_{\text{lin}}^H(\omega) = H^T \omega$ with some $H \in \mathbf{R}^{m \times \nu}$.*

i. *Then*

$$\sup_{x \in \mathcal{X}} \mathbf{E} \{ \|\hat{w}_{\text{lin}}^H(\omega) - Bx\| \} \leq \mathfrak{r}_1[H]$$

ii. *Furthermore, let*

$$\varkappa = 1 + \sqrt{2 \ln[\epsilon^{-1}]}.$$

Then

$$\text{Risk}_{\epsilon}[\hat{w}_{\text{lin}}^H | \mathcal{X}] \leq \mathfrak{r}_{\varkappa}[H]. \quad (7)$$

Furthermore, function $\mathfrak{r}_{\varkappa}[H]$ is a convex, continuous and coercive function of the contrast matrix, and can be efficiently minimized w.r.t. H .

Remarks.

■ As a consequence of the statement i of the proposition, the optimal value \mathfrak{r}^* of the (clearly solvable) convex optimization problem

$$\mathfrak{r}^* := \min_H \mathfrak{r}_1[H] = \min_{H, \lambda, \mu, \Theta} \left\{ \phi_{\mathcal{T}}(\mu) + \phi_{\mathcal{S}}(\lambda) + \sigma^2 \text{Tr}(\Theta) : \begin{array}{c|c|c} \lambda \geq 0, \mu \geq 0, & & \\ \hline \frac{\sum_{\ell} \lambda_{\ell} S_{\ell}}{\frac{1}{2}(B - H^T A)^T} & \frac{\frac{1}{2}(B - H^T A)}{\sum_k \mu_k T_k} & \frac{\frac{1}{2}H^T}{\Theta} \\ \hline \frac{1}{2}H & & \end{array} \succeq 0 \right\} \quad (8)$$

is an upper bound on the expected risk

$$\text{Risk}[\hat{w}_{\text{lin}}^* | \mathcal{X}] = \sup_{x \in \mathcal{X}} \mathbf{E} \{ \|\hat{w}_{\text{lin}}^*(Ax + \xi) - Bx\| \}$$

of the estimate $\widehat{w}_{\text{lin}}^*(\omega) = H_*^T \omega$ yielded by the H -component H_* of an optimal solution to the problem. Note that we do not need to assume that the noise ξ is sub-Gaussian for the above bound to hold: it would suffice to suppose that $\mathbf{E}\{\xi\xi^T\} \preceq \sigma^2 I$. Moreover (cf. [9, Proposition 4.16]), the value \mathfrak{r}^* is within moderate factor of the minimax-optimal $\|\cdot\|$ -risk

$$\text{RiskOpt}[\mathcal{X}] = \inf_{\widehat{w}(\cdot)} \text{Risk}[\widehat{w}|\mathcal{X}]$$

(here inf is taken over all estimates, linear and nonlinear alike):

$$\mathfrak{r}^* \leq O(1) \sqrt{\ln(K+1) \ln(L+1)} \text{RiskOpt}[\mathcal{X}] \quad (9)$$

(from now on, $O(1)$'s stands for an absolute constant).

- Observe that the ϵ -risk of the estimate $\widehat{w}_{\text{lin}}^*$ may be evaluated using the statement ii of Proposition 1.

Corollary 2. *Let $\epsilon \in (0, 1)$. The ϵ -risk of the estimate $\widehat{w}_{\text{lin}}^*(\omega)$ satisfies*

$$\text{Risk}_\epsilon[\widehat{w}_{\text{lin}}^*|\mathcal{X}] \leq \left(1 + \sqrt{2 \ln[\epsilon^{-1}]}\right) \mathfrak{r}^*.$$

- The bound $\bar{\psi}_\epsilon(\Theta) = \varkappa^2 \text{Tr}(\Theta)$ for the $1 - \epsilon$ quantile of the quadratic form $\xi^T \Theta \xi$ in the expression (6) can be replaced by the following tighter but harder to process bounds (cf, e.g., [1, Proposition A.1])

$$\begin{aligned} \psi_\epsilon(\Theta) &:= \min_{\alpha} \left\{ -\frac{\alpha}{2} \log \text{Det}(I - 2\alpha^{-1} \Theta) + \alpha \ln[\epsilon^{-1}], \alpha \geq 2\lambda_{\max}(\Theta) \right\} \\ &\leq \psi'_\epsilon(\Theta) := \min_{\alpha} \left\{ \text{Tr}(\Theta) + \text{Tr}(\Theta(\alpha I - 2\Theta)^{-1} \Theta) + \alpha \ln(\epsilon^{-1}) : \alpha \geq 2\lambda_{\max}(\Theta) \right\} \\ &\leq \tilde{\psi}_\epsilon(\Theta) := \text{Tr}(\Theta) + 2\|\Theta\|_{\text{Fro}} \sqrt{\ln[\epsilon^{-1}]} + 2\lambda_{\max}(\Theta) \ln[\epsilon^{-1}] \leq \bar{\psi}_\epsilon(\Theta). \end{aligned} \quad (10)$$

2.2.2 Polyhedral estimate

Polyhedral estimate is specified by $m \times m$ contrast matrix H satisfying

$$\sigma\chi \|\text{Col}_j[H]\|_2 \leq 1, \quad 1 \leq j \leq m, \quad (11)$$

with

$$\chi = [\chi(\epsilon/m)] = \sqrt{2 \ln[2\epsilon/m]};$$

in other words, H is normalized by the requirement that

$$\text{Prob}_{\xi \sim \mathcal{SG}(0, I)} \{ \sigma \|H^T \xi\|_\infty > 1 \} \leq \epsilon. \quad (12)$$

The associated with H polyhedral estimate $\widehat{w}_{\text{poly}}^H(\omega)$ is

$$\widehat{w}_{\text{poly}}^H(\omega) = B\bar{x}(\omega), \quad \bar{x}(\omega) = \text{argmin}_x \{ \|H^T(\omega - Ax)\|_\infty : x \in \mathcal{X} \} \quad (13)$$

It is shown in [9, Setion 5.1.5] that the ϵ -risk of $\widehat{w}_{\text{poly}}^H$ is upper-bounded by the quantity

$$\min_{\lambda, \mu, v} \left\{ 2 \left[\phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sum_j v_j \right] : \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2} B \\ \hline \frac{1}{2} B^T & A^T H \text{Diag}\{v\} H^T A + \sum_k \mu_k T_k \end{array} \right] \succeq 0 \right\} \quad (P[H])$$

A presumably good synthesis of the contrast H is yielded by feasible solutions to the convex optimization problem

$$\mathfrak{p}_\chi^* = \min_{\Theta} \left\{ \mathfrak{p}_\chi[\Theta] := \min_{\lambda, \mu} \left\{ 2 \left[\phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2(\epsilon/m) \text{Tr}(\Theta) \right] : \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2} B \\ \hline \frac{1}{2} B^T & A^T \Theta A + \sum_k \mu_k T_k \end{array} \right] \succeq 0 \right\} \right\} \quad (14)$$

Given a feasible solution (λ, μ, Θ) to (14), we set $H = [\sigma\chi(\epsilon/m)]^{-1} U$ where $\Theta = U \text{Diag}\{\nu\} U^T$ is the eigenvalue decomposition of Θ and $v = [\sigma\chi(\epsilon/m)]^2 \nu$. Note that such H satisfies (11) and, moreover, (λ, μ, v, H) is a

feasible solution to $(P[H])$ with values of respective objectives of both problems at these feasible solutions being equal to each other. As a result, the ϵ -risk of the polyhedral estimate \hat{w}_{poly} stemming, in the just explained fashion, from an optimal solution to the (clearly solvable) problem (14) is upper-bounded by \mathfrak{p}_χ^* . As shown in [9, Proposition 5.10], the resulting polyhedral estimate is nearly minimax-optimal:

$$\text{Risk}_\epsilon[\hat{w}_{\text{poly}}|\mathcal{X}] \leq \mathfrak{p}_\chi^* \leq O(1)\sqrt{\ln(K+1)\ln(L+1)\ln(2m/\epsilon)} \text{RiskOpt}_\epsilon[\mathcal{X}]$$

where $\text{RiskOpt}_\epsilon[\mathcal{X}]$ is the minimax ϵ -risk.

2.3 From polyhedral to linear estimate and back

Observe that problems (8) and (14) responsible for the design of nearly minimax-optimal under the circumstances linear and polyhedral estimates are well structured convex problems. State-of-the-art Interior Point Semidefinite Programming (SDP) algorithms may be used to compute high-accuracy solutions to these problems in a wide range of geometries of \mathcal{T} and \mathcal{S} . However, the presence of matrix variables H and Θ results in large design dimensions of the SDP's to be solved and make prohibitively time consuming processing problem instances of sizes m, n in the range of hundreds. The first goal of this paper is to show that matrix variables may be eliminated from (8) and (14) allowing for processing by dedicated First Order algorithms, resulting in significant extension of the ranges of problem sizes amenable for numerical processing.

Our first observation is that problems (8) and (14) are “nearly reducible” to each other. Indeed, let $(\lambda, \mu, H, \Theta)$ be feasible to (8). We clearly have $\Theta \succeq 0$. Let $G = \begin{bmatrix} I & & \\ & I & \\ & & A^T \end{bmatrix}$. By multiplying the semidefinite constraint of (8) by G on the left and G^T on the right we conclude that (λ, μ, Θ) is a feasible solution to (14) with the corresponding objective value

$$2[\phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta)] = 2\mathfrak{r}_\chi[H].$$

The converse is also true.

Lemma 3. *Let (λ, μ, Θ) be a feasible solution to (14). Then it can be augmented to the feasible solution $(2\lambda, \mu, H, \Theta)$ of (8) with the corresponding objective value*

$$2\phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta) \leq \mathfrak{p}_\chi[\Theta].$$

3 Designing polyhedral estimates by a First Order method

According to the results from the previous section, when speaking about numerical design of linear and polyhedral estimates, we can focus solely on solving problem (14).³ Next, projecting, if necessary, the observation ω onto the image space of A , we can assume w.l.o.g. that $m \leq n$ and the image space of A is the entire \mathbf{R}^m . In fact, we make here a stronger assumption:⁴

Assumption 0. Matrix $A \in \mathbf{R}^{n \times n}$ in (2) is nonsingular.

Under this assumption, we can carry out partial minimization in Θ in (14). Specifically, it is immediately seen that (14) is equivalent to the optimization problem

$$\min_{\lambda, \mu, \Theta} \left\{ \phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta) : \lambda > 0, \mu \geq 0, \Theta \succeq 0, \Theta \succeq \underbrace{A^{-T} \left[\frac{1}{4} B^T \left[\sum_{\ell} \lambda_{\ell} S_{\ell} \right]^{-1} B - \sum_k \mu_k T_k \right] A^{-1}}_{\mathfrak{T}(\lambda, \mu)} \right\} \quad (15)$$

In the latter problem partial minimization in Θ is as follows: given $\lambda > 0$ and $\mu \geq 0$ we compute the eigenvalue decomposition

$$\mathfrak{T}(\lambda, \mu) = U \text{Diag}\{v\} U^T$$

³ Strictly speaking, this is so if we assume that when looking for a linear estimate, we are ready to tolerate a moderate constant factor (namely, 2) in the risk bound of the resulting estimate.

⁴ We briefly describe the “conversion” of (14) into the form amenable for First Order algorithms in the case of singular A in Section C.

of $\mathfrak{T}(\lambda, \mu)$. The best in terms of the objective of (15) choice of Θ given λ, μ is

$$\Theta = U \text{Diag}\{v^+\}U^T. \quad ([\alpha^+ = \max[\alpha, 0]])$$

Therefore, (15) reduces to

$$\begin{aligned} \min_{\lambda > 0, \mu \geq 0} \left\{ \Upsilon(\lambda, \mu) := \phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \sum_i \lambda_i^+(\mathfrak{T}(\lambda, \mu)) \right\} \\ \mathfrak{T}(\lambda, \mu) = A^{-T} \left[\frac{1}{4} B^T \left[\sum_{\ell} \lambda_{\ell} S_{\ell} \right]^{-1} B - \sum_k \mu_k T_k \right] A^{-1} \end{aligned} \quad (16)$$

where $\lambda_1(Q) \geq \lambda_2(Q) \geq \dots \geq \lambda_p(Q)$ are the eigenvalues of symmetric $p \times p$ matrix Q . and $\lambda_i^+(Q) = \max[\lambda_i(Q), 0]$.

3.1 Setting up Composite Truncated Level algorithm

We intend to solve the problem of interest (14) by applying to (16) a First Order algorithm — Composite Truncated Level algorithm (CTL). Detailed description of the method is presented in Section A. CTL is aimed at solving convex optimization problems of the form

$$\text{Opt} = \min_{x \in X} \{ \phi(x) := \psi(x) + f(x) \} \quad (17)$$

where $X \subset \mathbf{R}^N$ is a nonempty bounded and closed convex set, and $\psi(\cdot)$ and $f(\cdot)$ are Lipschitz continuous convex functions on X with “simple” X and ψ (for details, see Section A). Note that in order to reduce the problem of interest (16) to the form (17), it suffices to set

$$\begin{aligned} X &= \left\{ x = [\lambda; \mu] \in \mathbf{R}_+^{L+K}, \lambda_{\ell} \geq \delta \ \forall \ell, \sum_{\ell} \lambda_{\ell} + \sum_k \mu_k \leq R \right\}, \\ \psi([\lambda; \mu]) &= \phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu), \\ f([\lambda; \mu]) &= \sigma^2 \chi^2 \sum_i \lambda_i^+(\mathfrak{T}(\lambda, \mu)). \end{aligned} \quad (18)$$

When solving (17), CTL “learns” the difficult part $f(x)$ of the objective via *oracle* which, given on input a query point $\bar{x} \in X$, returns a “simple” Lipschitz continuous convex function (model) $f_{\bar{x}}(\cdot)$ such that

$$f_{\bar{x}}(\bar{x}) = f(\bar{x}) \ \& \ f_{\bar{x}}(y) \leq f(y) \quad \forall y \in X.$$

Oracle \mathcal{O}_{ρ}

In the situation we are interested in with the data for (17) given by (18), the oracle may be built as follows:

1. Given query point $\bar{x} = [\bar{\lambda}; \bar{\mu}] \in X$, we compute the matrices $\bar{\Lambda} = \sum_{\ell} \bar{\lambda}_{\ell} S_{\ell}$ and $\bar{\mathfrak{T}} = \mathfrak{T}(\bar{\lambda}, \bar{\mu})$ along with the eigenvalue decomposition $\bar{\mathfrak{T}} = \bar{U} \text{Diag}\{\bar{v}\} \bar{U}^T$, $\bar{v}_1 \geq \bar{v}_2 \geq \dots \geq \bar{v}_n$ of $\bar{\mathfrak{T}}$.
2. We put

$$T(\lambda, \mu) = A^{-T} \left[\frac{1}{4} B^T \bar{\Lambda}^{-1} \left[2\bar{\Lambda} - \sum_{\ell} \lambda_{\ell} S_{\ell} \right] \bar{\Lambda}^{-1} B - \sum_k \mu_k T_k \right] A^{-1}$$

This function is obtained from $\mathfrak{T}(\lambda, \mu)$ by linearization in λ at $\lambda = \bar{\lambda}$ and clearly \succeq -underestimates $\mathfrak{T}(\lambda, \mu)$ on X , while $T(\bar{\lambda}, \mu) \equiv \mathfrak{T}(\bar{\lambda}, \mu)$. Consequently,

$$f([\lambda; \mu]) = \sigma^2 \chi^2 \sum_i \lambda_i^+(\mathfrak{T}(\lambda, \mu)) \geq f_{\bar{\lambda}}(\lambda, \mu) := \sigma^2 \chi^2 \sum_i \lambda_i^+(T(\lambda, \mu)),$$

the inequality being equality when $\lambda = \bar{\lambda}$.

Recall that for every symmetric convex function g on \mathbf{R}^n and every $n \times n$ symmetric matrix D one has [2, Proposition 4.2.1]

$$g([D_{11}; \dots; D_{nn}]) \leq g(\lambda(D)).$$

When specifying $g(s) = \sum_{i=1}^n s_i^+$ and denoting by $D_i(\lambda, \mu)$ the diagonal entries in the matrix $\bar{U}^T T(\lambda, \mu) \bar{U}$, $1 \leq i \leq n$, we get

$$[f([\lambda; \mu]) \geq] f_{\bar{\lambda}}(\lambda, \mu) \geq \sum_{i=1}^n D_i^+(\lambda, \mu) \quad \forall (\lambda > 0, \mu \geq 0)$$

with both inequalities becoming equalities for $\lambda = \bar{\lambda}$ and $\mu = \bar{\mu}$. Taking into account that functions $D_i(\lambda, \mu)$ are affine, we conclude that *piecewise linear function* $\sum_i D_i^+(\lambda, \mu)$ *underestimates* $f([\lambda; \mu])$ *in the domain* $\lambda > 0$ *and is equal to* $f([\lambda; \mu])$ *when* $(\lambda, \mu) = (\bar{\lambda}, \bar{\mu})$.

3. The above considerations justify the oracle \mathcal{O}_ϱ defined as follows:

$$f_{\bar{x}=[\bar{\lambda}; \bar{\mu}]}(x) = \sum_{\iota=1}^{\varrho} \max[\alpha_\iota(x), 0],$$

where ϱ , $1 \leq \varrho \leq n$, is the ‘‘complexity parameter’’ of the oracle, and $\alpha_\iota(x)$ are affine functions of $x = [\lambda; \mu]$ specified according to

- for $\iota < \varrho$, $\alpha_\iota(x) = D_\iota(x)$;
- $\alpha_\varrho(x) = \sum_{\iota \geq \varrho} \bar{D}_\iota(x)$, $\bar{D}_\iota(x) = \begin{cases} D_\iota(x), & D_\iota(\bar{x}) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$

By construction, $f_{\bar{x}}(x)$ is the sum of ϱ positive parts of linear forms, underestimates $f(x)$ everywhere, and coincides with $f(x)$ when $x = \bar{x}$.

3.2 Numerical illustration

Consider the situation in which $\|\cdot\|$ is $\|\cdot\|_2$. In this case problem (14) reads

$$\mathbf{p}_\chi^* = \min_{\lambda, \mu, \Theta} \left\{ \begin{array}{l} \lambda \geq 0, \mu \geq 0, \Theta \succeq 0 \\ 2[\lambda + \phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta)] : \left[\begin{array}{c|c} \lambda I_\nu & \frac{1}{2} B \\ \hline \frac{1}{2} B^T & A^T \Theta A + \sum_k \mu_k T_k \end{array} \right] \succeq 0 \end{array} \right\} \quad (19)$$

Observe that scaling a feasible solution (λ, μ, Θ) to the problem according to $(\lambda, \mu, \Theta) \mapsto (s\lambda, s^{-1}\mu, s^{-1}\Theta)$ with $s > 0$ preserves feasibility; the best in terms of the objective scaling of (λ, μ, Θ) corresponds to $s = \sqrt{[\phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta)]/\lambda}$ and results in the value of the objective $4\sqrt{\lambda[\phi_{\mathcal{T}}(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta)]}$. As a result, we can eliminate the variable λ , thus arriving at the problem

$$\bar{\mathbf{p}}_\chi^* = \min_{\bar{\mu}, \bar{\Theta}} \left\{ \begin{array}{l} \bar{\mu} \geq 0, \bar{\Theta} \succeq 0 \\ F(\bar{\mu}, \bar{\Theta}) := \phi_{\mathcal{T}}(\bar{\mu}) + \sigma^2 \chi^2 \text{Tr}(\bar{\Theta}) : \left[\begin{array}{c|c} I_\nu & \frac{1}{2} B \\ \hline \frac{1}{2} B^T & A^T \bar{\Theta} A + \sum_k \bar{\mu}_k T_k \end{array} \right] \succeq 0 \end{array} \right\} \quad (20)$$

A feasible (an optimal) solution $\bar{\mu}, \bar{\Theta}$ to (20) gives rise to feasible (resp., optimal) solution $\lambda = \sqrt{F(\bar{\mu}, \bar{\Theta})}$, $\mu = \bar{\mu}/\lambda$, $\Theta = \bar{\Theta}/\lambda$ to (19) with the value of the objective equal to $4\sqrt{F(\bar{\mu}, \bar{\Theta})}$; in particular,

$$\mathbf{p}_\chi^* = 4\sqrt{\bar{\mathbf{p}}_\chi^*}.$$

In our experiments, we used $B = I_n$, $A \in \mathbf{R}^{n \times n}$ with i.i.d. entries drawn from $\mathcal{N}(0, 1)$, we put $\sigma = 0.1$, $\epsilon = 0.05$, and used ellitopic signal set

$$\mathcal{X} = \{x \in \mathbf{R}^n : \sum_{i \in I_k} ix_i^2 \leq 1, 1 \leq i \leq K\}$$

where I_1, \dots, I_K are consecutive segments of the range $1 \leq i \leq n$ of cardinality n/K each.

Under the circumstances, problem (20) reads

$$\bar{\mathbf{p}}_\chi^* = \min_{\bar{\mu}, \bar{\Theta}} \left\{ \begin{array}{l} \sum_{k=1}^K \bar{\mu}_k + \underbrace{\sigma^2 \chi^2 (\epsilon/n)}_\gamma \text{Tr}(\bar{\Theta}) : \bar{\mu} \geq 0, \bar{\Theta} \succeq 0 \\ \left[\begin{array}{c|c} I_\nu & \frac{1}{2} I_n \\ \hline \frac{1}{2} I_n & A^T \bar{\Theta} A + D[\bar{\mu}] \end{array} \right] \succeq 0 \end{array} \right\} \quad (21)$$

■ **Table 1** Solving (22) by CTL with $\varrho = 10, \tau = 10$.

n	64	128	256	512	1024	1024
K	8	16	32	64	128	1024
calls	23	8	10	28	24	26
phases	11	4	5	14	11	11
CPU, sec	7	2	4	20	96	781
0.05-risk	3.215	4.283	4.519	4.169	4.216	7.868
$\ \cdot\ _2$ -risk	1.709	2.237	2.389	2.209	2.231	4.271

■ **Table 2** Performance of CTL vs. ϱ and τ , problem (22) with $n = 1024, K = 128$. Data in cells: # of calls/# of phases/CPU time, sec.

	$\varrho=1$	$\varrho=10$
$\tau=1$	50/11/275	31/12/136
$\tau=10$	26/11/109	24/11/96

where $D[\bar{\mu}]$ is diagonal $n \times n$ matrix with the i -th diagonal entry equal to $i\mu_k$ when $i \in I_k$.

After eliminating Θ by partial minimization (21) becomes

$$\min_{\bar{\mu}} \left\{ \sum_k \bar{\mu}_k + \gamma \sum_l \lambda_l^+ (A^{-T} [I_n - D[\bar{\mu}]] A^{-1}) : 0 \leq \bar{\mu}, \sum_k \bar{\mu}_k \leq R \right\} \quad (22)$$

(we have imposed a large enough upper bound on $\sum_k \bar{\mu}_k$ to make the optimization domain bounded). The resulting problem was processed by the CTL algorithm utilizing oracle \mathcal{O}_ϱ . ϱ was the first of the two control parameters used the experiments; the second parameter was the maximum cardinality τ of bundle allowed for CTL.⁵ We used “ ℓ_1/ℓ_2 proximal setup,” [11, Section 2.1], in which

$$\|\cdot\| = \|\cdot\|_1, \quad \omega(\bar{\mu}) = \kappa_n \|\mu\|_{p_n}^2, \quad p_n = 1 + 1/\ln n,$$

where κ_n is an easy to compute constant ensuring strong convexity, modulus 1, of $\omega(\cdot)$ w.r.t. $\|\cdot\|_1$ (see [11, Theorem 2.1]). The CTL parameters $\lambda_\ell, \bar{\theta}, \underline{\theta}$ were set to $1/2$, and the auxiliary problems (steps 4.2-4.3) were processed by Interior Point solver **Mosek** invoked via **CVX**, see [4]. When solving (22), computations were terminated when the best found so far value of the objective were within the factor 1.1 of the generated by the method lower bound on the optimal value.

We report on results of our experiments in Tables 1 and 2.⁶ To put these results in proper perspective, note that solving (21) by state-of-the-art **Mosek** Interior Point solver takes 35 sec when $n = 64, K = 8$ and 1785 sec when $n = 128, K = 16$; as applied to the same problem with $n = 256, K = 64$, **Mosek** runs out of memory.

A CTL — Composite Truncated Level algorithm

A.1 Situation and goal

CTL is a First Order method for solving optimization problems

$$\text{Opt} = \min_{x \in X} \phi(x) := \psi(x) + f(x), \quad (17)$$

where

- $X \subset \mathbf{R}^n$ is nonempty, convex, closed, and bounded
- $\psi : X \rightarrow \mathbf{R}$ and $f : X \rightarrow \mathbf{R}$ are Lipschitz continuous convex functions.

Our assumptions are as follows:

⁵ For description of CTL and related entities, see Section A.

⁶ MATLAB code for this experiment is available at GitHub repository <https://github.com/a11-fr/Algorithms-for-linear-and-polyhedral-estimates/tree/main>.

Assumption 1. X is equipped with a *proximal setup* composed of a norm $\|\cdot\|$ on \mathbf{R}^n and a *distance-generating function* $\omega : X \rightarrow \mathbf{R}$ which is continuously differentiable and strongly convex on X with convexity modulus 1 w.r.t. $\|\cdot\|$:

$$\langle \nabla\omega(x) - \nabla\omega(y), x - y \rangle \geq \|x - y\|^2 \quad \forall x, y \in X.$$

Proximal setup induces *Bregman distance* $V_x(y)$ on X and *Bregman diameter* Ω of X :

$$V_x(y) = \omega(y) - [\omega(x) - \langle \nabla\omega(x), y - x \rangle] \geq \frac{1}{2}\|x - y\|^2, \quad x, y \in X$$

$$\Omega = \left[2 \max_{x, y \in X} V_x(y) \right]^{1/2} \geq \max_{x, y \in X} \|x - y\|.$$

Assumption 2. We have at our disposal *oracle* \mathcal{O} which, given on input a query point $x \in X$, returns a *piece*: a Lipschitz continuous on X convex function

$$\phi_x(\cdot) = \psi(\cdot) + f_x(\cdot) : X \rightarrow \mathbf{R}$$

where $f_x(\cdot)$ belongs to some family \mathcal{F} of “simple” Lipschitz continuous convex functions on X . We suppose that

$$\forall (x, y \in X) : f_x(y) \leq f(y) \quad \& \quad f_x(x) = f(x)$$

(in particular, $\phi(x) = \phi_x(x)$) and that functions $\phi_x(\cdot) : X \rightarrow \mathbf{R}$ are uniformly in $x \in X$ Lipschitz continuous on X :

$$|\phi_x(u) - \phi_x(v)| \leq L_\phi \|u - v\| \quad \forall (u, v \in X) \tag{23}$$

for some $L_\phi < \infty$.

The simplest example of such oracle is that of family \mathcal{F} comprised of affine functions of \mathbf{R}^n , and $f_x(y) = f(x) + \langle f'(x), y - x \rangle$ where $f'(x)$ is a subgradient of f at x (“first order oracle”). In a less trivial example, $f(x)$ is the largest eigenvalue of a symmetric matrix $S(x)$ which affinely depends on x , while $f_x(y) = \max_{i \leq k} e_i^T S(y) e_i$, where $k \leq n$ is fixed and e_1, \dots, e_k are the k leading eigenvectors of $S(x)$.

Assumption 3. We assume that for some positive integer τ we are able to solve efficiently problems of the form

$$\min_y \left\{ \psi(y) + \max_{i \leq \tau} f_{x_i}(y) : y \in X, \alpha(y) \leq 0 \right\}$$

and

$$\min_y \left\{ V_x(y) : y \in X, \psi(y) + \max_{i \leq \tau} f_{x_i}(y) \leq \ell, \alpha(y) \leq 0 \right\}$$

where $\alpha(\cdot)$ is an affine function.

We remark that the algorithm to follow is the composite version of Non-Euclidean Restricted Memory Level method [3] operating with $\psi \equiv 0$ and the family of affine functions in the role of \mathcal{F} ; the Euclidean version of the latter algorithm is the minimization version of the Proximal level bundle method from [10].

A.1.1 The algorithm

The description of CTL is as follows.

0. Control parameters of the algorithms are $\lambda_\ell \in (0, 1), \bar{\theta} \in (0, 1), \underline{\theta} \in (0, 1)$.

1. At the beginning of an iteration of CTL, we have at our disposal

- 1.1.** *upper bound* $\bar{\phi}$ on Opt; the best (the smallest) of the values of ϕ observed at the query points processed so far. These upper bounds do not increase as the iteration count grows. The query point \hat{y} with $\bar{\phi} = \phi(\hat{y})$ is considered as the approximate solution generated so far.
- 1.2.** *lower bound* $\underline{\phi}$ on Opt; these lower bounds do not decrease as the iteration number grows
- 1.3.** *prox-center* $\bar{x} \in X$ and *level* $\ell \in (\underline{\phi}, \bar{\phi})$
- 1.4.** *query point* $x \in X$

- 1.5. *bundle* – a nonempty collection \mathcal{B} of $\tau_{\mathcal{B}} \leq \tau$ pieces $\phi_{\iota}(\cdot) = \psi(\cdot) + f^{\iota}(\cdot)$, $1 \leq \iota \leq \tau_{\mathcal{B}}$, with $f^{\iota} \in \mathcal{F}$; positive integer τ is a parameter of the algorithm.

The very first iteration is preceded by *initialization* where we call the oracle at a (whatever) point $x_{\text{ini}} \in X$ and set

$$\bar{\phi} = \phi(x_{\text{ini}}), \underline{\phi} = \min_{x \in X} [\psi(x) + f_{x_{\text{ini}}}(x)], \mathcal{B} = \{\psi(x) + f_{x_{\text{ini}}}(x)\}$$

Note that since the pieces reported by the oracle underestimate $\phi(\cdot)$, we do ensure $\underline{\phi} \leq \text{Opt}$.

2. Iterations are split into consecutive *phases*, with prox-center and level common to all iterations of a phase.

For a particular phase,

- 2.1. the prox-center \bar{x} is selected at the beginning of the first iteration of the phase and can be a whatever point of X ;
 2.2. the query point of the first iteration of the phase is $x = \bar{x}$,
 2.3. the level ℓ is selected at the beginning of the very first iteration of the phase as

$$\ell = \lambda_{\ell} \underline{\phi} + (1 - \lambda_{\ell}) \bar{\phi}.$$

- At the beginning of the first iteration of a phase, we set

$$\bar{\Delta} = \bar{\phi} - \ell, \underline{\Delta} = \ell - \underline{\phi}, \Delta = \bar{\Delta} + \underline{\Delta} = \bar{\phi} - \underline{\phi}.$$

Note that *the gap* Δ of a phase upper-bounds the inaccuracy in terms of the objective of the approximate solution available at the beginning of the phase.

3. At iterations of a phase, we maintain the relation

$$\phi(y) \geq \ell \text{ for all } y \in X \text{ such that } \langle \nabla \omega(x) - \nabla \omega(\bar{x}), y - x \rangle < 0 \quad (24)$$

where \bar{x}, ℓ are the prox-center and the level of the phase, and x is the query point of the iteration. Note that this relation takes place at the very first iteration of a phase, since for such an iteration $x = \bar{x}$.

4. An iteration of a phase is organized as follows:

- 4.1. We call the oracle at the query point x of the iteration, thus getting the value of the objective $\phi(x)$ and a piece $\phi_x(\cdot)$. After $\phi(x)$ is known, we
- update the upper bound $\bar{\phi}$ and the approximate solution \hat{x} :

$$(\bar{\phi}, \hat{x}) = \begin{cases} (\phi(x), x), & \text{if } \phi(x) < \bar{\phi} \\ (\bar{\phi}, \hat{x}), & \text{otherwise.} \end{cases}$$

- update the bundle \mathcal{B} by adding to it the piece $\phi_x(\cdot)$ and removing, if necessary, one of “old” pieces to keep the number of pieces $\phi_1, \dots, \phi_{\tau_{\mathcal{B}}}$ in the resulting bundle to be at most τ .

- 4.2. If $\phi(x) - \ell \leq \bar{\theta} \bar{\Delta}$ (“essential progress in upper bound on Opt”), we terminate the phase and pass to the next one. Otherwise we solve the auxiliary problem

$$\tilde{\phi} = \min_y \left\{ \hat{\phi}(y) := \max_{1 \leq \iota \leq \tau_{\mathcal{B}}} \phi_{\iota}(y) : y \in X, \langle \nabla \omega(x) - \nabla \omega(\bar{x}), y - x \rangle \geq 0 \right\} \quad (25)$$

(as usual, $\tilde{\phi} = +\infty$ when the right hand side problem is infeasible), and update the lower bound $\underline{\phi}$ on Opt according to

$$\underline{\phi} \mapsto \max \left[\underline{\phi}, \min[\tilde{\phi}, \ell] \right]$$

- **Note.** By (24), $\phi(y) \geq \ell$ at every point $y \in X$ which is not feasible for the optimization problem in (25). Besides this, the pieces $\phi_{\iota}(\cdot)$ in the bundle, and, consequently, the *model* $\hat{\phi}(\cdot)$, underestimate $\phi(\cdot)$ on X . As a result, the quantity $\min[\tilde{\phi}, \ell]$, and consequently the new value of $\underline{\phi}$, indeed is a lower bound on Opt. If updated $\underline{\phi}$ satisfies

$$\ell - \underline{\phi} \leq \underline{\theta} \underline{\Delta}$$

(“essential progress in lower bound on Opt”), we terminate the phase and pass to the next one.

4.3. If the iteration in question does not result in phase change, we solve the auxiliary problem

$$\min_y \left\{ V_{\bar{x}}(y) : y \in X, \widehat{\phi}(y) \leq \ell, \langle \nabla\omega(x) - \nabla\omega(\bar{x}), y - x \rangle \geq 0 \right\} \quad (26)$$

take its optimal solution, x_+ , as the new query point, and pass to the next iteration of the phase.

► **Note.** When we need to solve (26), we have, by construction, $\widetilde{\phi} \leq \ell$, so that the problem in (26) is feasible, a feasible solution being a minimizer in (25). Thus, the new query point x_+ is well defined. Besides this, from the definition of x_+ it follows that

$$\langle \nabla\omega(x_+) - \nabla\omega(\bar{x}), y - x_+ \rangle \geq 0$$

for every feasible solution y to (26). As a result, when $y \in X$ satisfies the relation

$$\langle \nabla\omega(x_+) - \nabla\omega(\bar{x}), y - x_+ \rangle < 0,$$

y is infeasible for (26), meaning that either $\widehat{\phi}(y) > \ell$, and in such case $\phi(y) \geq \widehat{\phi}(y) \geq \ell$, or y satisfies the premise in (24), implying that $\phi(y) \geq \ell$ by (24). We see that (24) holds true when x is replaced with x_+ , that is, (24) is maintained during the iterations.

A.1.2 Convergence analysis

Observe that

(!). *If a phase is finite, then the gap Δ_+ of the subsequent phase does not exceed a fixed fraction $\theta\Delta$ of the gap Δ of the phase in question, where*

$$\theta = \max [1 - \lambda_\ell \bar{\theta}, \underline{\theta} + \lambda_\ell(1 - \underline{\theta})] \in (0, 1);$$

Indeed, (!) is an immediate consequence of the phase termination rules in 4.2 combined with the facts that $\underline{\phi}$ does not decrease, and $\bar{\phi}$ does not increase as the iteration count grows.

The following observation is crucial:

(!!). *The number of iterations at a phase with gap Δ does not exceed*

$$\left\lceil \left(\frac{L_\phi \Omega}{\bar{\theta} \lambda_\ell \Delta} \right)^2 \right\rceil \quad (27)$$

(here $\lceil a \rceil$ stands for the upper integer part of a , the smallest integer greater or equal to a).

Indeed, let ℓ be the level of the phase. Assume that the phase contains more than $T \geq 1$ iterations, so that the upper bound $\bar{\phi}$, the lower bound ϕ on Opt, same as the model $\phi_t(\cdot)$ generated at iteration t of the phase are well defined for $t = 1, \dots, T$, and the query points x_t are well defined for $t = 1, \dots, T + 1$. By construction, for $1 \leq t \leq T$ we have

$$\phi_t(x_t) > \ell + \bar{\theta} \bar{\Delta}, \quad (28a)$$

$$\phi_t(x_{t+1}) \leq \ell, \quad (28b)$$

$$\langle \nabla V_{\bar{x}}(x_t), x_{t+1} - x_t \rangle \geq 0 \quad (28c)$$

where $\bar{x} = x_1$ is the prox-center of the phase. Indeed, when $t \leq T$,

- (28a) holds true since otherwise the phase would be terminated at its t -th iteration due to essential progress in upper bound on Opt, which is not the case when $t \leq T$;
- (28b) and (28c) hold because, by construction of x_{t+1} at a non-concluding iteration t of a phase, x_{t+1} minimizes continuously differentiable on X convex function $V_{\bar{x}}(\cdot)$ over the set

$$X \cap \{y \in X : \phi_t(y) \leq \ell\} \cap \{y \in Y : \langle \nabla V_{\bar{x}}(x_t), y - x_t \rangle \geq 0\}.$$

Now note that by construction of $\phi_t(\cdot)$ and due to Assumption Ass2, ϕ_t is Lipschitz continuous with constant L_ϕ w.r.t. $\|\cdot\|$, which combines with (28a) and (28b) to imply that

$$\|x_t - x_{t+1}\| > L_\phi^{-1} \bar{\theta} \cdot \bar{\Delta} = L_\phi^{-1} \bar{\theta} \lambda_\ell \Delta.$$

The latter relation, in turn, combines with (28c) and with inherited from $\omega(\cdot)$ strong convexity of $V_{\bar{x}}(\cdot)$ on X w.r.t. $\|\cdot\|$ to imply that

$$V_{\bar{x}}(x_{t+1}) > V_{\bar{x}}(x_t) + \frac{[\bar{\theta}\lambda_\ell]^2 \Delta^2}{2L_\phi^2}, 1 \leq t \leq T.$$

Taking into account that $V_{\bar{x}}(x_1) = 0$ and $V_{\bar{x}}(x) \leq \frac{1}{2}\Omega^2$ for all $x \in X$, we arrive at (27).

As an immediate consequence of (!) and (!!), we get the following efficiency estimate:

Proposition 4. *For every $\epsilon \in (0, L_\phi\Omega)$, the overall number of CTL iterations until an ϵ -optimal, as certified by current gap, solution to the minimization problem is built does not exceed*

$$N(\epsilon) = C(L_\phi\Omega/\epsilon)^2$$

with C depending solely on the control parameters λ_ℓ , θ , and $\underline{\theta}$.

B Proofs for Section 2

B.1 Proof of Proposition 1

To save notation, in this section we use shortcut notation \hat{w}_H for the linear estimate $\hat{w}_{\text{lin}}^H(Ax + \xi)$. Note that for all $x \in \mathcal{X}$, the loss $\|\hat{w}_H - Bx\|$ of the estimate \hat{w}_H satisfies

$$\begin{aligned} \|\hat{w}_H - Bx\| &= \|(H^T A - B)x + \sigma H^T \xi\| = \max_{u \in \mathcal{B}_*} \{u^T [(H^T A - B)x + \sigma H^T \xi]\} \\ &= \max_{u \in \mathcal{B}_*} \left\{ [u; x; \sigma \xi]^T \begin{bmatrix} & \frac{1}{2}(B - H^T A) & \frac{1}{2}H^T \\ \frac{1}{2}(B - H^T A)^T & & \\ \frac{1}{2}H & & \end{bmatrix} [u; x; \sigma \xi] \right\} \\ &\leq \max_{u \in \mathcal{B}_*, x \in \mathcal{X}} \left\{ u^T \left[\sum_\ell \lambda_\ell S_\ell \right] u + x^T \left[\sum_k \mu_k T_k \right] x + \sigma^2 \xi^T \Theta \xi \right\} \end{aligned}$$

where $\mu, \lambda \geq 0$ and $\Theta \in \mathbf{S}^m$ are such that

$$\begin{bmatrix} \sum_\ell \mu_\ell S_\ell & \frac{1}{2}(B - H^T A) & \frac{1}{2}H \\ \frac{1}{2}(B - H^T A)^T & \sum_k \lambda_k T_k & \\ \frac{1}{2}H^T & & \Theta \end{bmatrix} \succeq 0.$$

We conclude that for all $x \in \mathcal{X}$,

$$\|\hat{w}_H - Bx\| \leq \max_{s \in \mathcal{S}, t \in \mathcal{T}} \left\{ \sum_\ell \lambda_\ell s_\ell + \sum_k \mu_k t_k + \sigma^2 \xi^T \Theta \xi \right\} = \phi_{\mathcal{S}}(\lambda) + \phi_{\mathcal{T}}(\mu) + \sigma^2 \xi^T \Theta \xi.$$

Due to (3) we have $\mathbf{E}\{\xi\xi^T\} \preceq \sigma^2 I$. We conclude that $\mathbf{E}\{\xi^T \Theta \xi\} \leq \text{Tr}(\Theta)$ what implies the first claim of the proposition. To complete the proof it remains to recall the bound

$$\text{Prob}_\xi \{ \xi^T \Theta \xi \geq \varkappa^2 \text{Tr}(\Theta) \} \leq \epsilon$$

for deviations of the quadratic form of sub-Gaussian random vectors (cf., e.g., [5, 12, 13]). \blacktriangleleft

B.2 Proof of Corollary 2

Indeed, let $\varkappa = 1 + \sqrt{2 \ln[\epsilon^{-1}]}$, and let λ^*, μ^* and Θ^* be components of an optimal solution to (8). Notice that λ^*, μ^* and Θ^* are feasible for (6). Moreover, $\bar{\lambda}, \bar{\mu}$ and $\bar{\Theta}$ where

$$\bar{\lambda} = \varkappa \lambda^*, \quad \bar{\mu} = \varkappa^{-1} \mu^*, \quad \bar{\Theta} = \varkappa^{-1} \Theta^*$$

are also feasible, so, by item ii of Proposition 1, the value

$$\phi_{\mathcal{T}}(\bar{\mu}) + \phi_{\mathcal{S}}(\bar{\lambda}) + \sigma^2 \varkappa^2 \text{Tr}(\bar{\Theta}) \leq \varkappa (\phi_{\mathcal{T}}(\mu^*) + \phi_{\mathcal{S}}(\lambda^*) + \sigma^2 \text{Tr}(\bar{\Theta})) \leq \varkappa r^*$$

upper-bounds the ϵ -risk of \hat{w}_{lin}^* . \blacktriangleleft

B.3 Proof of Lemma 3

Let (λ, μ, Θ) be a feasible solution to (14) such that

$$\Lambda := \sum_{\ell} \lambda_{\ell} S_{\ell} \succ 0, \Xi := \sum_k \mu_k T_k \succ 0. \quad (29)$$

Note that every feasible solution to (14) remains feasible and satisfies (29) after replacing zero entries of λ_{ℓ} and μ_k , if any, with arbitrarily small positive entries.

Now, due to $\begin{bmatrix} \Lambda & \frac{1}{2}B \\ \frac{1}{2}B^T & A^T \Theta A + \Xi \end{bmatrix} \succeq 0$ we have

$$\left[\begin{array}{c|c} I_{\nu} & \frac{1}{2}\Lambda^{-1/2}B\Xi^{-1/2} \\ \hline \frac{1}{2}\Xi^{-1/2}B^T\Lambda^{-1/2} & \Xi^{-1/2}A^T\Theta A\Xi^{-1/2} + I_n \end{array} \right] \succeq 0.$$

When setting $\Theta^{1/2}A\Xi^{-1/2} = US$ with $UU^T = I_n$ and $S \succeq 0$, we get

$$\frac{1}{4}[\Lambda^{-1/2}B\Xi^{-1/2}]^T[\Lambda^{-1/2}B\Xi^{-1/2}] \preceq S^2 + I_n \preceq (S + I_n)^2$$

(recall that $S \succeq 0$). We conclude that there is $Q \in \mathbf{R}^{\nu \times n}$ of spectral norm $\|Q\|_{2,2} \leq 1$ such that

$$\frac{1}{2}[\Lambda^{-1/2}B\Xi^{-1/2}] = Q(S + I_n)$$

and

$$B - 2\Lambda^{1/2}QS\Xi^{1/2} = 2\Lambda^{1/2}Q\Xi^{1/2}.$$

When recalling what S is and setting $H = 2\Theta^{1/2}UQ^T\Lambda^{1/2}$ we have

$$B - H^T A = 2\Lambda^{1/2}Q\Xi^{1/2}.$$

Due to $\|Q\|_{2,2} \leq 1$, by Schur complement lemma, now it follows that

$$\left[\begin{array}{c|c} \Lambda & \frac{1}{2}(B - H^T A) \\ \hline \frac{1}{2}(B - H^T A)^T & \Xi \end{array} \right] \succeq 0. \quad (30)$$

Besides this, by construction,

$$\frac{1}{4}H\Lambda^{-1}H^T = \Theta^{1/2} \underbrace{UQ^TQU^T}_{\preceq I_n} \Theta^{1/2} \preceq \Theta,$$

that is,

$$\left[\begin{array}{c|c} \Lambda & \frac{1}{2}H^T \\ \hline \frac{1}{2}H & \Theta \end{array} \right] \succeq 0. \quad (31)$$

Finally, (30) together with (31) imply that matrix

$$\left[\begin{array}{c|c|c} 2\Lambda & \frac{1}{2}(B - H^T A) & \frac{1}{2}H^T \\ \hline \frac{1}{2}(B - H^T A)^T & \Xi & \\ \hline \frac{1}{2}H & & \Theta \end{array} \right]$$

is positive semidefinite, meaning that $2\lambda, \mu, H$ and Θ form a feasible solution to (8). The corresponding objective value is

$$\phi_S(2\lambda) + \phi_T(\mu) + \sigma^2 \varkappa^2 \text{Tr}(\Theta) = 2\phi_S(\lambda) + \phi_T(\mu) + \sigma^2 \varkappa^2 \text{Tr}(\Theta) \leq \mathfrak{p}_{\varkappa}[\Theta]. \quad \blacktriangleleft$$

C Contrast synthesis for the polyhedral estimate

Projecting, if necessary, the observations onto the image space of A , we reduce the situation to that in which this image space is the entire \mathbf{R}^m ; that is, $m \leq n$ with positive singular values $\sigma_1, \dots, \sigma_m$ of A . Let us assume that $n = m + d$ for some $d \geq 0$, and let

$$A = UDV^T, \quad D = \underbrace{[\text{Diag}(\sigma_1, \dots, \sigma_m), 0_{m \times d}]}_{=: D_m}, \quad U \in \mathbf{R}^{m \times m}, \quad V \in \mathbf{R}^{m \times n},$$

be the (full) singular-value decomposition of A . The starting point of the following computation is the bound (14) for the ϵ -risk of the polyhedral estimate: we have

$$\begin{aligned} \mathbf{p}_\chi^* &= 2 \min_{\lambda, \mu, \Theta} \left\{ \begin{array}{l} \lambda \geq 0, \mu \geq 0, \Theta \succeq 0, \\ \phi_S(\lambda) + \phi_T(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta) : \left[\begin{array}{c|c} \sum_\ell \lambda_\ell S_\ell & \frac{1}{2} B \\ \hline \frac{1}{2} B^T & \sum_k \mu_k T_k + A^T \Theta A \end{array} \right] \succeq 0 \end{array} \right\} \\ &= 2 \min_{\lambda, \mu, \Theta} \left\{ \begin{array}{l} \lambda > 0, \mu \geq 0, \Theta \succeq 0, \\ 2\phi_S(\lambda) + \phi_T(\mu) + \sigma^2 \chi^2 \text{Tr}(\Theta) : A^T \Theta A \succeq \frac{1}{4} B^T \left[\sum_\ell \lambda_\ell S_\ell \right]^{-1} B - \sum_k \mu_k T_k \end{array} \right\} \\ &= 2 \min_{\lambda, \mu, \bar{\Theta}} \left\{ \begin{array}{l} \lambda > 0, \mu \geq 0, \bar{\Theta} = U^T \Theta U \succeq 0, \\ F(\lambda, \mu, \bar{\Theta}) \\ := \phi_S(\lambda) + \phi_T(\mu) + \sigma^2 \chi^2 \text{Tr}(\bar{\Theta}) : \left[\begin{array}{c|c} D_m \bar{\Theta} D_m & \\ \hline & \end{array} \right] = D^T \bar{\Theta} D \\ \succeq V^T \left[\frac{1}{4} B^T \left[\sum_\ell \lambda_\ell S_\ell \right]^{-1} B - \sum_k \mu_k T_k \right] V \end{array} \right\} \\ &= 2 \min_{\lambda, \mu, \bar{\Theta}} \left\{ \begin{array}{l} \lambda > 0, \mu \geq 0, \bar{\Theta} \succeq 0, \\ F(\lambda, \mu, \bar{\Theta}) : \left[\begin{array}{c|c} \bar{\Theta} & \\ \hline & \end{array} \right] \succeq \underbrace{\left[\begin{array}{c|c} D_m & \\ \hline & I \end{array} \right]^{-1} V^T \left[\frac{1}{4} B^T \left[\sum_\ell \lambda_\ell S_\ell \right]^{-1} B - \sum_k \mu_k T_k \right] V \left[\begin{array}{c|c} D_m & \\ \hline & I \end{array} \right]^{-1}}_{=: C(\lambda, \mu)} \end{array} \right\} \quad (32) \end{aligned}$$

Observe that the matrix-valued function $C(\lambda, \mu)$ is \succeq -convex for $\lambda > 0$, and is negative definite for every fixed λ for all μ such that $\min_i \mu_i \geq \underline{\mu}$ large enough. On the other hand, when $Z(\lambda, \mu) \prec 0$ in the representation

$$C(\lambda, \mu) = \left[\begin{array}{c|c} X(\lambda, \mu) & Y(\lambda, \mu) \\ \hline Y^T(\lambda, \mu) & Z(\lambda, \mu) \end{array} \right]$$

the semidefinite constraint of (32) is satisfied if and only if

$$\bar{\Theta} \succeq W(\lambda, \mu) := X(\lambda, \mu) + Y^T(\lambda, \mu) Z(\lambda, \mu)^{-1} Y(\lambda, \mu).$$

As a result, when denoting $[M]_+$ the matrix obtained from a symmetric matrix M by replacing its eigenvalues with their positive parts in the eigenvalue decomposition of M , we conclude that

$$\mathbf{p}_\chi^* = 2 \min_{\lambda, \mu} \left\{ \phi_S(\lambda) + \phi_T(\mu) + \sigma^2 \chi^2 \text{Tr}[W(\lambda, \mu)]_+ : \lambda > 0, \mu \geq 0, Z(\lambda, \mu) \prec 0 \right\}.$$

The bottom line is that in the situation of this section, building Θ (and thus, the near-optimal polyhedral estimate) reduces to solving convex problem of design dimension $L + M$ with (relatively) easy-to-compute objective and constraints.

References

- 1 Yannis Bekri, Anatoli Juditsky, and Arkadi Nemirovski. On robust recovery of signals from indirect observations. <https://arxiv.org/abs/2309.06563>, 2023.
- 2 Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2001.

- 3 Aharon Ben-Tal and Arkadi Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Math. Program.*, 102(3):407–456, 2005.
- 4 Michael Grant and Stephen Boyd. The *cvx* Users' Guide. Release 2.1, 2014. <https://web.cvxr.com/cvx/doc/CVX.pdf>.
- 5 Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17: article no. 52 (6 pages), 2012.
- 6 Anatoli Juditsky and Arkadi Nemirovski. Near-optimality of linear recovery from indirect observations. *Math. Stat. Learn.*, 1(2):171–225, 2018.
- 7 Anatoli Juditsky and Arkadi Nemirovski. Near-optimality of linear recovery in gaussian observation scheme under $\|\cdot\|_2^2$ -loss. *Ann. Stat.*, 46(2):1603–1629, 2018.
- 8 Anatoli Juditsky and Arkadi Nemirovski. On polyhedral estimation of signals via indirect observations. *Electron. J. Stat.*, 14(1):458–502, 2020.
- 9 Anatoli Juditsky and Arkadi Nemirovski. *Statistical Inference via Convex Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2020.
- 10 Krzysztof C. Kiwiel. Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities. *Math. Program.*, 69(1):89–109, 1995.
- 11 Yurii Nesterov and Arkadi Nemirovski. On first-order algorithms for ℓ_1 /nuclear norm minimization. *Acta Numer.*, 22:509–575, 2013.
- 12 Mark Rudelson and Roman Vershynin. Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18: article no. 82 (9 pages), 2013.
- 13 Vladimir Spokoiny and Maya Zhilova. Sharp deviation bounds for quadratic forms. *Math. Methods Stat.*, 22(2):100–113, 2013.