NEAR-OPTIMALITY OF LINEAR RECOVERY IN GAUSSIAN OBSERVATION SCHEME UNDER $\|\cdot\|_2^2$-LOSS

BY ANATOLI JUDITSKY[1] AND ARKADI NEMIROVSKI[2]

Université Grenoble-Alpes and Georgia Institute of Technology

We consider the problem of recovering linear image Bx of a signal x known to belong to a given convex compact set $\mathcal{X}$ from indirect observation $\omega = Ax + \sigma\xi$ of x corrupted by Gaussian noise ξ. It is shown that under some assumptions on $\mathcal{X}$ (satisfied, e.g., when $\mathcal{X}$ is the intersection of K concentric ellipsoids/elliptic cylinders), an easy-to-compute linear estimate is near-optimal in terms of its worst case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$-loss. The main novelty here is that the result imposes no restrictions on A and B. To the best of our knowledge, preceding results on optimality of linear estimates dealt either with one-dimensional Bx (estimation of linear forms) or with the "diagonal case" where A, B are diagonal and $\mathcal{X}$ is given by a "separable" constraint like $\mathcal{X} = \{x : \sum_i a_i^2 x_i^2 \le 1\}$ or $\mathcal{X} = \{x : \max_i |a_i x_i| \le 1\}$.

1. Introduction. In this paper, we address one of the most basic problems of high-dimensional statistics, specifically, as follows: given indirect noisy observation

$$\omega = Ax + \sigma\xi \qquad \left[A : m \times n, \xi \sim \mathcal{N}(0, I_m)\right]$$

of unknown "signal" x known to belong to a given convex compact subset $\mathcal{X}$ of $\mathbf{R}^n$, we want to recover the image $w = Bx \in \mathbf{R}^\nu$ of x under a given linear mapping. We focus on the case where the quality of a candidate recovery $\omega \mapsto \widehat{w}(\omega)$ is quantified by its worst case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$-error, that is, by the risk

$$\text{Risk}[\widehat{w}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \left[\mathbf{E}_\xi\left\{\|\widehat{w}(Ax + \sigma\xi) - Bx\|_2^2\right\}\right]^{1/2}.$$

The simplest and the most studied type of recovery is affine one: $\widehat{w}(\omega) = H^T\omega + h$; assuming $\mathcal{X}$ symmetric w.r.t. the origin, we lose nothing when passing from affine estimates to linear ones—those of the form $\widehat{w}_H(\omega) = H^T\omega$. Starting from the pioneering works of Kuks and Olman (1971, 1972), linear estimates received much attention in the statistical literature [see, e.g., Arnold and Stahlecker (2000), Christopeit and Helmes (1996), Drygas (1996), Pilz (1986), Rao (1973, 1976) and

references therein]. An advantage of linear estimates, from the computational point of view, is that under favorable circumstances (e.g., when $\mathcal{X}$ is an ellipsoid), minimizing risk over linear estimates is an efficiently solvable problem. On the other hand, linear estimates are also of major importance to statistical theory. For instance, a huge body of results on *rate-optimality* of linear estimates on various signal classes (which arise from some classes of regular functions) form the backbone of classical nonparametric statistics [see, e.g., Ibragimov and Has'minskiĭ (1981), Tsybakov (2009), Wasserman (2006)]. Furthermore, for several important signal classes, linear estimates occur to be near-optimal on the class of *all* possible estimates. This is, for instance, the case for signal recovery from direct observations (the case of $B = A = I$) in the situation where the set $\mathcal{X}$ of signals is an ellipsoid or a box. The case of ellipsoidal $\mathcal{X}$ was studied first by M. S. Pinsker [see Pinsker (1980)], who showed that in the problem of recovery of the signal $x \in \mathcal{X}$ from direct observation $\omega = x + \sigma\xi$, $\mathcal{X}$ being a "Sobolev ellipsoid" of the form $\{x \in \mathbf{R}^n : \sum_j j^{2\alpha} x_j^2 \leq L^2\}$, the ratio of the risk of a properly selected linear estimate to the minimax risk $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] := \inf_{\widehat{w}(\cdot)} \mathrm{Risk}[\widehat{w}|\mathcal{X}]$ (the infimum is taken over all estimates, not necessarily linear) tends to 1, as $\sigma \to +0$, and this happens *uniformly in n*, α and L being fixed. Similar "asymptotic optimality" results are also known for ellipsoids related to classes of analytic functions [Golubev, Levit and Tsybakov (1996)] and for "diagonal" case, where $\mathcal{X}$ is the above ellipsoid/box and A, B are diagonal matrices [Efromovich and Pinsker (1996)] [see also Efromovich (1999) for a modern presentation of that approach]. The results on nonasymptotic near-optimality of linear estimates (up to a factor $1.11\ldots$) are also available for the case where $A = B = I$ and $\mathcal{X}$ is an ellipsoid ($\mathcal{X} = \{x \in \mathbf{R}^n : \sum_j a_j^2 x_j^2 \leq 1\}$ for given a_j) or a box ($\mathcal{X} = \{x \in \mathbf{R}^n : \max_j |a_j x_j| \leq 1\}$) [see, e.g., Donoho, Liu and MacGibbon (1990)] (the corresponding argument can be easily extended to the case of diagonal A and B). Note that the situation is quite different for the problem of estimation of a *linear form* $w = b^T x$ (i.e., the case of one-dimensional Bx). An exceptional from several points of view "general" (i.e., not imposing severe restrictions on how the geometries of $\mathcal{X}$, A and B are linked to each other) result on optimality of linear estimates in this case is due to D. Donoho who proved [Donoho (1994)] that when recovering a linear form, the best, over all linear estimates, risk is within the factor $1.11\ldots$ of the minimax risk.

The goal of this paper is to establish a rather general result on near-optimality of properly built linear estimates as compared to all possible estimates. Note that a result of this type is bounded to impose some restrictions on $\mathcal{X}$, since there are cases (e.g., the one of a high-dimensional $\|\cdot\|_1$-ball $\mathcal{X}$) where linear estimates are *by far* suboptimal. Our restrictions on the family of sets (we call them *ellitopes*) $\mathcal{X}$ reduce to the existence of a special type representation of $\mathcal{X}$ and are satisfied, for example, when $\mathcal{X}$ is the intersection of $K < \infty$ ellipsoids/elliptic cylinders:

$$(1) \qquad \mathcal{X} = \left\{x \in \mathbf{R}^n : x^T S_k x \leq 1, 1 \leq k \leq K\right\} \qquad \left[S_k \succeq 0, \sum_k S_k \succ 0\right].$$

In particular, $\mathcal{X}$ can be a symmetric w.r.t. the origin compact polytope given by $2K$ linear inequalities $-1 \leq s_k^T x \leq 1$, $1 \leq k \leq K$. Another instructive example is a set of the form $\mathcal{X} = \{x : \|Sx\|_p \leq L\}$, where $p \geq 2$ and S is a matrix with trivial kernel. It should be stressed that while imposing some restrictions on $\mathcal{X}$, *we require nothing from A and B*. Our main result (Theorem 2.1) states, in particular, that in the case of $\mathcal{X}$ given by (1) and arbitrary A, B, the risk of properly selected linear estimate $\widehat{w}_{H_*}$, with both H_* and the risk being efficiently computable, satisfies the bound

$$\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln\left(\frac{O(1)\|B\|^2 K^2 \kappa^{-1}}{\text{Risk}_{\text{opt}}^2[\mathcal{X}]}\right)} \, \text{Risk}_{\text{opt}}[\mathcal{X}], \qquad (*)$$

where $\|B\|$ is the spectral norm of B, κ is the minimal eigenvalue of $\sum_k S_k$, $\text{Risk}_{\text{opt}}[\mathcal{X}]$ is the minimax risk and $O(1)$ stands for an absolute constant. It should be mentioned that technique used to construct lower bound for optimal risks leads to more precise oracle inequalities when imposing constraints on the structure of the signal class $\mathcal{X}$ and matrices A, B; in particular, it allows to reproduce classical "asymptotic" optimality results, for example, in the situation considered in Efromovich and Pinsker (1996), Pinsker (1980). On the other hand, we do not know if the bound $(*)$ can be significantly improved in some important "simple cases," for instance, in the case where $B = I$ and $\mathcal{X}$ is an ellipsoid, without imposing any restrictions on A. In this work, however, we prefer to see our approach as "operational"—the *provably nearly optimal* estimate itself, its risk and even the lower risk bound involved are all given by an efficient computation, which is supposed to provide precise near-optimality guaranties for each set of the problem data. From this point of view, the oracle inequality $(*)$ can be viewed as a general indication of a "goodness" of linear estimates in a certain context, namely, where the signal set is an intersection of "not too large" number of ellipsoids/elliptic cylinders. This is in sharp contrast with traditional results of nonparametric statistics, where near-optimal estimates and their risks are given in a "closed analytical form," at the price of severe restrictions on the structure of the "data" $\mathcal{X}$, A and B.[3] This being said, it should be stressed that one of the crucial components of our construction is completely classical—this is the idea, going back to Pinsker (1980) to bound from below the minimax risk via Bayesian risk associated with properly selected Gaussian prior.[4]

The main body of the paper is organized as follows. Section 2 contains problem formulation (Section 2.1), construction of the linear estimate we deal with (Sec-

---

[3]Since this paper has been submitted, the proposed approach has been further developed in Juditsky and Nemirovski (2016). For instance, it is shown that similar near-optimality guaranties for linear estimators can be obtained for more general risks (e.g., ℓ_p-loss with $1 \leq p \leq 2$) and slightly more general sets $\mathcal{X}$, which are solution sets of systems of quadratic *matrix* inequalities, for deterministic bounded noises, etc.

[4]Pinsker (1980) addresses the problem of $\|\cdot\|_2$-recovery of a signal x from direct observations ($A = B = I$) in the case where $\mathcal{X}$ is a high-dimensional ellipsoid with "regularly decreasing half-

tion 2.2) and the central result on near-optimality of this estimate (Section 2.3). Section 3 contains some extensions. Specifically, we present a version of our main result for the case when the usual worst case expected $\| \cdot \|_2^2$-risk is replaced with properly defined *relative* risk (Section 3.1) and provide a robust, w.r.t. uncertainty in A, B, version of the estimate (Section 3.2). In Section 3.3, we show that the key argument underlying the proof of our main result can be used beyond the scope of statistics, specifically, when quantifying the approximation ratio of the semidefinite relaxation bound on the maximum of a quadratic form over an ellitope. Proofs are relegated to the Supplementary Material [Juditsky and Nemirovski (2018)].

2. Situation and main result.

2.1. *Situation and goal.*   Given $v \times n$ matrix B, consider the problem of estimating linear image Bx of unknown signal x known to belong to a given set $\mathcal{X} \subset \mathbf{R}^n$ via noisy observation

$$(2) \qquad\qquad\qquad\qquad \omega = Ax + \sigma\xi,$$

where an $m \times n$ matrix A and $\sigma > 0$ are given, and $\xi \sim \mathcal{N}(0, I_m)$ is the standard Gaussian observation noise. From now on, we assume that $\mathcal{X} \subset \mathbf{R}^n$ is a set given by

$$(3) \qquad \mathcal{X} = \left\{ x \in \mathbf{R}^n : \exists\bigl(y \in \mathbf{R}^{\bar{n}}, t \in \mathcal{T}\bigr) : x = Py, y^T S_k y \le t_k, 1 \le k \le K \right\},$$

where:

- P is an $n \times \bar{n}$ matrix,
- $S_k \succeq 0$ are $\bar{n} \times \bar{n}$ matrices with $\sum_k S_k \succ 0$,
- $\mathcal{T}$ is a nonempty computationally tractable[5] convex compact subset of $\mathbf{R}_+^K$ intersecting the interior of $\mathbf{R}_+^K$ and such that $\mathcal{T}$ is monotone, meaning that the relations $0 \le \tau \le t$ and $t \in \mathcal{T}$ imply that $\tau \in \mathcal{T}$.[6] Note that under our assumptions $\operatorname{int} \mathcal{T} \ne \varnothing$.

We assume that $BP \ne 0$, since otherwise one has $Bx = 0$ for all $x \in \mathcal{X}$ and the estimation problem is trivial. In the sequel, we refer to a set of the form (3) with data

---

axes," like $\mathcal{X} = \{ x \in \mathbf{R}^n : \sum_j j^{2\alpha} x_j^2 \le L^2 \}$ with $\alpha > 0$. In this case, Pinsker's construction shows that as $\sigma \to +0$, the risk of properly built linear estimate is uniformly in n, $(1 + o(1))$ times the minimax risk. This is much stronger than (∗), and it seems quite unlikely that a similarly strong result may hold true in the general case underlying (∗).

[5]For all practical purposes, it suffices to assume that $\mathcal{T}$ is given by an explicit *semidefinite representation*

$$\mathcal{T} = \{ t : \exists w : A(t, w) \succeq 0 \},$$

where $A(t, w)$ is a symmetric and affine in t, w matrix.

[6]The latter relation is "for free"—given a nonempty convex compact set $\mathcal{T} \subset \mathbf{R}_+^K$, the right-hand side of (3) remains intact when passing from $\mathcal{T}$ to its "monotone hull" $\{ \tau \in \mathbf{R}_+^K : \exists t \in \mathcal{T} : \tau \le t \}$, which already is a monotone convex compact set.

$[P, \{S_k, 1 \le k \le K\}, \mathcal{T}]$ satisfying just formulated assumptions as to an *ellitope*, and to (3)—as to *ellitopic representation* of $\mathcal{X}$. Here are instructive examples of ellitopes (in all these examples, P is the identity mapping):

- when $K = 1$, $\mathcal{T} = [0, 1]$ and $S_1 \succ 0$, $\mathcal{X}$ is the ellipsoid $\{x : x^T S_1 x \le 1\}$;
- when $K \ge 1$, $\mathcal{T} = \{t \in \mathbf{R}^K : 0 \le t_k \le 1, k \le K\}$, and $\mathcal{X}$ is the intersection

$$\bigcap_{1 \le k \le K} \{x : x^T S_k x \le 1\}$$

  of centered at the origin ellipsoids/elliptic cylinders. In particular, when U is a $K \times n$ matrix of rank n with rows u_k^T, $1 \le k \le K$, and $S_k = u_k u_k^T$, $\mathcal{X}$ is symmetric w.r.t. the origin polytope $\{x : \|Ux\|_\infty \le 1\}$;
- when U, u_k and S_k are as in the latter example and $\mathcal{T} = \{t \in \mathbf{R}_+^K : \sum_k t_k^{p/2} \le 1\}$ for some $p \ge 2$, we get $\mathcal{X} = \{x : \|Ux\|_p \le 1\}$.

It should be added that the family of ellitope-representable sets is quite rich: this family admits a "calculus," so that more ellitopes can be constructed by taking intersections, direct products, linear images (direct and inverse) or arithmetic sums of "basic ellitopes" given by the above examples. In fact, the property to be an ellitope is preserved by all basic operations with sets preserving convexity and symmetry w.r.t. the origin; see Section B of the Supplementary Material [Juditsky and Nemirovski (2018)].

As another instructive, in the context of nonparametric statistics, example of an ellitope, consider the situation where our signals x are discretizations of functions of continuous argument running through a compact d-dimensional domain D, and the functions f we are interested in are those satisfying a Sobolev-type smoothness constraint—an upper bound on the $L_p(D)$-norm of $\mathcal{L}f$, where $\mathcal{L}$ is a linear differential operator with constant coefficients. After discretization, this restriction can be modeled as $\|Lx\|_p \le 1$, with properly selected matrix L. As we already know from the above example, when $p \ge 2$, the set $\mathcal{X} = \{x : \|Lx\|_p \le 1\}$ is an ellitope, and as such is captured by our machinery. Note also that by the outlined calculus, imposing on the functions f in question *several Sobolev-type smoothness constraints* with parameters $p \ge 2$, still results in a set of signals which is an ellitope.

Estimates and their risks.   In the outlined situation, a candidate estimate is a Borel function $\widehat{w}(\cdot) : \mathbf{R}^m \to \mathbf{R}^\nu$; given observation (2) we recover $w = Bx$ as $\widehat{w}(\omega)$. In the sequel, we quantify the quality of an estimate by its worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$ recovery error:

$$\text{Risk}[\widehat{w}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \left[\mathbf{E}_{\xi \sim \mathcal{N}(0, I_m)} \{\|\widehat{w}(Ax + \sigma\xi) - Bx\|_2^2\}\right]^{1/2}$$

and define the optimal, or the *minimax*, risk as

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

where inf is taken over all Borel candidate estimates.

Our initial observation is that when replacing matrices A and B with AP and BP, respectively, we pass from the initial estimation problem of interest—one where the signal set $\mathcal{X}$ is given by (3), and we want to recover Bx, $x \in \mathcal{X}$, via observation (2), to the *transformed problem*, where the signal set is

$$\mathcal{Y} = \{y \in \mathbf{R}^{\bar{n}} : \exists t \in \mathcal{T} : y^T S_k y \leq t_k, 1 \leq k \leq K\},$$

and we want to recover $[BP]y$, $y \in \mathcal{Y}$, via observation

$$\omega = [AP]y + \sigma \xi.$$

It is obvious that the considered families of estimates (the family of all linear and the family of all estimates), same as the risks of the estimates, remain intact under this transformation; in particular,

$$\text{Risk}[\widehat{w}|\mathcal{X}] = \sup_{y \in \mathcal{Y}} \left[\mathbf{E}_\xi \{ \|\widehat{w}([AP]y + \sigma\xi) - [BP]y\|_2^2 \} \right]^{1/2}.$$

Therefore, to save notation, from now on and unless mentioned otherwise, we assume that matrix P is identity, so that $\mathcal{X}$ is the ellitope

$$(4) \qquad \mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T}, x^T S_k x \leq t_k, 1 \leq k \leq K\}.$$

The *Main goal* of what follows is to demonstrate that a *linear in ω* estimate

$$(5) \qquad \widehat{w}_H(\omega) = H^T \omega$$

with properly selected efficiently computable matrix H is near-optimal in terms of its risk. We start with building this estimate.

2.2. *Building linear estimate.* Restricting ourselves to linear estimates (5), we may be interested in the estimate with the smallest risk, that is, associated with a $\nu \times m$ matrix H which is an optimal solution to the optimization problem:

$$\min_H \{R(H) := \text{Risk}^2[\widehat{w}_H|\mathcal{X}]\}.$$

We have

$$R(H) = \max_{x \in \mathcal{X}} \mathbf{E}_\xi \|H^T \omega - Bx\|_2^2$$

$$= \mathbf{E}_\xi \|H^T \xi\|_2^2 + \max_{x \in \mathcal{X}} \|H^T Ax - Bx\|_2^2$$

$$= \sigma^2 \text{Tr}(H^T H) + \max_{x \in \mathcal{X}} x^T (H^T A - B)^T (H^T A - B)x.$$

As the maximum over x of convex quadratic functions of H, $R(H)$ is itself convex. However, as the maximum over $\mathcal{X}$ of a quadratic in x function, $R(H)$ is typically hard to compute.[7] For this reason, we use a linear estimate yielded by minimizing

---

[7]For instance, when $\mathcal{X}$ is a unit cube $\{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$, computing $R(0)$ in the case of general-type B is equivalent to maximizing over X a general-type convex quadratic form; it is known that solving the latter problem already within 4% accuracy is NP-hard.

an *efficiently computable convex upper bound* on $R(H)$ which is built as follows. Let $\phi_{\mathcal{T}}$ be the support function of $\mathcal{T}$:

$$\phi_{\mathcal{T}}(\lambda) = \max_{t \in \mathcal{T}} \lambda^T t : \mathbf{R}^K \to \mathbf{R}.$$

Observe that whenever $\lambda \in \mathbf{R}_+^K$ and H are such that

$$(6) \qquad (B - H^T A)^T (B - H^T A) \preceq \sum_k \lambda_k S_k,$$

for $x \in \mathcal{X}$ it holds

$$(7) \qquad \|Bx - H^T A x\|_2^2 \le \phi_{\mathcal{T}}(\lambda).$$

Indeed, in the case of (6) and with $x \in \mathcal{X}$, there exists $t \in \mathcal{T}$ such that $x^T S_k x \le t_k$ for all t, and consequently the vector $\bar{t}$ with the entries $\bar{t}_k = x^T S_k x$ also belongs to $\mathcal{T}$, whence

$$\|Bx - H^T A x\|_2^2 = \|Bx - H^T A x\|_2^2 \le \sum_k \lambda_k x^T S_k x = \lambda^T \bar{t} \le \phi_{\mathcal{T}}(\lambda),$$

which combines with (4) to imply (7).

From (7), it follows that if H and $\lambda \ge 0$ are linked by (6), then

$$\text{Risk}^2[\widehat{x}_H | \mathcal{X}] = \max_{x \in \mathcal{X}} \mathbf{E}\{\|Bx - H^T(Ax + \sigma\xi)\|_2^2\}$$
$$= \sigma^2 \text{Tr}(H^T H) + \max_{x \in \mathcal{X}} \|[B - H^T A]x\|_2^2$$
$$\le \sigma^2 \text{Tr}(H^T H) + \phi_{\mathcal{T}}(\lambda).$$

We see that the efficiently computable convex function

$$\widehat{R}(H) = \inf_{\lambda} \left\{ \sigma^2 \text{Tr}(H^T H) + \phi_{\mathcal{T}}(\lambda) : (B - H^T A)^T (B - H^T A) \preceq \sum_k \lambda_k S_k, \lambda \ge 0 \right\}$$

(which clearly is well defined due to compactness of $\mathcal{T}$ combined with $\sum_k S_k \succ 0$) is an upper bound on $R(H)$.[8] Therefore, the efficiently computable optimal

---

[8]It is well known that when $K = 1$ (i.e., $\mathcal{X}$ is an ellipsoid) the above bounding scheme is exact: $R(\cdot) \equiv \widehat{R}(\cdot)$. For more complicated $\mathcal{X}$'s, $\widehat{R}(\cdot)$ could be larger than $R(\cdot)$, although the ratio $\widehat{R}(\cdot)/R(\cdot)$ is bounded by $O(\log(K))$; see Section 3.3.

solution (H_*, λ_*) to the (clearly solvable) convex optimization problem

$$\text{Opt} = \min_{H,\lambda} \Big\{ \sigma^2 \text{Tr}(H^T H) + \phi_{\mathcal{T}}(\lambda) :$$

$$(B - H^T A)^T (B - H^T A) \preceq \sum_k \lambda_k S_k, \lambda \geq 0 \Big\}$$

(8)

$$= \min_{H,\lambda} \Big\{ \sigma^2 \text{Tr}(H^T H) + \phi_{\mathcal{T}}(\lambda) :$$

$$\begin{bmatrix} \sum_k \lambda_k S_k & B^T - A^T H \\ B - H^T A & I_\nu \end{bmatrix} \succeq 0, \lambda \geq 0 \Big\}$$

yields a linear estimate $\widehat{w}_{H_*}$ with the risk upper-bounded by $\sqrt{\text{Opt}}$.

2.3. *Lower-bounding optimal risk and near-optimality of* $\widehat{w}_{H_*}$. Let us consider the convex optimization problem

$$\text{Opt}_* = \max_{Q,t} \big\{ \varphi(Q) := \text{Tr}\big(B[Q - QA^T(\sigma^2 I_m + AQA^T)^{-1}AQ]B^T\big),$$

(9)

$$Q \succeq 0, t \in \mathcal{T}, \text{Tr}(QS_k) \leq t_k, 1 \leq k \leq K \big\}$$

$$= \max_{Q,t} \Big\{ \text{Tr}(BQB^T) - \text{Tr}(G) : \begin{bmatrix} G & BQA^T \\ AQB^T & \sigma^2 I_m + AQA^T \end{bmatrix} \succeq 0,$$

(10)

$$Q \succeq 0, t \in \mathcal{T}, \text{Tr}(QS_k) \leq t_k, 1 \leq k \leq K \Big\}.$$

Note that the function $\varphi(Q)$ has a transparent statistical interpretation. Specifically, given an $n \times n$ matrix $Q \succeq 0$, consider two independent Gaussian random vectors, $\xi \sim \mathcal{N}(0, I_m)$ and $\eta \sim \mathcal{N}(0, Q)$. We claim that

$$(11) \qquad \varphi(Q) = \inf_{G(\cdot):\mathbf{R}^m \to \mathbf{R}^\nu} \mathbf{E}_{[\xi,\eta]}\big\{ \|G(\sigma\xi + A\eta) - B\eta\|_2^2 \big\}.$$

Indeed, by the normal correlation theorem [see, e.g., Liptser and Shiryayev (1977), Theorem 13.1], the optimal, in terms of expected $\|\cdot\|_2^2$-error, recovery $G_*(\cdot)$ of $B\eta$ via observation $\sigma\xi + A\eta$—the conditional, given $\sigma\xi + A\eta$, expectation of $B\eta$—is linear, and the corresponding expected $\|\cdot\|_2^2$-error is exactly $\varphi(Q)$.

In the sequel, we set

$$\mathcal{Q} = \big\{ Q \in \mathbf{S}^n : Q \succeq 0, \exists t \in \mathcal{T} : \text{Tr}(QS_k) \leq t_k, 1 \leq k \leq K \big\}.$$

Note that $\mathcal{Q}$ is a convex compact set due to $\sum_k S_k \succ 0$ combined with compactness of $\mathcal{T}$.

Observe that if (Q, t) is feasible for (9), then the Gaussian random vector $\eta \sim \mathcal{N}(0, Q)$ belongs to $\mathcal{X}$ "on average"—it satisfies the constraints $\mathbf{E}\{\eta^T S_k \eta\} = \text{Tr}(QS_k) \leq t_k, k = 1, \ldots, K$, and $t \in \mathcal{T}$. The lower bounding scheme we intend

to implement goes back to Pinsker (1980) and heavily relies upon this fact—it bounds from below the minimax, over $x \in \mathcal{X}$, risk of estimating Bx by comparing this risk to the risk of optimal recovery of $B\eta$ in the Gaussian problem, where $\eta \in \mathcal{X}$ with "high probability," as is the case when $Q \in \rho\mathcal{Q}$ with appropriate $\rho < 1$. Specifically, we have the following result.

LEMMA 2.1. *Given a positive semidefinite $n \times n$ matrix Q and $\delta \in (0, 1/5]$, let $\eta \sim \mathcal{N}(0, Q)$ and $\xi \sim \mathcal{N}(0, I_m)$ be independent from each other Gaussian vectors. Assume that*

$$\operatorname{Prob}\{\eta \notin \mathcal{X}\} \le \delta \le 1/5.$$

Then

$$(12) \qquad \varphi(Q) \le \operatorname{Risk}^2_{\mathrm{opt}}[\mathcal{X}] + \left[M_* + \sqrt{2}q_{1-\delta/2}\sqrt{\operatorname{Tr}(BQB^T)}\right]^2 \delta,$$

where q_α is the α-quantile of the standard normal distribution:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q_\alpha} e^{-s^2/2} \, ds = \alpha,$$

and

$$(13) \qquad M_* = \sqrt{\max_{Q \in \mathcal{Q}} \operatorname{Tr}(BQB^T)}.$$

Further, if $Q \in \rho\mathcal{Q}$ for some $\rho \in (0, 1]$, then

$$\varphi(Q) \le \operatorname{Risk}^2_{\mathrm{opt}}[\mathcal{X}] + [1 + \sqrt{2\rho}q_{1-\delta/2}]^2 M_*^2 \delta.$$

For proof, see Section A.1.1 of the Supplementary Material [Juditsky and Nemirovski (2018)].

The second principal component of the construction of the lower bound for $\operatorname{Risk}_{\mathrm{opt}}$ is provided by the following statement.

LEMMA 2.2. *In the premise of this section, (10) is a conic problem which is strictly feasible and solvable, with the conic dual problem equivalent to (8). As a consequence, one has*

$$(14) \qquad \operatorname{Opt}_* = \operatorname{Opt}.$$

Let now (Q, t) be an optimal solution to (9), and let for $0 < \rho \le 1$, $Q_\rho = \rho Q$. Note that $\varphi(Q_\rho) \ge \rho\varphi(Q) = \rho\operatorname{Opt}$ (recall that φ is concave[9] with $\varphi(0_{m \times m}) = 0$),

---

[9]Concavity of φ can be verified directly; a transparent alternative verification is to notice that (11) implies that

$$\varphi(Q) = \min_H \mathbf{E}_{[\xi,\zeta] \sim \mathcal{N}(0, I_{m+n})}\{([B - H^T A]Q^{1/2}\zeta + H^T \sigma\xi)^T([B - H^T A]Q^{1/2}\zeta + H^T \sigma\xi)\}$$

$$= \min_H [\operatorname{Tr}([B - H^T A]^T Q[B - H^T A]) + \sigma^2 \operatorname{Tr}(H^T H)]$$

is a minimum of affine functions of Q and as such is concave.

and

$$\text{Tr}(BQ_\rho B^T) = \rho\,\text{Tr}(BQB^T) \le \rho M_*^2.$$

In view of Lemma 2.1 as applied with Q_ρ in the role of Q, whenever $\rho \in (0, 1]$ and there exists $\delta_\rho \le 1/5$ such that $\text{Prob}_{\eta\sim\mathcal{N}(0,Q_\rho)}\{\eta \notin \mathcal{X}\} \le \delta_\rho$, we have

$$(15) \qquad \rho\,\text{Opt} \le \varphi(Q_\rho) \le \text{Risk}_{\text{opt}}^2[\mathcal{X}] + [1 + \sqrt{2\rho}q_{1-\delta_\rho/2}]^2 M_*^2 \delta_\rho.$$

To proceed, we need an upper bound δ_ρ on the probability $\text{Prob}_{\eta\sim\mathcal{N}(0,Q_\rho)}\{\eta \notin \mathcal{X}\}$. It is given by the following simple result.

LEMMA 2.3. *Let S and Q be positive semidefinite $n \times n$ matrices with $\rho := \text{Tr}(SQ) \le 1$, and let $\eta \sim \mathcal{N}(0, Q)$. Then*

$$(16) \qquad \begin{aligned} \text{Prob}\{\eta^T S\eta > 1\} &\le \inf_{0\le\gamma<\min_i(2s_i)^{-1}}\left\{\exp\left(-\frac{1}{2}\sum_{i=1}^n \ln(1-2\gamma s_i) - \gamma\right)\right\} \\ &\le e^{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}}, \end{aligned}$$

where s_i are the eigenvalues of $Q^{1/2}SQ^{1/2}$.

Now we are done. Indeed, note that the matrix Q_ρ satisfies $\text{Tr}(S_k Q_\rho) \le \rho t_k$ for some $t \in \mathcal{T}$; applying Lemma 2.3 and taking into account (4), we conclude that

$$\text{Prob}_{\eta\sim\mathcal{N}(0,Q_\rho)}\{\eta \notin \mathcal{X}\} \le \sum_{k=1}^K \text{Prob}\{\eta^T S_k\eta > t_k\} \le K\exp\left\{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}\right\},$$

so we can set

$$(17) \qquad \delta_\rho := \min\left[K\exp\left\{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}\right\}, 1\right].$$

It is straightforward to verify that with the just-defined δ_ρ, for $0 < \rho < 1$ it holds

$$(18) \qquad [1 + \sqrt{2\rho}q_{1-\delta_\rho/2}]^2\delta_\rho \le 8K\exp\{-(3\rho)^{-1}\}.$$

Assuming that $\delta_\rho \le 1/5$, the latter bound combines with (15) to yield

$$(19) \qquad \rho\,\text{Opt} \le \text{Risk}_{\text{opt}}^2[\mathcal{X}] + 8KM_*^2\exp\{-(3\rho)^{-1}\}.$$

Let us choose

$$\bar\rho^{-1} = 3\ln\left(\frac{8KM_*^2}{\text{Risk}_{\text{opt}}^2[\mathcal{X}]}\right)$$

so that

$$8KM_*^2\exp\{-(3\bar\rho)^{-1}\} \le \text{Risk}_{\text{opt}}^2[\mathcal{X}].$$

Observe that by evident reasons $M_*^2 \geq \text{Risk}_{\text{opt}}^2[\mathcal{X}]$, whence $\bar{\rho}^{-1} \geq 3\ln(8K)$, which in view of (17) implies that $\delta_{\bar{\rho}} \leq 1/5$, so that (19) is applicable to $\rho = \bar{\rho}$, thus implying that

$$\text{Opt} \leq \frac{2}{\bar{\rho}} \text{Risk}_{\text{opt}}^2[\mathcal{X}] = 6\ln\left(\frac{8KM_*^2}{\text{Risk}_{\text{opt}}^2[\mathcal{X}]}\right)\text{Risk}_{\text{opt}}^2[\mathcal{X}].$$

Recalling that $\sqrt{\text{Opt}}$ upper-bounds $\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}]$, we have arrived at our main result.

THEOREM 2.1. *The efficiently computable linear estimate* $\widehat{w}_{H_*}(\omega) = H_*^T \omega$ *yielded by an optimal solution to the optimization problem* (8) *is nearly optimal in terms of its risk*:

$$(20) \qquad \text{Risk}[\widehat{w}_{H_*}|\mathcal{X}] \leq \sqrt{\text{Opt}} \leq \sqrt{6\ln\left(\frac{8M_*^2K}{\text{Risk}_{\text{opt}}^2[\mathcal{X}]}\right)}\text{Risk}_{\text{opt}}[\mathcal{X}]$$

with M_* *given by* (13).[10]

2.4. *Discussion.* The result of Theorem 2.1 merits few comments.

1. *Simplifying expression for nonoptimality factor.* Relation (20) states that when $\mathcal{X}$ is an ellitope (3), the risk $\sqrt{\text{Opt}}$ of the efficiently computable linear estimate yielded by (8) is just by a logarithmic in $\frac{M_*^2K}{\text{Risk}_{\text{opt}}^2[\mathcal{X}]}$ factor worse than the optimal risk $\text{Risk}_{\text{opt}}[\mathcal{X}]$. A minor shortcoming of (20) is that the "nonoptimality factor" is expressed in terms of unknown to us optimal risk. This can be easily cured. For example, setting

$$\bar{\rho}^{-1} = 6\ln\left(\frac{17KM_*^2}{\text{Opt}}\right),$$

it is immediately seen that

$$\frac{\bar{\rho}}{2}\text{Opt} \geq 6KM_*^2\exp\{-(3\bar{\rho})^{-1}\},$$

and $\delta_{\bar{\rho}}$ as given by (17) with $\rho = \bar{\rho}$ is $\leq 1/5$, implying by (19) that $\frac{1}{2}\bar{\rho}\,\text{Opt} \leq \text{Risk}_{\text{opt}}^2[\mathcal{X}]$, whence

$$(21) \qquad \text{Risk}_{\text{opt}}^2[\mathcal{X}] \geq \left[12\ln\left(\frac{17KM_*^2}{\text{Opt}}\right)\right]^{-1}\text{Opt}.$$

---

[10]Note that in the case of $\mathcal{X}$ defined as in (1), (20) implies the bound (∗) announced in the Introduction. Indeed, in this case for $Q \in \mathcal{Q}$ it holds $\kappa\,\text{Tr}(Q) \leq \text{Tr}([\sum_k S_k]Q) \leq K$ (recall that κ is the smallest eigenvalue of $\sum_k S_k$), implying that $\text{Tr}(Q) \leq K\kappa^{-1}$. Hence $M_*^2 = \max_{Q \in \mathcal{Q}}\text{Tr}(BQB^T) \leq \|B\|^2\max_{Q \in \mathcal{Q}}\text{Tr}(Q) \leq \|B\|^2K\kappa^{-1}$, so that (20) indeed implies (∗).

Note that all the quantities in the right-hand side of (21) are efficiently computable given the problem data, and that $\sqrt{\text{Opt}}$ is an upper bound on $\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}]$.

Furthermore, if a simple though less precise expression of the factor in terms of this data is required, it can be obtained as follows. Recall that two points $x = x_+$ and $x = -x_+$ of $\mathcal{X}$ can be distinguished through the observation $Ax + \sigma\xi$ with maximal probability of error $0 < \alpha < 1$ only if $\|Ax\|_2 \geq c_\alpha\sigma$, $c_\alpha > 0$;[11] by the standard argument one conclude that the risk of estimation of Bx satisfies, for some absolute constant $c > 0$:

$$(22) \qquad \text{Risk}_{\text{opt}}^2[\mathcal{X}] \geq \max\{\|Bx\|_2 : \|Ax\|_2 \leq c\sigma, x \in \mathcal{X}\}.$$

Now let $B = I$, and consider two typical for the traditional nonparametric statistics types of $\mathcal{X}$:

- $\mathcal{X}$ is the ellipsoid $\{x \in \mathbf{R}^n : \sum_i a_i^2 x_i^2 \leq 1\}$ with $0 < a_1 \leq a_2 \leq \cdots \leq a_n$ (for properly selected a_i this set models the restriction onto a regular n-point grid of functions from a Sobolev ball). Here, $K = 1$, $\mathcal{T} = [0, 1]$, $S_1 = \text{Diag}\{a_1^2, \ldots, a_n^2\}$. When choosing $x = te_1$, where e_1 is the first basic orth and $t \in]0, 1]$, using (22) we get $\text{Risk}_{\text{opt}}[\mathcal{X}] \geq \min[1/a_1, c\sigma/\|[A]_1\|_2]$ where $[A]_1$ is the first column of A. On the other hand, we have $M_*^2 = a_1^{-2}$, and the simplified risk bound reads

$$\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln\left(1 + \frac{\|[A]_1\|_2}{\sigma a_1}\right)} \text{Risk}_{\text{opt}}[\mathcal{X}].$$

- $\mathcal{X}$ is the box $\{x \in \mathbf{R}^n : a_i|x_i| \leq 1, 1 \leq i \leq n\}$, where, as above, $0 < a_1 \leq a_2 \leq \cdots \leq a_n$. Here, $K = n$, $\mathcal{T} = [0, 1]^n$, $x^T S_k x = a_k^2 x_k^2$, resulting in $M_*^2 = \sum_i a_i^{-2} \leq na_1^{-2}$. The same bound $\text{Risk}_{\text{opt}}[\mathcal{X}] \geq \min[1/a_1, c\sigma/\|[A]_1\|_2]$ holds in this case and, consequently,

$$\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln n + \ln\left(1 + \frac{\|[A]_1\|_2}{\sigma a_1}\right)} \text{Risk}_{\text{opt}}[\mathcal{X}].$$

Now let B be a general-type matrix, and assume for the sake of simplicity that B has trivial kernel. We associate with the data the following quantities:

- size of $\mathcal{T}$, $T = \max_{t \in \mathcal{T}} \sum_k t_k$, and $\varkappa$—the minimal eigenvalue of $\sum_k S_k$. Note that for any $x \in \mathcal{X}$, $\sum_k x^T S_k x \leq T$, thus the radius $r(\mathcal{X}) = \max_{x \in \mathcal{X}} \|x\|_2$ of $\mathcal{X}$ satisfies $r(\mathcal{X}) \leq \sqrt{T/\varkappa}$;
- ℓ_1/ℓ_∞-*condition number of* $\mathcal{T}$

$$\text{Cond}(\mathcal{T}) = \sqrt{\frac{T}{\max_{t \in \mathcal{T}} \min_{k \leq K} t_k}} = \sqrt{\frac{\max_{t \in \mathcal{T}} \sum_k t_k}{\max_{t \in \mathcal{T}} \min_{k \leq K} t_k}};$$

by our assumptions, $\mathcal{T}$ intersects the interior of $\mathbf{R}_+^K$ and thus $\sqrt{K} \leq \text{Cond}(\mathcal{T}) < \infty$;

---

[11] In fact, one can choose $c_\alpha = q_{1-\alpha}$, the $1 - \alpha$-quantile of the standard normal distribution.

- *condition number of B*: $\mathrm{Cond}(B) = \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)}$, where $\sigma_{\max}(B)$ and $\sigma_{\min}(B)$ are, respectively, the largest and the smallest singular values of B.

COROLLARY 2.1. *In the situation of this section*,

(23)
$$\mathrm{Risk}[\widehat{w}_{H_*} | \mathcal{X}]$$
$$\leq O(1) \sqrt{\ln \left(K \mathrm{Cond}^2(B) \left[\mathrm{Cond}^2(\mathcal{T}) + \frac{\|A\|^2 T}{\sigma^2 \varkappa} \right] \right)} \, \mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}];$$

here and in what follows, $O(1)$ *stands for a properly selected positive absolute constant*.

It is worth to note that, surprisingly, the logarithmic factor in (23) does not depend of the structure of singular spectrum of A, the entity which, as far as the role of A is concerned, is primarily responsible for $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$.

2. *Relaxing the symmetry requirement*. Sets $\mathcal{X}$ of the form (3)—we called them ellitopes—are symmetric w.r.t. the origin convex compacts of special structure. This structure is rather flexible, but the symmetry is "built in." We are about to demonstrate that, to some extent, the symmetry requirement can be relaxed. Specifically, assume instead of (3) that for some $\alpha \geq 1$ it holds

$$\underbrace{\left\{ x \in \mathbf{R}^n : \exists (y \in \mathbf{R}^{\bar{n}}, t \in \mathcal{T}) : x = Py \ \& \ y^T S_k y \leq t_k, 1 \leq k \leq K \right\}}_{\underline{\mathcal{X}}} \subset \mathcal{X} \subset \alpha \underline{\mathcal{X}},$$

with S_k and $\mathcal{T}$ possessing the properties postulated in Section 2.1. Let Opt and H_* be the optimal value and optimal solution of the optimization problem (8) associated with the data $S_1, \ldots, S_K, \mathcal{T}$ and matrices $\bar{A} = AP$, $\bar{B} = BP$ in the role of A, B, respectively. It is immediately seen that the risk $\mathrm{Risk}[\widehat{w}_{H_*} | \mathcal{X}]$ of the linear estimate $\widehat{w}_{H_*}(\omega)$ is at most $\alpha \sqrt{\mathrm{Opt}}$. On the other hand, we have $\mathrm{Risk}_{\mathrm{opt}}[\underline{\mathcal{X}}] \leq \mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$, and by Theorem 2.1 also $\sqrt{\mathrm{Opt}} \leq \sqrt{6 \ln(\frac{8M_*^2 K}{\mathrm{Risk}_{\mathrm{opt}}^2[\underline{\mathcal{X}}]})} \, \mathrm{Risk}_{\mathrm{opt}}[\underline{\mathcal{X}}]$. Taken together, these relations imply that

(24)
$$\mathrm{Risk}[\widehat{w}_{H^*} | \mathcal{X}] \leq \alpha \sqrt{6 \ln \left(\frac{8M_*^2 K \alpha}{\mathrm{Risk}_{\mathrm{opt}}^2[\mathcal{X}]} \right)} \, \mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}].$$

In other words, as far as the "level of nonoptimality" of efficiently computable linear estimates is concerned, signal sets $\mathcal{X}$ which can be approximated by ellitopes within a factor α *of order of 1* are nearly as good as the ellitopes. To give an example: it is known that whenever the intersection $\mathcal{X}$ of K elliptic cylinders $\{x : (x - c_k)^T S_k (x - c_k) \leq 1\}$, $S_k \succeq 0$, concentric or not, is bounded and has a nonempty interior, $\mathcal{X}$ can be approximated by an ellipsoid within the factor

$\alpha = K + 2\sqrt{K}$.[12] Assuming w.l.o.g. that the approximating ellipsoid is centered at the origin, the level of nonoptimality of a linear estimate is bounded by (24) with $O(1)K$ in the role of α. Note that bound (24) rapidly deteriorates when α grows, and this phenomenon to some extent "reflects the reality." For example, a perfect simplex $\mathcal{X}$ inscribed into the unit sphere in $\mathbf{R}^n$ is in-between two centered at the origin Euclidean balls with the ratio of radii equal to n (i.e., $\alpha = n$). It is immediately seen that with $A = B = I$, in the range $\sigma \leq n\sigma^2 \leq 1$ of values of n and σ, we have

$$\text{Risk}_{\text{opt}}[\mathcal{X}] \approx \sqrt{\sigma}, \qquad \text{Risk}_{\text{opt}}[\widehat{w}_{H_*}|\mathcal{X}] = O(1)\sqrt{n}\sigma,$$

with $\approx$ meaning "up to logarithmic in n/σ factor." In other words, for large $n\sigma$ linear estimates indeed are significantly [albeit not to the full extent of (24)] outperformed by nonlinear ones.

Another "bad for linear estimates" situation suggested by (20) is that where the description (3) of $\mathcal{X}$, albeit possible, requires a huge value of K. Here again, (20) reflects to some extent the reality: when $\mathcal{X}$ is the unit ℓ_1 ball in $\mathbf{R}^n$, (3) takes place with $K = 2^{n-1}$; consequently, the factor at $\text{Risk}_{\text{opt}}[\mathcal{X}]$ in the right-hand side of (20) becomes at least $\sqrt{n}$. On the other hand, in the range $\sigma \leq n\sigma^2 \leq 1$ of values of n, σ, and with $A = B = I$, the risks $\text{Risk}_{\text{opt}}[\mathcal{X}]$, $\text{Risk}_{\text{opt}}[\widehat{w}_{H_*}|\mathcal{X}]$ are basically the same as in the case of $\mathcal{X}$ being the perfect simplex inscribed into the unit sphere in $\mathbf{R}^n$, and linear estimates indeed are "heavily nonoptimal" when $n\sigma$ is large.

2.5. *Numerical illustration.* Observe that inequality (15) taken together with an efficiently computable upper bound δ_ρ for the probability that $\eta \notin \mathcal{X}$ for $\eta \sim \mathcal{N}(0, Q_\rho)$ yields a single-parametric family of lower bounds on $\text{Risk}_{\text{opt}}[\mathcal{X}]$:

$$\text{Risk}_{\text{opt}}^2[X] \geq \rho \, \text{Opt} - [1 + \sqrt{2\rho}q_{1-\delta_\rho/2}]^2 M_*^2 \delta_\rho.$$

We can compute the right-hand side for several values of ρ, take the largest of the resulting lower bounds on $\text{Risk}_{\text{opt}}[\mathcal{X}]$ and compare the result with the risk $\sqrt{\text{Opt}}$ of the efficiently computable linear estimate yielded by the optimal solution to (8). In this way, we hopefully will end up with less pessimistic assessment of the level of nonoptimality of linear estimates than the one yielded by (20). On the other hand, better lower bounds can be computed using directly inequality (12) of Lemma 2.1 along with an efficiently computable approximation of the constraint $\text{Prob}\{\eta \notin \mathcal{X}\} \leq \delta$ on the distribution $\mathcal{N}(0, Q)$ of η. Indeed, given $0 < \delta \leq 1/5$, suppose that $\mathcal{Q}_\delta$ is a convex subset of the positive semidefinite cone such that for

---

[12]Specifically, setting $F(x) = -\sum_{k=1}^K \ln(1 - (x - c_k)^T S_k(x - c_k)) : \text{int } \mathcal{X} \to \mathbf{R}$ and denoting by $\bar{x}$ the *analytic center* $\text{argmin}_{x \in \text{int } \mathcal{X}} F(x)$, one has

$$\{x : (x - \bar{x})^T F''(\bar{x})(x - \bar{x}) \leq 1\} \subset \mathcal{X} \subset \{x : (x - \bar{x})^T F''(\bar{x})(x - \bar{x}) \leq [K + 2\sqrt{K}]^2\}.$$

any $Q \in \mathcal{Q}_\delta$ and $\eta \sim \mathcal{N}(0, Q)$ one has $\mathrm{Prob}\{\eta \notin \mathcal{X}\} \leq \delta$. Then, according to (12), the quantity

$$(25) \qquad \mathrm{Opt}_\delta - [M_* + \sqrt{2} q_{1-\delta/2} \| BQ_\delta^{1/2} \|_2]^2 \delta,$$

where

$$\mathrm{Opt}_\delta = \max_{Q \in \mathcal{Q}_\delta} \varphi(Q)$$

and $\mathcal{Q}_\delta$ is the corresponding optimal solution, is a lower bound on $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$.

We have conducted two experiments aimed to compare the suboptimality factors obtained numerically with their theoretical counterparts. In both experiments, B and P are set to be $n \times n$ identity matrices, and $n \times n$ sensing matrix A is a randomly rotated matrix with singular values λ_j, $1 \leq j \leq n$, forming a geometric progression, with $\lambda_1 = 1$ and $\lambda_n = 0.01$. In the first experiment, the signal set $\mathcal{X}_1$ is an ellipsoid

$$\mathcal{X}_1 = \left\{ x \in \mathbf{R}^n : \sum_{j=1}^n j^2 x_j^2 \leq 1 \right\},$$

that is, $K = 1$, $S_1 = \sum_{j=1}^n j^2 e_j e_j^T$ (e_j are basic orths), and $\mathcal{T} = [0, 1]$. With two natural implementations of the outlined bounding scheme (for the sake of completeness, the details of the lower bound computation are provided in Section C of the Supplementary Material [Juditsky and Nemirovski (2018)]), we arrived at simulation results presented on Figures 1 and 2. It is worth to mention that the theoretical estimation of the "suboptimality factor" computed according to (21) varies in the interval [31.6, 73.7] in this experiment.
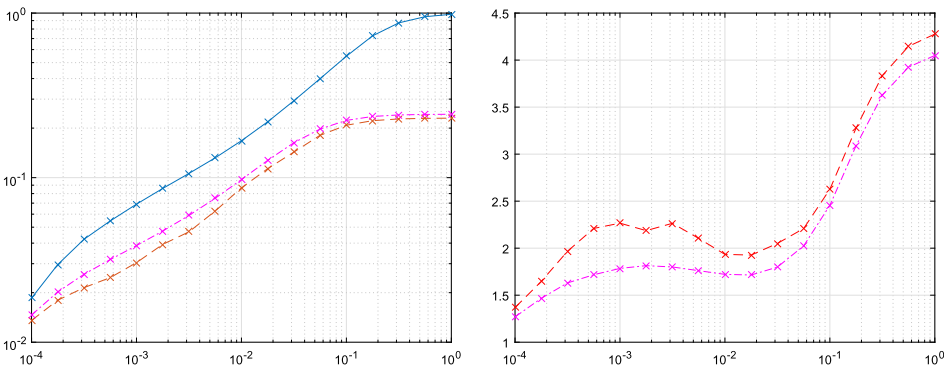

FIG. 1. *Recovery on ellipsoids: risk bounds as functions of the noise level σ, dimension $n = 32$. Left plot: upper bound of the risk of linear recovery (solid blue line); red dash line and magenta dash-dot line—lower bounds utilizing two implementations of the bounding scheme. Right plot: suboptimality ratios.*
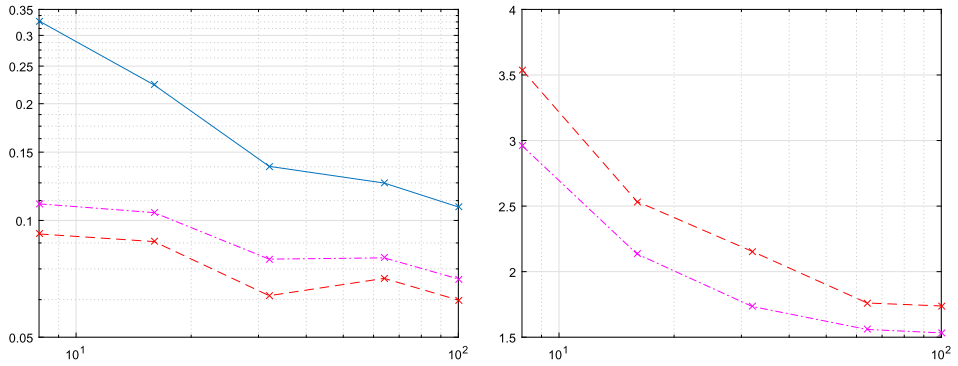
FIG. 2.   *Recovery on ellipsoids*: *risk bounds as functions of problem dimension n, noise level*
$\sigma = 0.01$. *Left plot*: *upper bound of the risk of linear recovery* (*solid blue line*), *red dash line and magenta dash-dot line—lower bounds on* $\mathrm{Risk}_{\mathrm{opt}}$ *utilizing two implementations of the bounding scheme.*
Right plot: *suboptimality ratios*.

In the second experiment, the signal set $\mathcal{X}$ is the box circumscribed around the above ellipsoid:

$$\mathcal{X} = \left\{ x \in \mathbf{R}^n : j|x_j| \leq 1, 1 \leq j \leq n \right\}$$
$$\left[K = n, S_k = k^2 e_k e_k^T, k = 1, \dots, K, \mathcal{T} = [0, 1]^K \right].$$

In this case, only one implementation of the bounding scheme is used. The simulation results of the second experiment are given on Figures 3 and 4. In this experiment also, the theoretical estimation of the nonoptimality of the linear estimate is very conservative—for different values of parameters the factor in the bound (21) varies between 73.2 and 115.4.
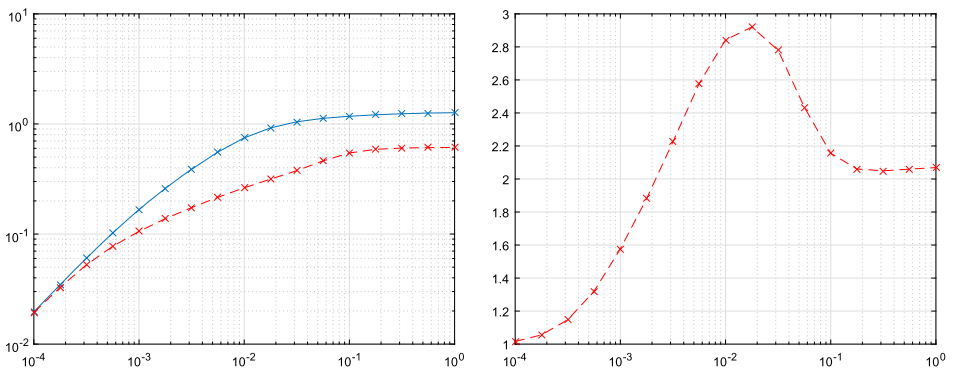

FIG. 3.   *Recovery on a box*: *risk bounds as functions of the noise level* σ, *dimension* $n = 32$. *Left plot*: *upper bound of the risk of linear recovery* (*solid blue line*) *and lower risk bound* (*red dash line*).
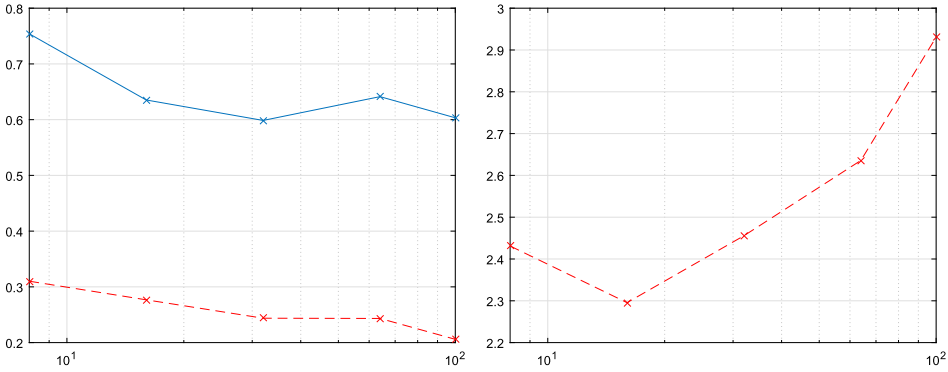*Right plot*: *suboptimality ratios*.

FIG. 4. *Recovery on a box*: risk bounds as functions of problem dimension n, noise level $\sigma = 0.01$. *Left plot*: upper bound of the risk of linear recovery (*solid blue line*) and lower bound on $\mathrm{Risk}_{\mathrm{opt}}$ (*red dash line*). *Right plot*: suboptimality ratio.

3. Extensions.

3.1. *Estimation in relative scale.* In this section, we consider the setting as follows. Assume that, same as in Section 2, we are given a $\nu \times n$ matrix B, and a noisy observation

$$\omega = Ax + \sigma\xi, \qquad \xi \sim \mathcal{N}(0, I_m),$$

of a signal $x \in \mathcal{X}$ with known $m \times n$ matrix A and $\sigma > 0$, and we aim to recover $w = Bx$. We are given a positive semidefinite symmetric $n \times n$ matrix S, and we quantify the quality of a candidate estimate $\widehat{w}(\cdot)$ by its *S-risk*—the quantity

$$\mathrm{RiskS}[\widehat{w}|\mathcal{X}] = \inf\{\sqrt{\tau} : \mathbf{E}\{\|\widehat{w}(Ax + \sigma\xi) - Bx\|_2^2\} \le \tau(1 + x^T S x)$$

(26)

$$\forall x \in \mathcal{X}\}.$$

The S-risk can be seen as risk with respect to the scale given by the "regularity parameter" $x^T S x$ of the unknown signal x. In particular, when $S = B^T B$, squared S-risk can be thought of as *relative* risk—the worst, over $x \in \mathbf{R}^n$, expected $\|\cdot\|_2^2$-error of recovering Bx scaled by $\|Bx\|_2^2$; when $S = 0$, we arrive at the usual risk $\mathrm{Risk}[\widehat{w}|\mathcal{X}]$.

As is the same as in Section 2, we assume w.l.o.g. that $\mathcal{X}$ is an ellitope given by (4).[13] Besides this, we assume that $B \neq 0$—otherwise the estimation problem is trivial.

We are about to prove that in the situation in question, efficiently computable linear estimate is near-optimal.

---

[13]To reduce the general case (3) to this one with $P = I$ it suffices to "lift" A, B, S to the y-space according to $A \mapsto \bar{A} = AP$, $B \mapsto \bar{B} = BP$, $S \mapsto \bar{S} = P^T S P$ and then replace $\mathcal{X}$ with the set $\mathcal{Y} = \{y \in \mathbf{R}^{\bar{n}} : \exists t \in \mathcal{T} : y^T S_k y \le t_k, 1 \le k \le K\}$.

3.1.1. *Building linear estimate.* Given a linear estimate $\widehat{w}_H(\omega) = H^T\omega$ and $\tau \geq 0$, let $\lambda \geq 0$ be such that $[B - H^T A]^T [B - H^T A] \preceq \sum_k \lambda_k S_k + \tau S$ [see (3)], implying that for all $x \in \mathcal{X}$, there exists $t = t_x \in \mathcal{T}$ such that

$$\mathbf{E}_\xi \{ \|\widehat{w}_H(Ax) - Bx\|_2^2 \} \leq x^T \Big[\sum_k \lambda_k S_k + \tau S \Big] x + \sigma^2 \operatorname{Tr}(H^T H)$$

$$\leq \sum_k t_k \lambda_k + x^T S x + \sigma^2 \operatorname{Tr}(H^T H),$$

so that for all $x \in \mathcal{X}$

$$\mathbf{E}_\xi \{ \|\widehat{w}_H(Ax + \sigma\xi) - Bx\|_2^2 \} \leq \phi_\mathcal{T}(\lambda) + \tau x^T S x + \sigma^2 \operatorname{Tr}(H^T H),$$

where $\phi_\mathcal{T}$ is the support function of $\mathcal{T}$. As a result, whenever H, $\tau \geq 0$ and $\lambda \geq 0$ are such that

$$\sigma^2 \operatorname{Tr}(H^T H) + \phi_\mathcal{T}(\lambda) \leq \tau, \qquad (H^T A - B)^T (H^T A - B) \preceq \sum_k \lambda_k S_k + \tau S,$$

we have

$$\operatorname{RiskS}[\widehat{w}_H | \mathcal{X}] \leq \sqrt{\tau}.$$

We arrive at the convex problem

$$(27) \qquad \operatorname{Opt} = \min_{\tau, H, \lambda} \left\{ \tau : \begin{bmatrix} \sum_k \lambda_k S_k + \tau S & B^T - A^T H \\ B - H^T A & I_v \end{bmatrix} \succeq 0, \\ \sigma^2 \operatorname{Tr}(H^T H) + \phi_\mathcal{T}(\lambda) \leq \tau, \lambda \geq 0 \right\}.$$

The H-component H_* of an optimal solution to this problem yields linear estimate $\widehat{w}_{H_*}(\omega) = H_*^T \omega$ with S-risk $\leq \sqrt{\operatorname{Opt}}$.

3.1.2. *Lower-bounding the optimal S-risk and near-optimality of $\widehat{w}_{H_*}$.* Consider the problem
(28)

$$\operatorname{Opt}_* = \max_{W, G, s, v} \left\{ \operatorname{Tr}(BWB^T) - \operatorname{Tr}(G) : \begin{bmatrix} G & BWA^T \\ AWB^T & \sigma^2 s I_m + AWA^T \end{bmatrix} \succeq 0, \\ W \succeq 0, \operatorname{Tr}(WS_k) \leq v_k, 1 \leq k \leq K, \\ \operatorname{Tr}(WS) + s \leq 1, [v; s] \in \mathbf{T} \right\},$$

where

$$(29) \qquad \mathbf{T} = \operatorname{cl}\{[t; \tau] \in \mathbf{R}^K \times \mathbf{R} : \tau > 0, \tau^{-1}t \in \mathcal{T}\} \subset \mathbf{R}_+^{K+1}$$

is a closed and pointed convex cone in $\mathbf{R}^{K+1}$ with a nonempty interior. We have the following counterpart of Lemma 2.2 for the present setting.

LEMMA 3.1.   *Problem* (28) *is strictly feasible and solvable. Furthermore*, *if* $(W, G, [v; s])$ *is an optimal solution to* (28), *then* $s > 0$, *and*

$$(30) \qquad \text{Opt} = \text{Opt}_* = \text{Tr}(B[W - WA^T(\sigma^2 s I_m + AWA^T)^{-1}AW]B^T).$$

Now let W, v and s stem from an optimal solution to (28). Then, as we have seen, $s > 0$, and we can set $t = v/s$, so that $t \in \mathcal{T}$. Let also $\rho \in (0, 1]$, and let us put $Q_\rho = \rho W/s$ and $\eta \sim \mathcal{N}(0, Q_\rho)$. We have $S^{-1}W \succeq 0$ and $\text{Tr}(s^{-1}WS_k) \le t_k$, $k \le K$, so that $s^{-1}W \in \mathcal{Q}$ and, therefore, $Q_\rho \in \rho\mathcal{Q}$. Hence, same as in the case of the usual risk, by Lemma 2.3,

$$(31) \qquad \text{Prob}\{\eta \notin \mathcal{X}\} \le \delta_\rho := \min\left[K\exp\left\{-\frac{1 - \rho + \rho\ln(\rho)}{2\rho}\right\}, 1\right].$$

We also have the following analog of Lemma 2.1.

LEMMA 3.2.   *Given* $\rho \in (0, 1]$, $Q \in \rho\mathcal{Q}$ *and* $\delta \le 1/5$, *let* $\eta \sim \mathcal{N}(0, Q)$ *and* $\xi \sim \mathcal{N}(0, I_m)$ *be independent from each other Gaussian vectors. Assume that*

$$\text{Prob}\{\eta \notin \mathcal{X}\} \le \delta.$$

Then

$$(32) \qquad \varphi(Q) \le \text{RiskS}^2_{\text{opt}}[\mathcal{X}](1 + \text{Tr}(QS)) + [1 + \sqrt{2\rho}q_{1-\delta/2}]^2 M_*^2\delta,$$

where M_* *is given by* (13), q_α, *same as in Lemma* 2.1, *is the* α-*quantile of the standard normal distribution, and*

$$\text{RiskS}_{\text{opt}}[\mathcal{X}] = \inf_{\widehat{w}(\cdot)} \text{RiskS}[\widehat{w}|\mathcal{X}]$$

is the minimax S-risk associated with $\mathcal{X}$.

For the proof, see Section A.1.1 of the Supplementary Material [Juditsky and Nemirovski (2018)].

Now note that

$$\begin{aligned}
\varphi(Q_\rho) &= \text{Tr}(B[Q_\rho - Q_\rho A^T(\sigma^2 I_m + AQ_\rho A^T)^{-1}AQ_\rho]B^T) \\
&= \frac{\rho}{s}\text{Tr}(B[W - \rho WA^T(s\sigma^2 I_m + \rho AWA^T)^{-1}AW]B^T) \\
&\ge \frac{\rho}{s}\text{Opt}_* = \frac{\rho}{s}\text{Opt}
\end{aligned}$$

[we have used (30) and the positivity of s]. Thus, when applying Lemma 3.2 with Q_ρ and δ_ρ in the role of Q and δ, we obtain for all $0 < \rho \le 1$ such that $\delta_\rho \le 1/5$

$$(33) \qquad \begin{aligned}
\frac{\rho}{s}\text{Opt} &\le \text{RiskS}^2_{\text{opt}}[\mathcal{X}](1 + \text{Tr}(Q_\rho S)) + [1 + \sqrt{2\rho}q_{1-\delta_\rho/2}]^2 M_*^2\delta_\rho \\
&= \text{RiskS}^2_{\text{opt}}[\mathcal{X}]\left(1 + \frac{\rho}{s}\text{Tr}(WS)\right) + [1 + \sqrt{\rho}q_{1-\delta_\rho/2}]^2 M_*^2\delta_\rho.
\end{aligned}$$

Similar to Section 2.3, setting

$$\bar{\rho}^{-1} = 3 \ln\left(\frac{8KM_*^2}{\text{RiskS}_{\text{opt}}^2[\mathcal{X}]}\right)$$

we ensure that

$$8KM_*^2 \exp\{-(3\bar{\rho})^{-1}\} \leq \text{RiskS}_{\text{opt}}^2[\mathcal{X}].$$

Now, same as in the case of usual risk, we clearly have $M_*^2 \geq \text{RiskS}_{\text{opt}}^2[\mathcal{X}]$, whence $\delta_{\bar{\rho}} \leq \exp\{-\frac{1}{3\bar{\rho}}\} \leq 1/5$, see (31), so that (33) is applicable with $\rho = \bar{\rho}$, thus implying that

$$\frac{\bar{\rho}}{s} \text{Opt} \leq \text{RiskS}_{\text{opt}}^2[\mathcal{X}]\left(1 + \frac{\bar{\rho}}{s}\text{Tr}(WS)\right) + 8KM_*^2 \exp\left\{-\frac{1}{3\bar{\rho}}\right\}$$

and

$$\bar{\rho} \text{Opt} \leq \text{RiskS}_{\text{opt}}^2[\mathcal{X}](s + \bar{\rho}\text{Tr}(WS)) + 8sKM_*^2 \exp\left\{-\frac{1}{3\bar{\rho}}\right\}$$

$$\leq \text{RiskS}_{\text{opt}}^2[\mathcal{X}] + 8KM_*^2 \exp\left\{-\frac{1}{3\bar{\rho}}\right\} = 2\text{RiskS}_{\text{opt}}^2[\mathcal{X}]$$

[note that $s + \bar{\rho}\text{Tr}(WS) \leq s + \text{Tr}(WS) \leq 1$ by constraints in (28)]. Recalling that $\sqrt{\text{Opt}}$ upper-bounds $\text{RiskS}[\widehat{w}_{H_*}|\mathcal{X}]$, we arrive at the following.

PROPOSITION 3.1. *The efficiently computable linear estimate* $\widehat{w}_{H_*}(\omega) = H_*^T \omega$ *yielded by an optimal solution to the optimization problem in* (27) *is nearly optimal in terms of S-risk*:

$$\text{RiskS}[\widehat{w}_{H_*}|\mathcal{X}] \leq \sqrt{6\ln\left(\frac{8KM_*^2}{\text{RiskS}_{\text{opt}}^2[X]}\right)} \text{RiskS}_{\text{opt}}[X],$$

where M_* is given by (13).

3.1.3. *The case of $\mathcal{X} = \mathbf{R}^n$.* The problem of minimizing the worst case, over $x \in \mathcal{X}$, S-risk over linear/all possible estimates makes sense for unbounded $\mathcal{X}$'s as well as for bounded ones. We intend to consider the case where $\mathcal{X} = \mathbf{R}^n$ and to show that in this case an efficiently computable linear estimate is *exactly optimal*.

Similar to (27), the problem of building the best, in terms of its worst case over $x \in \mathbf{R}^n$ S-risk, linear estimate reads

$$(34) \qquad \text{Opt} = \min_{\tau, H}\left\{\tau : \begin{bmatrix} \tau S & B^T - A^T H \\ B - H^T A & I_\nu \end{bmatrix} \succeq 0, \sigma^2 \text{Tr}(H^T H) \leq \tau\right\};$$

a feasible solution (τ, H) to this problem produces an estimate $\widehat{w}_H(\omega) = H^T \omega$ with $\text{RiskS}[\widehat{w}_H|\mathbf{R}^n] \leq \sqrt{\tau}$. We are about to demonstrate the following.

PROPOSITION 3.2. *Assuming problem* (34) *feasible, the problem is solvable, and its optimal solution* (Opt, H_*) *induces linear estimate* $\widehat{w}_{H_*}$ *which is minimax optimal*:

$$(35) \qquad \text{RiskS}[\widehat{w}_{H_*}|\mathbf{R}^n] = \sqrt{\text{Opt}} = \inf_{\widehat{w}(\cdot)} \text{RiskS}[\widehat{w}(\cdot)|\mathbf{R}^n].$$

It may be interesting to compare the optimal S-risk $\text{RiskS}[\widehat{w}_{H_*}|\mathbf{R}^n] = \sqrt{\text{Opt}}$ to the maximal risk $\text{Risk}[\widehat{w}_{H^*}|\mathcal{X}_S]$ of the optimal linear estimation of Bx over the ellipsoid $\mathcal{X}_S = \{x \in \mathbf{R}^n : x^T S x \leq 1\}$, so that H^* is the optimal solution to (8) with $K = 1$, $S_1 = S$ and $\mathcal{T} = [0, 1]$; note that in this case the optimal value in (8) is exactly $\text{Risk}[\widehat{w}_{H^*}|\mathcal{X}_S]$, and not just an upper bound on this risk. When comparing (8) with (34), one can easily see that both risks are equivalent up to a factor $\sqrt{2}$:

$$\text{RiskS}[\widehat{w}_{H_*}|\mathbf{R}^n] \leq \text{Risk}[\widehat{w}_{H^*}|\mathcal{X}_S] \leq \sqrt{2}\,\text{RiskS}[\widehat{w}_{H_*}|\mathbf{R}^n].$$

Note also that by the definition of S-risk, we have

$$\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}_S] \leq \sqrt{2}\,\text{RiskS}[\widehat{w}_{H_*}|\mathcal{X}_S] \leq \sqrt{2}\,\text{RiskS}[\widehat{w}_{H_*}|\mathbf{R}^n],$$

which combines with the above inequalities to imply that

$$\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}_S] \leq \sqrt{2}\,\text{Risk}[\widehat{w}_{H^*}|\mathcal{X}_S].$$

However, the estimate $\widehat{w}_{H_*}$ cannot be seen as *adaptive* over the family of "coaxial" ellipsoids $\mathcal{X}_S^\kappa = \{x \in \mathbf{R}^n : x^T S x \leq \kappa\}$, $\kappa \in K \subset \mathbf{R}_+$; see, for example, Lepskiĭ (1990). For instance, the maximal over $\mathcal{X}_S^\kappa$ risk $\text{Risk}[\widehat{w}_{H_*}|\mathcal{X}_S^\kappa]$ does not scale correctly for $\kappa \ll 1$ and $\kappa \gg 1$.

3.1.4. *Numerical illustration.* In the above considerations, we treated matrix S as part of the data. In fact, we can make S a variable restricted to reside in a given computationally tractable convex subset $\mathcal{S}$ of the positive semidefinite cone, and look for minimal, over linear estimates *and matrices* $S \in \mathcal{S}$, S-risk. This can be done as follows. We consider a parametric family of problems with τ in (27) being a parameter rather than a variable, and S being a variable restricted to reside in $\mathcal{S}$; then we apply bisection in τ to find the smallest value of τ for which the problem is feasible. With S and linear estimate yielded by this procedure, the S-risk of the estimate clearly possesses near-optimality properties completely similar to those we have just established for the case of fixed S.

As an illustration of these ideas, consider the following experiment. Let $[r; v]$ be state of pendulum with friction—the 2-dimensional continuous time dynamical system obeying the equations

$$\dot{r} = v,$$
$$\dot{v} = -\nu^2 r - \kappa v + w,$$

where w is the external input. Assuming this input constant on consecutive time intervals of duration Δ, the sequence $z_\tau = [r(\tau\Delta); v(\tau\Delta)]$, $\tau = 0, 1, \ldots$, obeys

finite-difference equation

$$z_\tau = P z_{\tau-1} + Q w_\tau, \qquad \tau = 1, 2, \ldots$$

with

$$P = \exp\left\{ \Delta \overbrace{\begin{bmatrix} 0 & 1 \\ -\nu^2 & -\kappa \end{bmatrix}}^{\vartheta} \right\},$$

$$Q = \int_0^\Delta \exp\{s\vartheta\} \begin{bmatrix} 0 \\ 1 \end{bmatrix} ds;$$

here, w_τ is the value of $w(\cdot)$ on the (continuous-time) interval $((\tau - 1)\Delta, \tau\Delta)$. Assume that we are observing corrupted by noise positions $r_\tau = r(\tau\Delta)$ of the pendulum on the discrete-time horizon $1 \le \tau \le T$ and want to recover the inputs w_s, $T - K + 1 \le s \le T$. Denoting by $x = [z_0; w_1; w_2; \ldots; w_T]$ the "signal" underlying our observations, we can easily build a $T \times (T + 2)$ matrix A and $1 \times (T + 2)$ matrices B_t such that the trajectory $r := [r_1; \ldots; r_T]$ of pendulum's positions is given by $r = Ax$, and $w_t = B_t x$. Given noisy observations

$$\omega = Ax + \sigma\xi, \qquad \xi \sim \mathcal{N}(0, I_T)$$

of pendulum's (discrete time) trajectory, we want to recover inputs w_t, $1 \le t \le T$, and their collections $w^K = [w_{T-K+1}; w_{T-K+2}; \ldots; w_T] = B^{(K)} x$.[14]

We intend to process our estimation problems by building the best, in terms of its S-risk taken over the entire space $\mathbf{R}^{T+2}$ of signals, estimate; in our design, S is not fixed in advance, but is instead restricted to be positive semidefinite with trace ≤ 1. Thus, the problems we want to solve are of the form [cf. (34)]:

$$(36) \qquad \mathrm{Opt}[B] = \min_{\tau, H, S} \left\{ \tau : \begin{array}{l} \begin{bmatrix} \tau S & B^T - A^T H \\ B - H^T A & I_T \end{bmatrix} \succeq 0, \\ \sigma^2 \mathrm{Tr}(H^T H) \le \tau, S \succeq 0, \mathrm{Tr}(S) \le 1 \end{array} \right\},$$

where B depends on what we want to recover ($B = B_t$ when recovering w_t, and $B = B^{(K)}$ when recovering w^K). By Proposition 3.2, the linear estimate $H_{B,*}^T \omega$ yielded by an optimal solution $(\mathrm{Opt}[B], H_{B,*}, S_{B,*})$ to the above (clearly solvable) problem is minimax optimal in terms of its S-risk $\mathrm{RiskS}[\cdot | \mathbf{R}^{T+2}]$ *taken with respect to $S = S_{B,*}$, and the corresponding minimax optimal risk is exactly* $\sqrt{\mathrm{Opt}[B]}$.

---

[14]Note that estimating w^K is not the same as a "stand alone" estimation of each individual entry in w^K.

The rationale behind restricting S to have its trace ≤ 1 is as follows. Imagine that we have reasons to believe that the entries in x "are of order of 1;" the simplest way to model this belief is to assume that x is uniformly distributed over the sphere $\mathcal{S}$ of radius $\sqrt{\dim x} = \sqrt{T+2}$. Under this assumption, the claim that an estimate $\widehat{w}(\cdot)$ has S-risk, taken over the entire space w.r.t. a matrix $S \succeq 0$ with $\operatorname{Tr}(S) \leq 1$, at most $\sqrt{\tau}$ means that

$$\mathbf{E}_{\xi \sim \mathcal{N}(0,I_T)}\{\|\widehat{w}(Ax + \sigma\xi) - B^K x\|_2^2\} \leq \tau(1 + x^T S x) \qquad \forall x.$$

This relation, after taking expectation over the uniformly distributed over $\mathcal{S}$ signal x, implies that the expectation, over both ξ and x, of the squared recovery risk is at most 2τ. Thus, optimising the S-risk over the linear estimates *and* $S \succeq 0$, $\operatorname{Tr}(S) \leq 1$, can be interpreted as safe minimization of the Bayesian risk taken w.r.t. a specific Bayesian prior (uniform distribution on $\mathcal{S}$). In this context, "safety" means that along with guarantees on the Bayesian risk, we get some meaningful upper bound on the expected $\|\cdot\|_2^2$-error of recovery applicable to *every individual* signal.

In view of the above considerations, with some terminology abuse, below we refer to the optimal value of (36) as to the *Bayesian risk* of recovering Bx.

In the experiment, we are about to report that we use $\Delta = 1$, $\kappa = 0.05$ and select ν to make the eigenfrequency of the pendulum equal to 1/8; free motion of the pendulum in the (r, v)-coordinates is shown on Figure 5. We used $\sigma = 0.075$, $T = 32$, and solved problem (36) for several "B-scenarios." The results are presented on Figure 5(b)–(d). Plots (b) and (c) show the bound $\sqrt{2 \operatorname{Opt}[B]}$ (see above) on the Bayesian risk along with the optimal value of the risk of optimal linear recovery of Bx for signals x from the ball $\mathcal{X}$ of radius $\sqrt{T+2}$, as given by the optimal values of (8) (blue). Plot (b) shows what happens when recovering individual inputs ($B = B_t$, $t = 1, 2, \ldots, T$) and displays the risks as functions of t; plot (c) shows the risks of recovering blocks $u^K = B^{(K)}x$ of inputs as functions of $K = 1, 2, 4, \ldots, 32$. Finally, plot (d) shows the eigenvalues of the S-components of optimal solutions to problems (36) with $B = B^{(K)}$.[15]

3.2. *Adding robustness.* In this section, we address the situation where the data A, B of problems (8) and (27) is not known exactly, and we are looking for estimates which are robust w.r.t. the corresponding data uncertainties. We lose nothing when restricting ourselves with problem (27), since (8) is the particular case $S = 0$ of (27), with ellitope $\mathcal{X}$ given by (3). We intend to focus on the simplest case of *unstructured norm-bounded uncertainty*

(37)
$$[A; B] := \begin{bmatrix} A \\ B \end{bmatrix} \in \mathcal{U}_r$$
$$= \{[A; B] = [A_*; B_*] + E^T \Delta F : \Delta \in \mathbf{R}^{p \times q}, \|\Delta\| \leq r\};$$

---

[15]With $B = B_t$, S-components of optimal solutions to (36) turn out to be of rank 1 for all t.
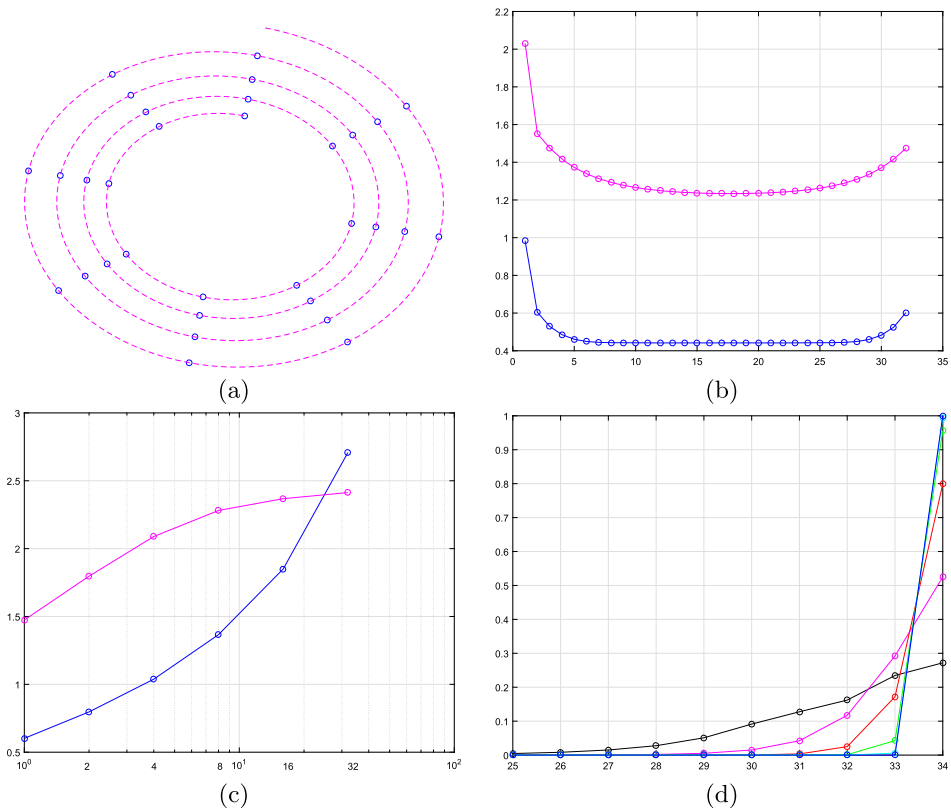
FIG. 5.   *Numerical illustration, Section* 3.1.4. (a): *free motion* ($w \equiv 0$) *of pendulum in* (r, v)-*plane in continuous* (*dashed line*) *and discrete* (*circles*) *time.* (b): *Bayesian* (*blue*) *and worst case* (*magenta*) *risks of recovering* w_t *vs.* $t = 1, 2, \ldots, 32$. (c): *Bayesian* (*blue*) *and worst-case* (*magenta*) *risks of recovering* $w^K := [w_{T-K+1}; w_{T-K+2}; \ldots; w_T]$ *vs.* K. (d): *eigenvalues* $\lambda_i(S_K)$ *of* S_K ($K = 32$—*black,* $K = 16$—*magenta,* $K = 8$—*red,* $K = 4$—*green,* $K = 2$—*cyan,* $K = 1$—*blue*); *we plot* 10 *largest eigenvalues of the S-matrices*; *the preceding* 24 *eigenvalues for all these matrices vanish.*

here $A_* \in \mathbf{R}^{m \times n}$, $B_* \in \mathbf{R}^{v \times n}$ are given *nominal data*, and $E \in \mathbf{R}^{p \times (m+v)}$, $F \in \mathbf{R}^{q \times n}$ are given matrices.[16] Our goal is to solve the *robust counterpart*

$$(38) \quad \text{RobOpt} = \min_{\tau, H, \lambda} \left\{ \tau : \begin{bmatrix} \sum_k \lambda_k S_k + \tau S & B^T - A^T H \\ B - H^T A & I_v \end{bmatrix} \succeq 0, \forall [A; B] \in \mathcal{U} \\ \sigma^2 \operatorname{Tr}(H^T H) + \phi_{\mathcal{T}}(\lambda) \leq \tau, \lambda \geq 0 \right\}$$

[16]Recall that in the case of $P \neq I$ we have to replace matrices A, B and S with AP, BP and $P^T SP$, respectively, and modify the definition of $\mathcal{U}_r$ accordingly: namely, when $[A; B]$ runs through the set $\mathcal{U}_r$, $[AP; BP]$ runs through

$$\overline{\mathcal{U}}_r = \{ [A; B] = [A_* P; B_* P] + E^T \Delta F P : \Delta \in \mathbf{R}^{p \times q}, \|\Delta\| \leq r \};$$

where A_*, B_* E and F are as in (37).

of problem (27). Plugging into (38) the parameterization of $[A; B]$ via Δ, the uncertainty-affected semidefinite constraint becomes

$$M(\lambda, \tau, H) + \mathcal{E}^T[H]\Delta\mathcal{F} + \mathcal{F}^T\Delta^T\mathcal{E}[H] \succeq 0 \qquad \forall(\Delta : \|\Delta\| \leq r),$$

$$(39) \quad M(\lambda, \tau, H) = \begin{bmatrix} \sum_k \lambda_k S_k + \tau S & B_*^T - A_*^T H \\ B_* - H^T A_* & I_\nu \end{bmatrix},$$

$$\mathcal{E}[H] = [0_{p\times n}, E_B - E_A H], \qquad \mathcal{F} = [F, 0_{q\times\nu}],$$

where

$$E = [E_A, E_B]$$

is the partitioning of the $p \times (m + \nu)$-matrix E into the blocks comprised by the first m and the last ν columns. A well-known result of Boyd et al. (1994) [see also Ben-Tal, El Ghaoui and Nemirovski (2009), Section 8.2.1] states that when $\mathcal{F} \neq 0$ (this is the only nontrivial case), the semi-infinite Linear Matrix Inequality in (39) holds true if and only if there exists μ such that

$$\begin{bmatrix} M(\lambda, \tau, H) - r^2\mu\mathcal{F}^T\mathcal{F} & [\mathcal{E}[H]]^T \\ \mathcal{E}[H] & \mu I_p \end{bmatrix} \succeq 0.$$

It follows that the semi-infinite convex problem (38) is equivalent to the explicit convex program

$$(40) \quad \text{RobOpt} = \min_{\tau, H, \lambda, \mu} \{\tau : \mathcal{G}(H, \mu, \tau) \succeq 0,$$

$$\sigma^2 \text{Tr}(H^T H) + \phi_{\mathcal{T}}(\lambda) \leq \tau, \lambda \geq 0\},$$

where

$$\mathcal{G}(H, \mu, \tau) = \begin{bmatrix} \sum_k \lambda_k S_k + \tau S - \mu r^2 F^T F & B_*^T - A_*^T H & \\ B_* - H^T A_* & I_\nu & E_B^T - H^T E_A^T \\ & E_B - E_A H & \mu I_p \end{bmatrix}.$$

The H-component of optimal solution to (40) yields robust w.r.t. uncertainty (37) estimate $H^T\omega$ of Bx via observation $Ax + \sigma\xi$, and the expected $\|\cdot\|_2^2$-error of this estimate does not exceed RobOpt, whatever be $x \in \mathcal{X}$ and $[A; B] \in \mathcal{U}$.

3.3. *Byproduct on semidefinite relaxation.* A byproduct of our main observation (Section 2.3) we are about to present has nothing to do with statistics; it relates to the quality of the standard semidefinite relaxation. Specifically, given a quadratic from $x^T C x$ and an ellitope $\mathcal{X}$ represented by (3), consider the problem:

$$(41) \quad \text{Opt}_* = \max_{x\in\mathcal{X}} x^T C x = \max_{y\in\mathcal{Y}} y^T P^T C P y.$$

This problem can be NP-hard (this is already so when $\mathcal{X}$ is the unit box and C is positive semidefinite); however, Opt admits an efficiently computable upper bound given by *semidefinite relaxation* as follows: whenever $\lambda \geq 0$ is such that

$$P^T C P \preceq \sum_{k=1}^{K} \lambda_k S_k,$$

for $y \in \mathcal{Y}$ we clearly have

$$[Py]^T C P y \leq \sum_k \lambda_k y^T S_k y \leq \phi_{\mathcal{T}}(\lambda)$$

due to the fact that the vector with the entries $y^T S_k y$, $1 \leq k \leq K$, belongs to $\mathcal{T}$. As a result, the efficiently computable quantity

$$(42) \qquad \text{Opt} = \min_{\lambda} \left\{ \phi_{\mathcal{T}}(\lambda) : \lambda \geq 0, P^T C P \preceq \sum_k \lambda_k S_k \right\}$$

is an upper bound on Opt_*. We have the following.

PROPOSITION 3.3. *Let C be a symmetric $n \times n$ matrix and $\mathcal{X}$ be given by elliptopic representation* (3), *and let* Opt_* *and* Opt *be given by* (41) *and* (42). *Then*

$$(43) \qquad \frac{\text{Opt}}{2 \ln K + 2\sqrt{\ln K} + 1} \leq \text{Opt}_* \leq \text{Opt}.$$

SUPPLEMENTARY MATERIAL

Proofs and auxiliary materials (DOI: 10.1214/17-AOS1596SUPP; .pdf). In the supplement, we provide full proofs of the results appearing in Sections 2 and 3. Because some of the proofs heavily rely upon Conic Duality, for the reader's convenience we present a brief outline of the related background. Other sections provide details of the calculus of ellitopes, and computations behind numerical lower bounds of Section 2.5.

REFERENCES

ARNOLD, B. F. and STAHLECKER, P. (2000). Another view of the Kuks–Olman estimator. *J. Statist. Plann. Inference* **89** 169–174. MR1794419

BEN-TAL, A., EL GHAOUI, L. and NEMIROVSKI, A. (2009). *Robust Optimization*. Princeton Univ. Press, Princeton, NJ. MR2546839

BOYD, S., EL GHAOUI, L., FERON, E. and BALAKRISHNAN, V. (1994). *Linear Matrix Inequalities in System and Control Theory*. SIAM Studies in Applied Mathematics **15**. SIAM, Philadelphia, PA. MR1284712

CHRISTOPEIT, N. and HELMES, K. (1996). Linear minimax estimation with ellipsoidal constraints. *Acta Appl. Math.* **43** 3–15. MR1385975

DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082

DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437. MR1062717

DRYGAS, H. (1996). Spectral methods in linear minimax estimation. *Acta Appl. Math.* **43** 17–42. MR1385976

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation*: *Methods*, *Theory*, *and Applications*. Springer, New York. MR1705298

EFROMOVICH, S. and PINSKER, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6** 925–942. MR1422411

GOLUBEV, Y. K., LEVIT, B. Y. and TSYBAKOV, A. B. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2** 167–181. MR1410136

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. *Applications of Mathematics* **16**. Springer, Berlin. MR0620321

JUDITSKY, A. and NEMIROVSKI, A. (2016). Near-optimality of linear recovery in Gaussian observation scheme under $\| \cdot \|_2$-loss. ArXiv Preprint ArXiv:1602.01355.

JUDITSKY, A. and NEMIROVSKI, A. (2018). Supplement to "Near-optimality of linear recovery in Gaussian observation scheme under $\| \cdot \|_2^2$-loss." DOI:10.1214/17-AOS1596SUPP.

KUKS, J. A. and OLMAN, W. (1971). Minimax linear estimation of regression coefficients (I). *Iswestija Akademija Nauk Estonskoj SSR* **20** 480–482.

KUKS, JA. and OLMAN, V. (1972). Linear minimax estimation of regression coefficients. *Eesti NSV Tead. Akad. Toim.*, *Füüs. Mat.* **21** 66–72. (errata insert). MR0319297

LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202

LIPTSER, R. S. and SHIRYAYEV, A. N. (1977). *Statistics of Random Processes*. *I*: *General Theory*. Springer, New York. MR0474486

PILZ, J. (1986). Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plann. Inference* **13** 297–318. MR0835614

PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 120–133.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York. MR0346957

RAO, C. R. (1976). Estimation of parameters in a linear model. *Ann. Statist.* **4** 1023–1037. MR0420979

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729

LJK
UNIVERSITÉ GRENOBLE-ALPES
700 AVENUE CENTRALE
38401 DOMAINE UNIVERSITAIRE
  DE SAINT-MARTIN-D'HÈRES
FRANCE
E-MAIL: anatoli.juditsky@univ-grenoble-alpes.fr

STEWART SCHOOL OF INDUSTRIAL
  AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
USA
E-MAIL: nemirovs@isye.gatech.edu