Journal of Machine Learning Research 17 (2016) 1-40 Submitted 7/14; Revised 6/15; Published 4/16

Loss Minimization and Parameter Estimation with Heavy Tails

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Department of Computer Science
Columbia University
New York, NY 10027, USA

Sivan Sabato SABATOS@CS.BGU.AC.IL
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva 8410501, Israel

Editor: David Dunson

Abstract

This work studies applications and generalizations of a simple estimation technique that provides exponential concentration under heavy-tailed distributions, assuming only bounded low-order moments. We show that the technique can be used for approximate minimization of smooth and strongly convex losses, and specifically for least squares linear regression. For instance, our d-dimensional estimator requires just $\tilde{O}(d \log(1/\delta))$ random samples to obtain a constant factor approximation to the optimal least squares loss with probability $1 - \delta$, without requiring the covariates or noise to be bounded or subgaussian. We provide further applications to sparse linear regression and low-rank covariance matrix estimation with similar allowances on the noise and covariate distributions. The core technique is a generalization of the median-of-means estimator to arbitrary metric spaces.

Keywords: Heavy-tailed distributions, unbounded losses, linear regression, least squares

1. Introduction

The minimax principle in statistical estimation prescribes procedures (*i.e.*, estimators) that minimize the worst-case risk over a large class of distributions generating the data. For a given loss function, the risk is the expectation of the loss of the estimator, where the expectation is taken over the data examined by the estimator. For example, for a large class of loss functions including squared loss, the empirical mean estimator minimizes the worst-case risk over the class of Gaussian distributions with known variance (Wolfowitz, 1950). In fact, Gaussian distributions with the specified variance are essentially the worst-case family of distributions for squared loss, at least up to constants (see, *e.g.*, Catoni, 2012, Proposition 6.1).

In this work, we are interested in estimators whose deviations from expected behavior are controlled with very high probability over the random draw of the data examined by the estimator. Deviations of the behavior of the estimator from its expected behavior are worrisome especially when data come from unbounded and/or heavy-tail distributions, where only very low order moments may be finite. For example, the Pareto distributions with shape parameter $\alpha > 0$ are unbounded and have finite moments only up to orders $< \alpha$; these distributions are commonly associated with the modeling of extreme events that manifest in data. Bounds on the expected behavior of an estimator are insufficient in these cases, since the high-probability guarantees that may be derived from such bounds (say, using Markov's inequality) are rather weak. For example, if the risk (*i.e.*, expected loss) of an estimator is bounded by ϵ, then all that we may derive from Markov's inequality is that the loss is no more than ϵ/δ with probability at least $1 - \delta$. For small values of $\delta \in (0, 1)$, the guarantee is not very reassuring, but it may be all one can hope for in these extreme scenarios—see Remark 7 in Section 3.1 for an example where this is tight. Much of the work in statistical learning theory is also primarily concerned with such high probability guarantees, but the bulk of the work makes either boundedness or subgaussian tail assumptions that severely limit the applicability of the results even in settings as simple as linear regression (see, *e.g.*, Srebro et al., 2010; Shamir, 2014).

Recently, it has been shown that it is possible to improve on methods which are optimal for expected behavior but suboptimal when high-probability deviations are concerned (Audibert and Catoni, 2011; Catoni, 2012; Brownlees et al., 2014). These improvements, which are important when dealing with heavy-tailed distributions, suggest that new techniques (*e.g.*, beyond empirical risk minimization) may be able to remove the reliance on boundedness or control of high-order moments. Bubeck et al. (2013) show how a more robust mean estimator can be used for solving the stochastic multi-armed bandit problem under heavy-tailed distributions.

This work applies and generalizes a technique for controlling large deviations from the expected behavior with high probability, assuming only bounded low-order moments such as variances. We show that the technique is applicable to minimization of smooth and strongly convex losses, and derive specific loss bounds for least squares linear regression, which match existing rates, but without requiring the noise or covariates to be bounded or subgaussian. This contrasts with several recent works (Srebro et al., 2010; Hsu et al., 2014; Shamir, 2014) concerned with (possibly regularized) empirical risk minimizers that require such assumptions. It is notable that in finite dimensions, our result implies that a constant factor approximation to the optimal loss can be achieved with a sample size that is independent of the size of the optimal loss. This improves over the recent work of Mahdavi and Jin (2013), which has a logarithmic dependence on the optimal loss, as well as a suboptimal dependence on specific problem parameters (namely condition numbers). We also provide a new generalization of the basic technique for general metric spaces, which we apply to least squares linear regression with heavy tail covariate and noise distributions, yielding an improvement over the computationally expensive procedure of Audibert and Catoni (2011).

The basic technique, found in the textbook of Nemirovsky and Yudin (1983, p. 243), is very simple, and can be viewed as a generalization of the median-of-means estimator used by Alon et al. (1999) and many others. The idea is to repeat an estimate several times, by splitting the sample into several groups, and then selecting a single estimator out of the resulting list of candidates. If an estimator from one group is good with noticeably better-than-fair chance, then the selected estimator will be good with probability exponentially close to one. This is remininscant of techniques from *robust statistics* (Huber, 1981), although our aim is expressly different in that our aim is good performance on the same probability distribution generating the data, rather than an uncontaminated or otherwise better behaved distribution. Our new technique can be cast as a simple selection problem in general metric spaces that generalizes the scalar median.

We demonstrate the versatility of our technique by giving further examples in sparse linear regression (Tibshirani, 1996) under heavy-tailed noise and low-rank covariance covariance matrix approximation (Koltchinskii et al., 2011) under heavy-tailed covariate distributions. We also show that for prediction problems where there may not be a reasonable metric on the predictors, one can achieve similar high-probability guarantees by using median aggregation in the output space.

The initial version of this article (Hsu and Sabato, 2013, 2014) appeared concurrently with the simultaneous and independent work of Minsker (2013), which develops a different generalization of the median-of-means estimator for Banach and Hilbert spaces. We provide a new analysis and comparison of this technique to ours in Section 7. We have also since become aware of the earlier work by Lerasle and Oliveira (2011), which applies the median-of-means technique to empirical risks in various settings much like the way we do in Algorithm 3, although our metric formulation is more general. Finally, the recent work of Brownlees et al. (2014) vastly generalizes the techniques of Catoni (2012) to apply to much more general settings, although they retain some of the same deficiencies (such as the need to know the noise variance for the optimal bound for least squares regression), and hence their results are not directly comparable to ours.

2. Overview of Main Results

This section gives an overview of the main results.

2.1 Preliminaries

Let $[n] := \{1, 2, \ldots, n\}$ for any natural number $n \in \mathbb{N}$. Let $\mathbb{1}\{P\}$ take value 1 if the predicate P is true, and 0 otherwise. Assume an example space $\mathcal{Z}$, and a distribution $\mathcal{D}$ over the space. Further assume a space of predictors or estimators $\mathbb{X}$. We consider learning or estimation algorithms that accept as input an i.i.d. sample of size n drawn from $\mathcal{D}$ and a confidence parameter $\delta \in (0, 1)$, and return an estimator (or predictor) $\hat{\boldsymbol{w}} \in \mathbb{X}$. For a (pseudo) metric ρ on $\mathbb{X}$, let $B_\rho(\boldsymbol{w}_0, r) := \{\boldsymbol{w} \in \mathbb{X} : \rho(\boldsymbol{w}_0, \boldsymbol{w}) \leq r\}$ denote the ball of radius r around $\boldsymbol{w}_0$.

We assume a loss function $\ell : \mathcal{Z} \times \mathbb{X} \to \mathbb{R}_+$ that assigns a non-negative number to a pair of an example from $\mathcal{Z}$ and a predictor from $\mathbb{X}$, and consider the task of finding a predictor that has a small loss in expectation over the distribution of data points, based on an input sample of n examples drawn independently from $\mathcal{D}$. The expected loss of a predictor $\boldsymbol{w}$ on the distribution is denoted $L(\boldsymbol{w}) = \mathbb{E}_{Z\sim D}(\ell(Z, \boldsymbol{w}))$. Let $L_\star := \inf_{\boldsymbol{w}} L(\boldsymbol{w})$. Our goal is to find $\hat{\boldsymbol{w}}$ such that $L(\hat{\boldsymbol{w}})$ is close to $L_\star$.

In this work, we are interested in performance guarantees that hold with high probability over the random draw of the input sample and any internal randomization used by the estimation algorithm. Thus, for a given allowed probability of failure $\delta \in (0, 1)$, we study excess loss $L(\hat{\boldsymbol{w}}) - L_\star$ achieved by the predictor $\hat{\boldsymbol{w}} \equiv \hat{\boldsymbol{w}}(\delta)$ returned by the algorithm on a $1 - \delta$ probability subset of the sample space. Ideally, the excess loss only depends sublogarithmically on $1/\delta$, which is the dependence achieved when the distribution of the excess loss has exponentially decreasing tails. Note that we assume that the value of δ is provided as input to the estimation algorithm, and only demand the probabilistic guarantee for this given value of δ. Therefore, strictly speaking, the excess loss need not exhibit exponential concentration. Nevertheless, in this article, we shall say that an estimation algorithm achieves exponential concentration whenever it guarantees, on input δ, an excess loss that grows only as $\log(1/\delta)$.

2.2 Robust Distance Approximation

Consider an estimation problem, where the goal is to estimate an unknown parameter of the distribution, using a random i.i.d. sample from that distribution. We show throughout this work that for many estimation problems, if the sample is split into non-overlapping subsamples, and estimators are obtained independently from each subsample, then with high probability, this generates a set of estimators such that some fraction of them are close, under a meaningful metric, to the true, unknown value of the estimated parameter. Importantly, this can be guaranteed in many cases even under under heavy-tailed distributions.

Having obtained a set of estimators, a fraction of which are close to the estimated parameter, the goal is now to find a single good estimator based on this set. This goal is captured by the following general problem, which we term *Robust Distance Approximation.* A Robust Distance Approximation procedure is given a set of points in a metric space and returns a single point from the space. This single point should satisfy the following condition: If there is an element in the metric space that a certain fraction of the points in the set are close to, then the output point should also be close to the same element. Formally, let $(\mathbb{X}, \rho)$ be a metric space. Let $W \subseteq \mathbb{X}$ be a (multi)set of size k and let $w_\star$ be a distinguished element in $\mathbb{X}$. For $\alpha \in (0, \frac{1}{2})$ and $w \in \mathbb{X}$, denote by $\Delta_W(w, \alpha)$ the minimal number r such that $|\{v \in W \mid \rho(w, v) \leq r\}| > k(\frac{1}{2} + \alpha)$. We often omit the subscript W and write simply Δ when W is known.

We define the following problem:

Definition 1 (Robust Distance Approximation) *Fix $\alpha \in (0, \frac{1}{2})$. Given W and $(\mathbb{X}, \rho)$ as input, return $y \in \mathbb{X}$ such that $\rho(y, w_\star) \leq C_\alpha \cdot \Delta_W(w_\star, \alpha)$, for some constant $C_\alpha \geq 0$. C_α is the* approximation factor *of the procedure.*

In some cases, learning with heavy-tailed distributions requires using a metric that depends on the distribution. Then, the Robust Distance Estimation procedure has access only to noisy measurements of distances in the metric space, and is required to succeed with high probability. In Section 3 we formalize these notions, and provide simple implementations of Robust Distance Approximation for general metric spaces, with and without direct access to the metric. For the case of direct access to the metric our formulation is similar to that of Nemirovsky and Yudin (1983).

2.3 Convex Loss Minimization

The general approach to estimation described above has many applications. We give here the general form of our main results for applications, and defer the technical definitions and results to the relevant sections. Detailed discussion of related work for each application is also provided in the appropriate sections.

First, we consider smooth and convex losses. We assume that the parameter space $\mathbb{X}$ is a Banach space with a norm $\|\cdot\|$ and a dual norm $\|\cdot\|_*$. We prove the following result:[1]

Theorem 2 *There exists an algorithm that accepts as input an i.i.d. sample of size n drawn from $\mathcal{D}$ and a confidence parameter $\delta \in (0,1)$, and returns $\hat{\boldsymbol{w}} \in \mathbb{X}$, such that if the following conditions hold:*

- *the dual norm $\|\cdot\|_*$ is γ-smooth;*
- *there exists $\alpha > 0$ and sample size n_α such that, with probability at least $1/2$, the empirical loss $\boldsymbol{w} \mapsto \hat{L}(\boldsymbol{w})$ is α-strongly convex with respect to $\|\cdot\|$ whenever the sample is of size at least n_α;*
- *$n \geq C \log(1/\delta) \cdot n_\alpha$ for some universal constant $C > 0$;*
- *$\boldsymbol{w} \mapsto \ell(z, \boldsymbol{w})$ is β-smooth with respect to $\|\cdot\|$ for all $z \in \mathcal{Z}$;*
- *$\boldsymbol{w} \mapsto L(\boldsymbol{w})$ is $\bar{\beta}$-smooth with respect to $\|\cdot\|$;*

then with probability at least $1-\delta$, for another universal constant $C' > 0$,

$$L(\hat{\boldsymbol{w}}) \leq \left(1 + \frac{C'\beta\bar{\beta}\gamma\lceil\log(1/\delta)\rceil}{n\alpha^2}\right) L_\star.$$

This gives a constant approximation of the optimal loss with a number of samples that does not depend on the value of the optimal loss. The full results for smooth convex losses are provided in Section 4. Theorem 2 is stated in full as Corollary 16, and we further provide a result with more relaxed smoothness requirements. As apparent in the result, the only requirements on the distribution are those that are implied by the strong convexity and smoothness parameters. This allows support for fairly general heavy-tailed distributions, as we show below.

2.4 Least Squares Linear Regression

A concrete application of our analysis of smooth convex losses is linear regression. In linear regression, $\mathbb{X}$ is a Hilbert space with an inner product $\langle\cdot,\cdot\rangle_{\mathbb{X}}$, and it is both the data space and the parameter space. The loss $\ell \equiv \ell^{\mathrm{sq}}$ is the squared loss

$$\ell^{\mathrm{sq}}((\boldsymbol{x}, y), \boldsymbol{w}) := \frac{1}{2}(\boldsymbol{x}^\top \boldsymbol{w} - y)^2.$$

L^{sq} and $L_\star^{\mathrm{sq}}$ are defined similarly to L and $L_\star$.

Unlike standard high-probability bounds for regression, we give bounds that make no assumption on the range or the tails of the distribution of the response variables, other than a trivial requirement that the optimal squared loss be finite. The assumptions on the distribution of the covariates are also minimal.

1. Formal definitions of terms used in the conditions are given in Section 4.

Let Σ be the second-moment operator $\boldsymbol{a} \mapsto \mathbb{E}(\boldsymbol{X}\langle \boldsymbol{X}, \boldsymbol{a}\rangle_{\mathbb{X}})$, where $\boldsymbol{X}$ is a random data point from the marginal distribution of $\mathcal{D}$ on $\mathbb{X}$. For a finite-dimensional $\mathbb{X}$, Σ is simply the (uncentered) covariance matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$. First, consider the finite-dimensional case, where $\mathbb{X} = \mathbb{R}^d$, and assume Σ is not singular. Let $\|\cdot\|_2$ denote the Euclidean norm in $\mathbb{R}^d$. Under only bounded $4+\epsilon$ moments of the marginal on $\mathbb{X}$ (a condition that we specify in full detail in Section 5), we show the following guarantee.

Theorem 3 *Assume the marginal of $\mathbb{X}$ has bounded $4+\epsilon$ moments. There is a constant $C > 0$ and an algorithm that accepts as input a sample of size n and a confidence parameter $\delta \in (0,1)$, and returns $\hat{\boldsymbol{w}} \in \mathbb{X}$, such that if $n \geq Cd\log(1/\delta)$, with probability at least $1-\delta$,*

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq L^{\mathrm{sq}}_\star + O\left(\frac{\mathbb{E}(\|\Sigma^{-1/2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)\|_2^2)\log(1/\delta)}{n}\right).$$

This theorem is stated in full as Theorem 19 in Section 5. Under standard finite fourth-moment conditions, this result translates to the bound

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{d\log(1/\delta)}{n}\right)\right)L^{\mathrm{sq}}_\star,$$

with probability $\geq 1-\delta$. These results improve over recent results by Audibert and Catoni (2011), Catoni (2012), and Mahdavi and Jin (2013). We provide a full comparison to related work in Section 5.

Theorem 3 can be specialized for specific cases of interest. For instance, suppose $\boldsymbol{X}$ is bounded and well-conditioned in the sense that there exists $R < \infty$ such that $\Pr[\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X} \leq R^2] = 1$, but Y may still be heavy-tailed. Under this assumption we have the following result.

Theorem 4 *Assume Σ is not singular. There exists an algorithm that accepts as input a sample of size n and a confidence parameter $\delta \in (0,1)$, and returns $\hat{\boldsymbol{w}} \in \mathbb{X}$, such that with probability at least $1-\delta$, for $n \geq O(R^2\log(R)\log(e/\delta))$,*

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{R^2\log(1/\delta)}{n}\right)\right)L^{\mathrm{sq}}_\star.$$

This theorem is stated in full as Theorem 20 in Section 5. Note that

$$\mathbb{E}(\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{X}) = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{X}) = \mathrm{tr}(\mathrm{Id}) = d,$$

so $R = \Omega(\sqrt{d})$. R^2 is closely related to a *condition number* for the distribution of $\boldsymbol{X}$. For instance, if $\mathbb{P}[\|\boldsymbol{X}\| = 1] = 1$, then $R^2 \leq d\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$. This result is minimax optimal up to logarithmic factors (see, *e.g.*, Nussbaum, 1999). We also remark that the boundedness assumption can be replaced by a subgaussian assumption on $\boldsymbol{X}$, in which case the sample size requirement becomes $O(d\log(1/\delta))$. We give analogous guarantees for the case of regularized least squares in a possibly finite-dimensional Hilbert space in Theorem 21, Section 5.

2.5 Other Applications, Comparisons, and Extensions

The general method studied here allows handling heavy tails in other applications as well. We give two examples in Section 6. First, we consider parameter estimation using L^1-regularized linear least squares regression (Lasso) under random subgaussian design. We show that using the above approach, parameter estimation bounds can be guaranteed for general bounded variance noise, including heavy-tailed noise. This contrasts with standard results that assume sub-Gaussian noise. Second, we show that low-rank covariance matrix approximation can be obtained for heavy-tailed distributions, under a bounded $4 + \epsilon$ moment assumption. These two applications have been analyzed also in the independent and simultaneous work of Minsker (2013).

All the results above are provided using a specific solution to the Robust Distance Approximation problem, which is easy to implement for any metric space. For the case of a fully known metric, in a Banach or a Hilbert space, Minsker (2013) proposed a different solution, which is based on the geometric median. In Section 7, we provide a detailed comparison of the approximation factor achieved by each approach, as well as some general lower bounds. Several interesting open questions remain regarding this general problem.

Lastly, in Section 8, we give a short proof to the intuitive fact that in some prediction problem, one can replace Robust Distance Approximation with taking the median of the predictions of the input estimators. This gives a possible improper-learning algorithm for relevant learning settings.

All of the techniques we have developed in this work are simple enough to implement and empirically evaluate, and indeed in some simulated experiments, we have verified the improvements over standard methods such as the empirical mean when the data follow heavy-tailed distributions. However, at present, the relatively large constant factors in our bounds are real enough to restrict the empirical improvements only to settings where very high confidence (*i.e.*, small values of δ) is required. By contrast, with an appropriately determined noise variance, the techniques of Catoni (2012) and Brownlees et al. (2014) may yield improvements more readily. Nevertheless, since our techniques are more general in some respects, it is worth investigating whether they can be made more practical (*e.g.*, with greater sample reuse or overlapping groups), and we plan to do this in future work.

3. The Core Techniques

In this section we present the core technique used for achieving exponential concentration. We first demonstrate the underlying principle via the median-of-means estimator, and then explain the generalization to arbitrary metric spaces. Finally, we show a new generalization that supports noisy feature measurements.

3.1 Warm-up: Median-of-Means Estimator

We first motivate the estimation procedure by considering the special case of estimating a scalar population mean using a *median-of-means* estimator, given in Algorithm 1. This estimator, heavily used in the streaming algorithm literature (Alon et al., 1999) (though a similar technique also appears in Nemirovsky and Yudin (1983) as noted in Levin (2005)), partitions a sample into k equal-size groups, and returns the median of the sample means

Algorithm 1 Median-of-means estimator

input Sample $S \subset \mathbb{R}$ of size n, number of groups $k \in \mathbb{N}$ such that $k \leq n/4$.
output Population mean estimate $\hat{\mu} \in \mathbb{R}$.

1: Randomly partition S into k subsets $S_1, S_2, \ldots, S_k$, each of size at least $\lfloor n/k \rfloor$.
2: For each $i \in [k]$, let $\mu_i \in \mathbb{R}$ be the sample mean of S_i.
3: Return $\hat{\mu} := \text{median}\{\mu_1, \mu_2, \ldots, \mu_k\}$.

of each group. Note that the possible non-uniqueness of the median does not affect the result; the arguments below apply to any one of them. The input parameter k should be thought of as a constant determined by the desired confidence level (*i.e.*, $k = \Theta(\log(1/\delta))$ for confidence $\delta \in (0, 1)$). It is well known that the median-of-means achieves estimation with exponential concentration. The following proposition gives a simple statement and proof. The constant 6 in the statement (see Eq. (1) below) is lower the constant in the analysis of Lerasle and Oliveira (2011, Proposition 1), which is $2\sqrt{6e} \approx 8.08$, but we require a larger value of n. By requiring an even larger n, the constant in the statement below can approach $3\sqrt{3}$.

Proposition 5 *Let x be a random variable with mean μ and variance $\sigma^2 < \infty$, and let S be a set of n independent copies of x. Assume $k \leq n/2$. With probability at least $1 - e^{-k/4.5}$, the estimate $\hat{\mu}$ returned by Algorithm 1 on input (S, k) satisfies $|\hat{\mu} - \mu| \leq \sigma\sqrt{8k/n}$. Therefore, if $k = 4.5\lceil \log(1/\delta) \rceil$ and $n \geq 18\lceil \log(1/\delta) \rceil$, then with probability at least $1 - \delta$,*

$$|\hat{\mu} - \mu| \leq 6\sigma\sqrt{\frac{\lceil \log(1/\delta) \rceil}{n}}. \tag{1}$$

Proof First, assume k divides n. Pick any $i \in [k]$, and observe that S_i is an i.i.d. sample of size n/k. Therefore, by Chebyshev's inequality, $\Pr[|\mu_i - \mu| \leq \sqrt{6\sigma^2 k/n}] \geq 5/6$. For each $i \in [k]$, let $b_i := \mathbb{1}\{|\mu_i - \mu| \leq \sqrt{6\sigma^2 k/n}\}$. Note that the b_i are independent indicator random variables, each with $\mathbb{E}(b_i) \geq 5/6$. By Hoeffding's inequality, $\Pr[\sum_{i=1}^k b_i > k/2] \geq 1 - e^{-k/4.5}$. In the event that $\sum_{i=1}^k b_i > k/2$, at least half of the μ_i are within $\sqrt{6\sigma^2 k/n}$ of μ, which means that the same holds for the median of the μ_i. If k does not divide n then the analysis can be carried out by substituting n with $\lfloor n/k \rfloor k \geq n - k \geq \frac{3}{4}n$, which scales the guarantee by a factor of $\sqrt{4/3}$. ■

Using the terminology of Robust Distance Approximation with the metric $\rho(x, y) = |x - y|$, the proof shows that with high probability over the choice of W, $\Delta_W(\mu, 0) \leq \sqrt{12\sigma^2 k/n}$. The result then immediately follows because on the space $(\mathbb{R}, \rho)$, the median is a Robust Distance Approximation procedure with $C_0 = 1$.

Remark 6 (Catoni's M-estimator) *Catoni (2012) proposes a mean estimator $\hat{\mu}$ that satisfies $|\hat{\mu} - \mu| = O(\sigma\sqrt{\log(1/\delta)/n})$ with probability at least $1 - \delta$. Remarkably, the leading constant in the bound is asymptotically optimal: it approaches $\sqrt{2}$ as $n \to \infty$. However, the estimator takes both δ and σ as inputs. Catoni also presents an estimator that takes only σ as an input; this estimator guarantees a $O(\sigma \log(1/\delta)/\sqrt{n})$ bound for all values of $\delta > \exp(1 - n/2)$ simultaneously.*

Remark 7 (Empirical mean) *Catoni (2012) shows that the empirical mean cannot provide a qualitatively similar guarantee. Specifically, for any $\sigma > 0$ and $\delta \in (0, 1/(2e))$, there is a distribution with mean zero and variance σ^2 such that the empirical average $\hat{\mu}_{\mathrm{emp}}$ of n i.i.d. draws satisfies*

$$\Pr\left[|\hat{\mu}_{\mathrm{emp}}| \geq \frac{\sigma}{\sqrt{2n\delta}}\left(1 - \frac{2e\delta}{n}\right)^{\frac{n-1}{2}}\right] \geq 2\delta. \tag{2}$$

Therefore the deviation of the empirical mean necessarily scales with $1/\sqrt{\delta}$ rather than $\sqrt{\log(1/\delta)}$ (with probability $\Omega(\delta)$).

3.2 Generalization to Arbitrary Metric Spaces

We now consider a simple generalization of the median-of-means estimator for arbitrary metric spaces, first mentioned in Nemirovsky and Yudin (1983). Let $\mathbb{X}$ be the parameter (solution) space, $\boldsymbol{w}_\star \in \mathbb{X}$ be a distinguished point in $\mathbb{X}$ (the target solution), and ρ a metric on $\mathbb{X}$ (in fact, a pseudometric suffices).

The first abstraction captures the generation of candidate solutions obtained from independent subsamples. We assume there is an oracle $\mathrm{APPROX}_{\rho,\varepsilon}$ which satisfies the following assumptions.

Assumption 1 *A query to* $\mathrm{APPROX}_{\rho,\varepsilon}$ *returns a random $\boldsymbol{w} \in \mathbb{X}$ such that*

$$\Pr\Big[\rho(\boldsymbol{w}_\star, \boldsymbol{w}) \leq \varepsilon\Big] \geq 2/3.$$

Note that the 2/3 could be replaced by another constant larger than half; we have not optimized the constants. The second assumption regards statistical independence. For an integer k, let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ be responses to k separate queries to $\mathrm{APPROX}_{\rho,\varepsilon}$.

Assumption 2 *$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are statistically independent.*

The proposed procedure, given in Algorithm 2, generates k candidate solutions by querying $\mathrm{APPROX}_{\rho,\varepsilon}$ k times, and then selecting a single candidate using a generalization of the median. Specifically, for each $i \in [k]$, the smallest ball centered at $\boldsymbol{w}_i$ that contains more than half of $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$ is determined; the $\boldsymbol{w}_i$ with the smallest such ball is returned. If there are multiple such $\boldsymbol{w}_i$ with the smallest radius ball, any one of them may be selected. This selection method is a Robust Distance Approximation procedure. The proof is given below and illustrated in Figure 1. Nemirovsky and Yudin (1983) proposed a similar technique, however their formulation relies on knowledge of ε.

Proposition 8 *Let $r_i := \min\{r \geq 0 : |B_\rho(\boldsymbol{w}_i, r) \cap W| > k/2\}$. Selecting $\boldsymbol{w}_{i_\star}$ such that $i_\star = \operatorname{argmin}_i r_i$ is a Robust Distance Approximation procedure with $C_0 = 3$.*

Proof Assume that $\Delta(w_\star, 0) \leq \varepsilon$. Then $|B_\rho(\boldsymbol{w}_\star, \varepsilon) \cap W| > k/2$. For any $\boldsymbol{v} \in B_\rho(\boldsymbol{w}_\star, \varepsilon) \cap W$, by the triangle inequality, $|B_\rho(\boldsymbol{v}, 2\varepsilon) \cap W| > k/2$. This implies that $r_{i_\star} \leq 2\varepsilon$, and so $|B_\rho(\boldsymbol{w}_{i_\star}, 2\varepsilon) \cap W| > k/2$. By the pigeonhole principle, $B_\rho(\boldsymbol{w}_\star, \varepsilon) \cap B_\rho(\boldsymbol{w}_{i_\star}, 2\varepsilon) \neq \emptyset$. Therefore, by the triangle inequality again, $\rho(\boldsymbol{w}_\star, \boldsymbol{w}_{i_\star}) \leq 3\varepsilon$. ■

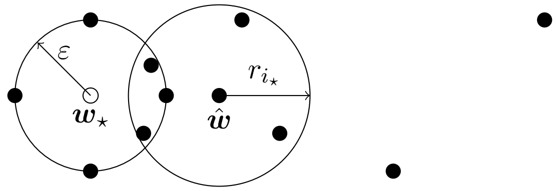

Figure 1: The argument in the proof of Proposition 8, illustrated on the Euclidean plane. If more than half of the $\boldsymbol{w}_i$ (depicted by full circles) are within ε of $\boldsymbol{w}_\star$ (the empty circle), then the selected $\boldsymbol{w}_{i_\star}$ is within $\varepsilon + r_{i_\star} \leq 3\varepsilon$ of $\boldsymbol{w}_\star$.

Algorithm 2 Robust approximation

input Number of candidates k, query access to $\text{APPROX}_{\rho,\varepsilon}$.
output Approximate solution $\hat{\boldsymbol{w}} \in \mathbb{X}$.

1: Query $\text{APPROX}_{\rho,\varepsilon}$ k times. Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$ be the responses to the queries; set $W := \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$.
2: For each $i \in [k]$, let $r_i := \min\{r \geq 0 : |B_\rho(\boldsymbol{w}_i, r) \cap W| > k/2\}$; set $i_\star := \arg\min_{i \in [k]} r_i$.
3: Return $\hat{\boldsymbol{w}} := \boldsymbol{w}_{i_\star}$.

Again, the number of candidates k determines the resulting confidence level. The following theorem provides a guarantee for Algorithm 2. We note that the resulting constants here might not be optimal in specific applications, since they depend on the arbitrary constant in Assumption 1.

Proposition 9 *Suppose that Assumption 1 and Assumption 2 hold. Then, with probability at least $1 - e^{-k/18}$, Algorithm 2 returns $\hat{\boldsymbol{w}} \in \mathbb{X}$ satisfying $\rho(\boldsymbol{w}_\star, \hat{\boldsymbol{w}}) \leq 3\varepsilon$.*

Proof For each $i \in [k]$, let $b_i := \mathbb{1}\{\rho(\boldsymbol{w}_\star, \boldsymbol{w}_i) \leq \varepsilon\}$. Note that the b_i are independent indicator random variables, each with $\mathbb{E}(b_i) \geq 2/3$. By Hoeffding's inequality, $\Pr[\sum_{i=1}^k b_i > k/2] \geq 1 - e^{-k/18}$. In the event that $\sum_{i=1}^k b_i > k/2$, more than half of the $\boldsymbol{w}_i$ are contained in the ball of radius ε around $\boldsymbol{w}_\star$, that is $\Delta_W(w_\star, 0) \leq \varepsilon$. The result follows from Proposition 8. ∎

3.3 Random Distance Measurements

In some problems, the most appropriate metric on $\mathbb{X}$ in which to measure accuracy is not directly computable. For instance, the metric may depend on population quantities which can only be estimated; moreover, the estimates may only be relatively accurate with some constant probability. For instance, this is the case when the metric depends on the population covariance matrix, a situation we consider in Section 5.2.3.

To capture such cases, we assume access to a metric estimation oracle as follows. Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$ be responses to k queries to $\text{APPROX}_{\rho,\epsilon}$. The metric estimation oracle, denoted DIST^j_ρ, provides (possibly via a random process) a function $f_j : \mathbb{X} \to \mathbb{R}_+$. $f_j(\boldsymbol{v})$ will be used as an estimate of $\rho(\boldsymbol{v}, \boldsymbol{w}_j)$. This estimate is required to be weakly accurate, as captured by the following definition of the random variables Z_j. Let $f_1, \ldots, f_k$ be responses to queries to $\mathrm{DIST}^1_\rho, \ldots, \mathrm{DIST}^k_\rho$, respectively. For $j \in [k]$, define

$$Z_j := \mathbb{1}\{\forall \boldsymbol{v} \in \mathbb{X}, \quad (1/2)\rho(\boldsymbol{v}, \boldsymbol{w}_j) \leq f_j(\boldsymbol{v}) \leq 2\rho(\boldsymbol{v}, \boldsymbol{w}_j)\}.$$

$Z_j = 1$ indicates that f_j provides a sufficiently accurate estimate of the distances from $\boldsymbol{w}_j$. Note that f_j need not correspond to a metric. We assume the following.

Assumption 3 *For any $j \in [k]$, $\Pr[Z_j = 1] \geq 8/9$.*

We further require the following independence assumption.

Assumption 4 *The random variables $Z_1, \ldots, Z_k$ are statistically independent.*

Note that there is no assumption on the statistical relationship between $Z_1, \ldots, Z_k$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$.

Algorithm 3 is a variant of Algorithm 2 that simply replaces computation of ρ-distances with computations using the functions returned by querying the DIST^j_ρ's. The resulting selection procedure is, with high probability, a Robust Distance Approximation.

Lemma 10 *Consider a run of Algorithm 3, with output $\hat{\boldsymbol{w}}$. Let $Z_1, \ldots, Z_k$ as defined above, and suppose that Assumption 3 and Assumption 4 hold. Then, with probability at least $1 - e^{-k/648}$,*

$$\rho(\hat{\boldsymbol{w}}, \boldsymbol{w}_\star) \leq 9 \cdot \Delta_W(\boldsymbol{w}_\star, \frac{5}{36}),$$

where $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$.

Proof By Assumptions 3 and 4, and by Hoeffding's inequality,

$$\Pr\left[\sum_{j=1}^{k} Z_j > \frac{31}{36}k\right] \geq 1 - e^{-k/648} \tag{3}$$

Assume this event holds, and denote $\varepsilon = \Delta_W(\boldsymbol{w}_\star, \frac{5}{36})$. We have $|B(\boldsymbol{w}_\star, \varepsilon) \cap W| \geq \frac{23}{36}k$.

Let $i \in [k]$ such that $\boldsymbol{w}_i \in B_\rho(\boldsymbol{w}_\star, \varepsilon)$. Then, for any $j \in [k]$ such that $\boldsymbol{w}_j \in B_\rho(\boldsymbol{w}_\star, \varepsilon)$, by the triangle inequality $\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \leq 2\varepsilon$. There are at least $\frac{23}{36}k$ such indices j, therefore for more than $k/2$ of the indices j, we have

$$\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \leq 2\varepsilon \text{ and } Z_j = 1.$$

For j such that this holds, by the definition of Z_j, $f_j(\boldsymbol{w}_i) \leq 4\varepsilon$. It follows that $r_i := \mathrm{median}\{f_j(\boldsymbol{w}_i) \mid j \in [k]\} \leq 4\epsilon$.

Now, let $i \in [k]$ such that $\boldsymbol{w}_i \notin B(\boldsymbol{w}_\star, 9\epsilon)$. Then, for any $j \in [k]$ such that $\boldsymbol{w}_j \in B_\rho(\boldsymbol{w}_\star, \varepsilon)$, by the triangle inequality $\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \geq \rho(\boldsymbol{w}_\star, \boldsymbol{w}_i) - \rho(\boldsymbol{w}_\star, \boldsymbol{w}_j) > 8\varepsilon$. As above, for more than $k/2$ of the indices j,

$$\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) > 8\varepsilon \text{ and } Z_j = 1.$$

For j such that this holds, by the definition of Z_j, $f_j(\boldsymbol{w}_i) > 4\varepsilon$. It follows that $r_i := \mathrm{median}\{f_j(\boldsymbol{w}_i) \mid j \in [k]\} > 4\epsilon$.

By Eq. (3), We conclude that with probability at least $1 - \exp(-k/648)$,

Algorithm 3 Robust approximation with random distances

input Number of candidates k, query access to $\text{APPROX}_{\rho,\varepsilon}$, query access to DIST_{ρ}.

output Approximate solution $\hat{\boldsymbol{w}} \in \mathbb{X}$.

1: Query $\text{APPROX}_{\rho,\varepsilon}$ k times. Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_k$ be the responses to the queries; set $W := \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$.

2: For $i \in [k]$, let f_i be the response of DIST_{ρ}^{j}, and set $r_i := \text{median}\{f_j(\boldsymbol{w}_i) : j \in [k]\}$; set $i_\star := \arg\min_{i \in [k]} r_i$.

3: Return $\hat{\boldsymbol{w}} := \boldsymbol{w}_{i_\star}$.

1. $r_i \leq 4\varepsilon$ for all $\boldsymbol{w}_i \in W \cap B_\rho(\boldsymbol{w}_\star, \varepsilon)$, and

2. $r_i > 4\varepsilon$ for all $\boldsymbol{w}_i \in W \setminus B_\rho(\boldsymbol{w}_\star, 9\varepsilon)$.

In this event the $\boldsymbol{w}_i \in W$ with the smallest r_i satisfies $\boldsymbol{w}_i \in B_\rho(\boldsymbol{w}_\star, 9\varepsilon)$. ■

The properties of the approximation procedure and of $\text{APPROX}_{\rho,\epsilon}$ are combined to give a guarantee for Algorithm 3.

Theorem 11 *Suppose that Assumptions 1,2,3,4 all hold. With probability at least* $1 - 2e^{-k/648}$, *Algorithm 3 returns* $\hat{\boldsymbol{w}} \in \mathbb{X}$ *satisfying* $\rho(\boldsymbol{w}_\star, \hat{\boldsymbol{w}}) \leq 9\varepsilon$.

Proof For each $i \in [k]$, let $b_i := \mathbb{1}\{\rho(\boldsymbol{w}_\star, \boldsymbol{w}_i) \leq \varepsilon\}$. By Assumptions 1 and 2, the b_i are independent indicator random variables, each with $\mathbb{E}(b_i) \geq 2/3$. By Hoeffding's inequality, $\Pr[\sum_{i=1}^{k} b_i > \frac{23}{36}k] \geq 1 - e^{-k/648}$. The result follows from Lemma 10 and a union bound. ■

In the following sections we show several applications of these general techniques.

4. Minimizing Strongly Convex Losses

In this section we apply the core techniques to the problem of approximately minimizing strongly convex losses, which includes least squares linear regression as a special case.

4.1 Preliminaries

Suppose $(\mathbb{X}, \|\cdot\|)$ is a Banach space, with the metric ρ induced by the norm $\|\cdot\|$. We sometimes denote the metric by $\|\cdot\|$ as well. Denote by $\|\cdot\|_*$ the dual norm, so $\|\boldsymbol{y}\|_* = \sup\{\langle \boldsymbol{y}, \boldsymbol{x}\rangle \colon \boldsymbol{x} \in \mathbb{X}, \|\boldsymbol{x}\| \leq 1\}$ for $\boldsymbol{y} \in \mathbb{X}^*$.

The derivative of a differentiable function $f \colon \mathbb{X} \to \mathbb{R}$ at $\boldsymbol{x} \in \mathbb{X}$ in direction $\boldsymbol{u} \in \mathbb{X}$ is denoted by $\langle \nabla f(\boldsymbol{x}), \boldsymbol{u}\rangle$. We say f is *α-strongly convex with respect to* $\|\cdot\|$ if

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \langle \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}$; it is *$\beta$-smooth with respect to* $\|\cdot\|$ if for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}$

$$f(\boldsymbol{x}) \leq f(\boldsymbol{x}') + \langle \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}'\rangle + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

We say $\|\cdot\|$ is *γ-smooth* if $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|^2$ is γ-smooth with respect to $\|\cdot\|$. We define n_α to be the smallest sample size such that the following holds: With probability $\geq 5/6$ over the choice of an i.i.d. sample T of size $|T| \geq n_\alpha$ from $\mathcal{D}$, for all $\boldsymbol{w} \in \mathbb{X}$,

$$L_T(\boldsymbol{w}) \geq L_T(\boldsymbol{w}_\star) + \langle \nabla L_T(\boldsymbol{w}_\star), \boldsymbol{w} - \boldsymbol{w}_\star \rangle + \frac{\alpha}{2}\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2. \tag{4}$$

In other words, the sample T induces a loss L_T which is α-strongly convex around $\boldsymbol{w}_\star$.[2] We assume that $n_\alpha < \infty$ for some $\alpha > 0$.

We use the following facts in our analysis.

Proposition 12 (Srebro et al., 2010) *If a non-negative function $f \colon \mathbb{X} \to \mathbb{R}_+$ is β-smooth with respect to $\|\cdot\|$, then $\|\nabla f(\boldsymbol{x})\|_*^2 \leq 4\beta f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{X}$.*

Proposition 13 (Juditsky and Nemirovski, 2008) *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent copies of a zero-mean random vector $\boldsymbol{X}$, and let $\|\cdot\|$ be γ-smooth. Then $\mathbb{E}\|n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\|^2 \leq (\gamma/n)\mathbb{E}\|\boldsymbol{X}\|^2$.*

Recall that $\mathcal{Z}$ is a data space, and $\mathcal{D}$ is a distribution over $\mathcal{Z}$. Let Z be a $\mathcal{Z}$-valued random variable with distribution $\mathcal{D}$. Let $\ell \colon \mathcal{Z} \times \mathbb{X} \to \mathbb{R}_+$ be a non-negative loss function, and for $\boldsymbol{w} \in \mathbb{X}$, let $L(\boldsymbol{w}) := \mathbb{E}(\ell(Z, \boldsymbol{w}))$ be the expected loss. Also define the empirical loss with respect to a sample T from $\mathcal{Z}$, $L_T(\boldsymbol{w}) := |T|^{-1}\sum_{z \in T} \ell(z, \boldsymbol{w})$. To simplify the discussion throughout, we assume ℓ is differentiable, which is anyway our primary case of interest. We assume that L has a unique minimizer $\boldsymbol{w}_\star := \arg\min_{\boldsymbol{w} \in \mathbb{X}} L(\boldsymbol{w})$.[3] Let $L_\star := \min_{\boldsymbol{w}} L(\boldsymbol{w})$. Set $\boldsymbol{w}_\star$ such that $L_\star = L(\boldsymbol{w}_\star)$.

4.2 Subsampled Empirical Loss Minimization

To use Algorithm 2, we implement $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ based on loss minimization over subsamples, as follows: Given a sample $S \subseteq \mathcal{Z}$, randomly partition S into k groups $S_1, S_2, \ldots, S_k$, each of size at least $\lfloor |S|/k \rfloor$, and let the response to the i-th query to $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ be the loss minimizer on S_i, *i.e.*, $\boldsymbol{w}_i = \arg\min_{\boldsymbol{w} \in \mathbb{X}} L_{S_i}(\boldsymbol{w})$. We call this implementation *subsampled empirical loss minimization*. Clearly, if S is an i.i.d. sample from $\mathcal{D}$, then $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ are statistically independent, and so Assumption 2 holds. Thus, to apply Proposition 9, it is left to show that Assumption 1 holds as well.[4]

The following lemma proves that Assumption 1 holds under these assumptions with

$$\varepsilon := \sqrt{\frac{32\gamma k \mathbb{E}\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2}{n\alpha^2}}. \tag{5}$$

Lemma 14 *Let ε be as defined in Eq. (5). Assume $k \leq n/4$, and that S is an i.i.d. sample from $\mathcal{D}$ of size n such that $\lfloor n/k \rfloor \geq n_\alpha$. Then subsampled empirical loss minimization using the sample S is a correct implementation of* $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ *for up to k queries.*

2. Technically, we only need the sample size to guarantee Eq. (4) for all $\boldsymbol{w} \in B_{\|\cdot\|}(\boldsymbol{w}_\star, r)$ for some $r > 0$.
3. This holds, for instance, if L is strongly convex.
4. An approach akin to the bootstrap technique (Efron, 1979) could also seem natural here: In this approach, $S_1, \ldots, S_k$ would be generated by randomly sub-sampling from S, with possible overlap between the sub-samples. However, this approach does not satisfy Assumption 2, since loss minimizers of overlapping samples are not statistically independent.

Proof Let $T = \lfloor n/k \rfloor$. Since $n \geq 4k$, we have $\lfloor n/k \rfloor k \geq n - k \geq \frac{3}{4}n$, therefore $1/T \leq \frac{4k}{3n}$. It is clear that $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k$ are independent by the assumption. Fix some $i \in [k]$. Observe that $\nabla L(\boldsymbol{w}_\star) = \mathbb{E}(\nabla \ell(Z, \boldsymbol{w}_\star)) = 0$, and therefore by Proposition 13:

$$\mathbb{E}\|\nabla L_{S_i}(\boldsymbol{w}_\star)\|_*^2 \leq (\gamma/T)\mathbb{E}\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2 \leq \frac{4\gamma k}{3n}\mathbb{E}\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2.$$

By Markov's inequality,

$$\Pr\left[\|\nabla L_{S_i}(\boldsymbol{w}_\star)\|_*^2 \leq \frac{8\gamma k}{n}\mathbb{E}(\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2)\right] \geq \frac{5}{6}.$$

Moreover, the assumption that $\lfloor n/k \rfloor \geq n_\alpha$ implies that with probability at least $5/6$, Eq. (4) holds for $T = S_i$. By a union bound, both of these events hold simultaneously with probability at least $2/3$. In the intersection of these events, letting $\boldsymbol{w}_i := \arg\min_{\boldsymbol{w} \in \mathbb{X}} L_{S_i}(\boldsymbol{w})$,

$$\begin{aligned}(\alpha/2)\|\boldsymbol{w}_i - \boldsymbol{w}_\star\|^2 &\leq -\langle \nabla L_{S_i}(\boldsymbol{w}_\star), \boldsymbol{w}_i - \boldsymbol{w}_\star \rangle + L_{S_i}(\boldsymbol{w}_i) - L_{S_i}(\boldsymbol{w}_\star) \\ &\leq \|\nabla L_{S_i}(\boldsymbol{w}_\star)\|_* \|\boldsymbol{w}_i - \boldsymbol{w}_\star\|,\end{aligned}$$

where the last inequality follows from the definition of the dual norm, and the optimality of $\boldsymbol{w}_i$ on L_{S_i}. Rearranging and combining with the above probability inequality implies

$$\Pr\left[\|\boldsymbol{w}_i - \boldsymbol{w}_\star\| \leq \varepsilon\right] \geq \frac{2}{3}$$

as required. ■

Combining Lemma 14 and Proposition 9 gives the following theorem.

Theorem 15 *Let n_α be as defined in Section 4.1, and assume that $\|\cdot\|_*$ is γ-smooth. Also, assume $k := 18\lceil \log(1/\delta) \rceil$, $n \geq 72\lceil \log(1/\delta) \rceil$, and that S is an i.i.d. sample from $\mathcal{D}$ of size n such that $\lfloor n/k \rfloor \geq n_\alpha$. Finally, assume Algorithm 3 uses the subsampled empirical loss minimization to implement $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$, where ε is as in Eq. (5). Then with probability at least $1 - \delta$, the parameter $\hat{\boldsymbol{w}}$ returned by Algorithm 2 satisfies*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\| \leq 72\sqrt{\frac{\gamma\lceil \log(1/\delta) \rceil \mathbb{E}\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2}{n\alpha^2}}.$$

We give an easy corollary of Theorem 15 for the case where ℓ is smooth. This is the full version of Theorem 2.

Corollary 16 *Assume the same conditions as Theorem 15, and also that:*

- *$\boldsymbol{w} \mapsto \ell(z, \boldsymbol{w})$ is β-smooth with respect to $\|\cdot\|$ for all $z \in \mathcal{Z}$;*
- *$\boldsymbol{w} \mapsto L(\boldsymbol{w})$ is $\bar{\beta}$-smooth with respect to $\|\cdot\|$.*

Then with probability at least $1 - \delta$,

$$L(\hat{\boldsymbol{w}}) \leq \left(1 + \frac{10368\beta\bar{\beta}\gamma\lceil \log(1/\delta) \rceil}{n\alpha^2}\right) L(\boldsymbol{w}_\star).$$

Proof This follows from Theorem 15 by first concluding that $\mathbb{E}[\|\nabla\ell(Z,\boldsymbol{w}_\star)\|_*^2] \leq 4\beta L(\boldsymbol{w}_\star)$, using the β-strong smoothness assumption on ℓ and Proposition 12, and then noting that $L(\hat{\boldsymbol{w}}) - L(\boldsymbol{w}_\star) \leq \frac{\bar{\beta}}{2}\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|^2$, due to the strong smoothness of L and the optimality of $L(\boldsymbol{w}_\star)$. ∎

Corollary 16 implies that for smooth losses, Algorithm 2 provides a constant factor approximation to the optimal loss with a sample size $\max\{n_\alpha, \gamma\beta\bar{\beta}/\alpha^2\} \cdot O(\log(1/\delta))$ (with probability at least $1-\delta$). In subsequent sections, we exemplify cases where the two arguments of the max are roughly of the same order, and thus imply a sample size requirement of $O(\gamma\bar{\beta}\beta/\alpha^2 \log(1/\delta))$. Note that there is no dependence on the optimal loss $L(\boldsymbol{w}_\star)$ in the sample size, and the algorithm has no parameters besides $k = O(\log(1/\delta))$.

We can also obtain a variant of Theorem 15 based on Algorithm 3 and Theorem 11, in which we assume that there exists some sample size $n_{k,\mathrm{DIST}_{\|\cdot\|}}$ that allows $\mathrm{DIST}_{\|\cdot\|}$ to be correctly implemented using an i.i.d. sample of size at least $n_{k,\mathrm{DIST}_{\|\cdot\|}}$. Under such an assumption, essentially the same guarantee as in Theorem 15 can be afforded to Algorithm 3 using the subsampled empirical loss minimization to implement $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ (for ε as in Eq. (5)) and the assumed implementation of $\mathrm{DIST}_{\|\cdot\|}$. Note that since Theorem 11 does not require $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ and $\mathrm{DIST}_{\|\cdot\|}$ to be statistically independent, both can be implemented using the same sample.

Theorem 17 *Let n_α be as defined in Section 4.1, $n_{k,\mathrm{DIST}_{\|\cdot\|}}$ be as defined above, and assume that $\|\cdot\|_*$ is γ-smooth. Also, assume $k := 648\lceil\log(2/\delta)\rceil$, S is an i.i.d. sample from $\mathcal{D}$ of size n such that $n \geq \max\{4k, n_{k,\mathrm{DIST}_{\|\cdot\|}}\}$, and $\lfloor n/k\rfloor \geq n_\alpha$. Further, assume Algorithm 3 implements $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ using S with subsampled empirical loss minimization, where ε is as in Eq. (5), and implements $\mathrm{DIST}_{\|\cdot\|}$ using S as well. Then with probability at least $1-\delta$, the parameter $\hat{\boldsymbol{w}}$ returned by Algorithm 3 satisfies*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\| \leq 1296\sqrt{\frac{\gamma\lceil\log(2/\delta)\rceil\mathbb{E}\|\nabla\ell(Z,\boldsymbol{w}_\star)\|_*^2}{n\alpha^2}}.$$

Remark 18 (Mean estimation and empirical risk minimization) *The problem of estimating a scalar population mean is a special case of the loss minimization problem, where $\mathcal{Z} = \mathbb{X} = \mathbb{R}$, and the loss function of interest is the square loss $\ell(z,w) = (z-w)^2$. The minimum population loss in this setting is the variance σ^2 of Z,* i.e., *$L(w_\star) = \sigma^2$. Moreover, in this setting, we have $\alpha = \beta = \bar{\beta} = 2$, so the estimate $\hat{w}$ returned by Algorithm 2 satisfies, with probability at least $1-\delta$,*

$$L(\hat{w}) = \left(1 + O\Big(\frac{\log(1/\delta)}{n}\Big)\right)L(w_\star).$$

*In Remark 7 a result from Catoni (2012) is quoted which implies that if $n = o(1/\delta)$, then the empirical mean $\hat{w}_{\mathrm{emp}} := \arg\min_{w\in\mathbb{R}} L_S(w) = |S|^{-1}\sum_{z\in S} z$ (*i.e., *empirical risk (loss) minimization for this problem) incurs loss*

$$L(\hat{w}_{\mathrm{emp}}) = \sigma^2 + (\hat{w}_{\mathrm{emp}} - w_\star)^2 = (1+\omega(1))L(w_\star)$$

with probability at least 2δ. Therefore empirical risk minimization cannot provide a qualitatively similar guarantee as Corollary 16. It is easy to check that minimizing a regularized *objective also does not work, since any non-trivial regularized objective necessarily provides an estimator with a positive error for some distribution with zero variance.*

In the next section we use the analysis for general smooth and convex losses to derive new algorithms and bounds for linear regression.

5. Least Squares Linear Regression

In linear regression, the parameter space $\mathbb{X}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{X}}$, and $\mathcal{Z} := \mathbb{X} \times \mathbb{R}$, where in the finite-dimensional case, $\mathbb{X} = \mathbb{R}^d$ for some finite integer d. The loss here is the squared loss, denoted by $\ell = \ell^{\mathrm{sq}}$, and defined as

$$\ell^{\mathrm{sq}}((\boldsymbol{x}, y), \boldsymbol{w}) := \frac{1}{2}(\boldsymbol{x}^\top \boldsymbol{w} - y)^2.$$

The regularized squared loss, for $\lambda \geq 0$, is denoted

$$\ell^{\lambda}((\boldsymbol{x}, y), \boldsymbol{w}) := \frac{1}{2}(\langle \boldsymbol{x}, \boldsymbol{w} \rangle_{\mathbb{X}} - y)^2 + \frac{1}{2}\lambda \langle \boldsymbol{w}, \boldsymbol{w} \rangle_{\mathbb{X}}.$$

Note that $\ell^0 = \ell^{\mathrm{sq}}$. We analogously define L^{sq}, L_T^{sq}, $L_\star^{\mathrm{sq}}$, L^λ, *etc.* as the squared-loss equivalents of $L, L_T, L_\star$. Finally, denote by Id the identity operator on $\mathbb{X}$.

The proposed algorithm for regression (Algorithm 4) is as follows. Set $k = C \log(1/\delta)$, where C is a universal constant. First, draw k independent random samples i.i.d. from $\mathcal{D}$, and perform linear regression with λ-regularization on each sample separately to obtain k linear regressors. Then, use the same k samples to generate k estimates of the covariance matrix of the marginal of $\mathcal{D}$ on the data space. Finally, use the estimated covariances to select a single regressor from among the k at hand. The slightly simpler variants of steps 4 and 5 can be used in some cases, as detailed below.

In Section 5.1, the full results for regression, mentioned in Section 2, are listed in full detail, and compared to previous work. The proofs are provided in Section 5.2.

5.1 Results

Let $\boldsymbol{X} \in \mathbb{X}$ be a random vector drawn according to the marginal of $\mathcal{D}$ on $\mathbb{X}$, and let $\Sigma : \mathbb{X} \to \mathbb{X}$ be the second-moment operator $\boldsymbol{a} \mapsto \mathbb{E}(\boldsymbol{X} \langle \boldsymbol{X}, \boldsymbol{a} \rangle_{\mathbb{X}})$. For a finite-dimensional $\mathbb{X}$, Σ is simply the (uncentered) covariance matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$. For a sample $T := \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m\}$ of m independent copies of $\boldsymbol{X}$, denote by $\Sigma_T : \mathbb{X} \to \mathbb{X}$ the empirical second-moment operator $\boldsymbol{a} \mapsto m^{-1} \sum_{i=1}^m \boldsymbol{X}_i \langle \boldsymbol{X}_i, \boldsymbol{a} \rangle_{\mathbb{X}}$.

Consider first the finite-dimensional case, where $\mathbb{X} = \mathbb{R}^d$, and assume Σ is not singular. Let $\|\cdot\|_2$ denote the Euclidean norm in $\mathbb{R}^d$. In this case we obtain a guarantee for ordinary least squares with $\lambda = 0$. The guarantee holds whenever the empirical estimate of Σ is close to the true Σ *in expectation*, a mild condition that requires only bounded low-order moments. For concreteness, we assume the following condition.[5]

5. As shown by Srivastava and Vershynin (2013), Condition 1 holds for various heavy-tailed distributions (*e.g.*, when $\boldsymbol{X}$ has a product distribution with bounded $4+\epsilon$ moments for some $\epsilon > 0$). Condition 1 may be easily substituted with other moment conditions, yielding similar results, at least up to logarithmic factors.

Algorithm 4 Regression for heavy-tails

input $\lambda \geq 0$, sample size n, confidence $\delta \in (0,1)$.
output Approximate predictor $\hat{\boldsymbol{w}} \in \mathbb{X}$.

1: Set $k := \lceil C \ln(1/\delta) \rceil$.
2: Draw k random i.i.d. samples $S_1, \dots, S_k$ from D, each of size $\lfloor n/k \rfloor$.
3: For each $i \in [k]$, let $\boldsymbol{w}_i \in \operatorname{argmin}_{\boldsymbol{w} \in \mathbb{X}} L^{\lambda}_{S_i}(\boldsymbol{w})$.
4: For each $i \in [k]$, $\Sigma_{S_i} \leftarrow \frac{1}{|S_i|} \sum_{(\boldsymbol{x},\cdot) \in S_i} \boldsymbol{x}\boldsymbol{x}^\top$.
[**Variant**: $S \leftarrow \cup_{i \in [k]} S_i$; $\Sigma_S \leftarrow \frac{1}{|S|} \sum_{(\boldsymbol{x},\cdot) \in S} \boldsymbol{x}\boldsymbol{x}^\top$].
5: For each $i \in [k]$, let r_i be the median of the values in

$$\{\langle \boldsymbol{w}_i - \boldsymbol{w}_j, (\Sigma_{S_j} + \lambda \operatorname{Id})(\boldsymbol{w}_i - \boldsymbol{w}_j) \rangle \mid j \in [k] \setminus \{i\}\}.$$

[**Variant**: Use Σ_S instead of Σ_{S_j}].
6: Set $i_\star := \arg\min_{i \in [k]} r_i$.
7: Return $\hat{\boldsymbol{w}} := \boldsymbol{w}_{i_\star}$.

Condition 1 (Srivastava and Vershynin 2013) *There exists $c, \eta > 0$ such that*

$$\Pr\left[\|\Pi \Sigma^{-1/2} \boldsymbol{X}\|_2^2 > t\right] \;\leq\; c t^{-1-\eta}, \quad \textit{for } t > c \cdot \operatorname{rank}(\Pi)$$

for every orthogonal projection Π in $\mathbb{R}^d$.

Under this condition, we show the following guarantee for least squares regression.

Theorem 19 *Assume Σ is not singular. If $\boldsymbol{X}$ satisfies Condition 1 with some fixed parameters $c > 0$ and $\eta > 0$, then if Algorithm 4 is run with $n \geq O(d \log(1/\delta))$ and $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq L^{\mathrm{sq}}_\star + O\left(\frac{\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)\|_2^2 \log(1/\delta)}{n}\right).$$

Our loss bound is given in terms of the following population quantity

$$\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)\|_2^2 \tag{6}$$

which we assume is finite. This assumption only requires bounded low-order moments of $\boldsymbol{X}$ and Y and is essentially the same as the conditions from Audibert and Catoni (2011) (see the discussion following their Theorem 3.1). Define the following finite fourth-moment conditions:

$$\kappa_1 := \frac{\sqrt{\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}\|_2^4}}{\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}\|_2^2} = \frac{\sqrt{\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}\|_2^4}}{d} < \infty \quad \text{and}$$
$$\kappa_2 := \frac{\sqrt{\mathbb{E}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)^4}}{\mathbb{E}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)^2} = \frac{\sqrt{\mathbb{E}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)^4}}{L^{\mathrm{sq}}_\star} < \infty.$$

Under these conditions, $\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)\|_2^2 \leq \kappa_1\kappa_2 d L_\star^{\mathrm{sq}}$ (via Cauchy-Schwartz); if κ_1 and κ_2 are constant, then we obtain the bound

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{d\log(1/\delta)}{n}\right)\right) L_\star^{\mathrm{sq}}$$

with probability $\geq 1 - \delta$. In comparison, the recent work of Audibert and Catoni (2011) proposes an estimator for linear regression based on optimization of a robust loss function which achieves essentially the same guarantee as Theorem 19 (with only mild differences in the moment conditions, see the discussion following their Theorem 3.1). However, that estimator depends on prior knowledge about the response distribution, and removing this dependency using Lepski's adaptation method (Lepski, 1991) may result in a suboptimal convergence rate. It is also unclear whether that estimator can be computed efficiently.

Other analyses for linear least squares regression and ridge regression by Srebro et al. (2010) and Hsu et al. (2014) consider specifically the empirical minimizer of the squared loss, and give sharp rates of convergence to $L_\star^{\mathrm{sq}}$. However, both of these require either boundedness of the loss or boundedness of the approximation error. In Srebro et al. (2010), the specialization of the main result to square loss includes additive terms of order $O(\sqrt{L(\boldsymbol{w}_\star) b \log(1/\delta)/n} + b\log(1/\delta)/n)$, where $b > 0$ is assumed to bound the square loss of any predictions almost surely. In Hsu et al. (2014), the convergence rate includes an additive term involving almost-sure bounds on the approximation error/non-subgaussian noise (The remaining terms are comparable to Eq. (9) for $\lambda = 0$, and Eq. (7) for $\lambda > 0$, up to logarithmic factors). The additional terms preclude multiplicative approximations to $L(\boldsymbol{w}_\star)$ in cases where the loss or approximation error is unbounded. In recent work, Mendelson (2014) proposes a more subtle 'small-ball' criterion for analyzing the performance of the risk minimizer. However, as evident from the lower bound in Remark 18, the empirical risk minimizer cannot obtain the same type of guarantees as our estimator.

The next result is for the case where there exists $R < \infty$ such that $\Pr[\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X} \leq R^2] = 1$ (and, here, we do not assume Condition 1). In contrast, Y may still be heavy-tailed. Then, the following result can be derived using Algorithm 4. Moreover, the simpler **variant** of Algorithm 4 suffices here.

Theorem 20 *Assume Σ is not singular. Let $\hat{\boldsymbol{w}}$ be the output of the variant of Algorithm 4 with $\lambda = 0$. With probability at least $1 - \delta$, for $n \geq O(R^2 \log(R)\log(1/\delta))$,*

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{R^2\log(2/\delta)}{n}\right)\right) L_\star^{\mathrm{sq}}.$$

Note that $\mathbb{E}(\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X}) = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X}) = \mathrm{tr}(\mathrm{Id}) = d$, therefore $R = \Omega(\sqrt{d})$. If indeed $R = \Theta(\sqrt{d})$, then a total sample size of $O(d\log(d)\log(1/\delta))$ suffices to guarantee a constant factor approximation to the optimal loss. This is minimax optimal up to logarithmic factors (Nussbaum, 1999). We also remark that the boundedness assumption can be replaced by a subgaussian assumption on $\boldsymbol{X}$, in which case the sample size requirement becomes $O(d\log(1/\delta))$.

In recent work of Mahdavi and Jin (2013), an algorithm based on stochastic gradient descent obtains multiplicative approximations to $L_\star$, for general smooth and strongly convex losses ℓ, with a sample complexity scaling with $\log(1/\tilde{L})$. Here, $\tilde{L}$ is an upper bound on $L_\star$, which must be known by the algorithm. The specialization of Mahdavi and Jin's main result to square loss implies a sample complexity of $\tilde{O}(dR^8 \log(1/(\delta L_\star^{\mathrm{sq}}))$ if $L_\star^{\mathrm{sq}}$ is known. In comparison, Theorem 20 shows that $\tilde{O}(R^2 \log(1/\delta))$ suffice when using our estimator. It would be interesting to understand whether the bound for the stochastic gradient method of Mahdavi and Jin (2013) can be improved, and whether knowledge of $L_\star$ is actually necessary in the stochastic oracle model. We note that the main result of Mahdavi and Jin (2013) can be more generally applicable than Theorem 15, because Mahdavi and Jin (2013) only assumes that the population loss $L(\boldsymbol{w})$ is strongly convex, whereas Theorem 15 requires the empirical loss $L_T(\boldsymbol{w})$ to be strongly convex for large enough samples T. While our technique is especially simple for the squared loss, it may be more challenging to implement well for other losses, because the local norm around $\boldsymbol{w}_\star$ may be difficult to approximate with an observable norm. We thus leave the extension to more general losses as future work.

Finally, we also consider the case where $\mathbb{X}$ is a general, infinite-dimensional Hilbert space, $\lambda > 0$, the norm of $\boldsymbol{X}$ is bounded, and Y again may be heavy-tailed.

Theorem 21 *Let $V > 0$ such that $\Pr[\langle \boldsymbol{X}, \boldsymbol{X}\rangle_{\mathbb{X}} \leq V^2] = 1$. Let $\hat{\boldsymbol{w}}$ be the output of the variant of Algorithm 4 with $\lambda > 0$. With probability at least $1-\delta$, as soon as $n \geq O((V^2/\lambda)\log(V/\sqrt{\lambda})\log(2/\delta))$,*

$$L^\lambda(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{(1+V^2/\lambda)\log(2/\delta)}{n}\right)\right) L_\star^\lambda.$$

If the optimal unregularized squared loss $L_\star^{\mathrm{sq}}$ is achieved by $\bar{\boldsymbol{w}} \in \mathbb{X}$ with $\langle \bar{\boldsymbol{w}}, \bar{\boldsymbol{w}}\rangle_{\mathbb{X}} \leq B^2$, the choice $\lambda = \Theta(\sqrt{L_\star^{\mathrm{sq}} V^2 \log(2/\delta)/(B^2 n)})$ yields that if $n \geq \tilde{O}(B^2V^2\log(2/\delta)/L_\star^{\mathrm{sq}})$ then

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) \leq L_\star^{\mathrm{sq}} + O\left(\sqrt{\frac{L_\star^{\mathrm{sq}} B^2 V^2 \log(1/\delta)}{n}} + \frac{(L_\star^{\mathrm{sq}} + B^2V^2)\log(1/\delta)}{n}\right). \tag{7}$$

By this analysis, a constant factor approximation for $L_\star^{\mathrm{sq}}$ is achieved with a sample of size $\tilde{O}(B^2V^2\log(1/\delta)/L_\star^{\mathrm{sq}})$. As in the finite-dimensional setting, this rate is known to be optimal up to logarithmic factors (Nussbaum, 1999). It is interesting to observe that in the non-parametric case, our analysis, like previous analyses, does require knowledge of $L_\star$ if λ is to be set correctly, as in Mahdavi and Jin (2013).

5.2 Analysis

We now show how the analysis of Section 4 can be applied to analyze Algorithm 4. For a sample $T \subseteq \mathcal{Z}$, if L_T is twice-differentiable (which is the case for squared loss), by Taylor's theorem, for any $\boldsymbol{w} \in \mathbb{X}$, there exist $t \in [0,1]$ and $\tilde{\boldsymbol{w}} = t\boldsymbol{w}_\star + (1-t)\boldsymbol{w}$ such that

$$L_T(\boldsymbol{w}) = L_T(\boldsymbol{w}_\star) + \langle \nabla L_T(\boldsymbol{w}_\star), \boldsymbol{w} - \boldsymbol{w}_\star\rangle_{\mathbb{X}} + \frac{1}{2}\langle \boldsymbol{w} - \boldsymbol{w}_\star, \nabla^2 L_T(\tilde{\boldsymbol{w}})(\boldsymbol{w} - \boldsymbol{w}_\star)\rangle_{\mathbb{X}},$$

Therefore, to establish a bound on n_α, it suffices to control

$$\Pr\left[\inf_{\boldsymbol{\delta} \in \mathbb{X}\setminus\{\boldsymbol{0}\}, \tilde{\boldsymbol{w}} \in \mathbb{R}^d} \frac{\langle \boldsymbol{\delta}, \nabla^2 L_T(\tilde{\boldsymbol{w}})\boldsymbol{\delta}\rangle_{\mathbb{X}}}{\|\boldsymbol{\delta}\|^2} \geq \alpha\right] \tag{8}$$

for an i.i.d. sample T from $\mathcal{D}$. The following lemma allows doing just that.

Lemma 22 (Specialization of Lemma 1 from Oliveira 2010) *Fix any* $\lambda \geq 0$, *and assume* $\langle \boldsymbol{X}, (\Sigma + \lambda \operatorname{Id})^{-1} \boldsymbol{X} \rangle_{\mathbb{X}} \leq r_\lambda^2$ *almost surely. For any* $\delta \in (0,1)$, *if* $m \geq 80 r_\lambda^2 \ln(4m^2/\delta)$, *then with probability at least* $1-\delta$, *for all* $\boldsymbol{a} \in \mathbb{X}$,

$$\frac{1}{2}\langle \boldsymbol{a}, (\Sigma + \lambda \operatorname{Id})\boldsymbol{a}\rangle_{\mathbb{X}} \leq \langle \boldsymbol{a}, (\Sigma_T + \lambda \operatorname{Id})\boldsymbol{a}\rangle_{\mathbb{X}} \leq 2\langle \boldsymbol{a}, (\Sigma + \lambda \operatorname{Id})\boldsymbol{a}\rangle_{\mathbb{X}}.$$

We use the boundedness assumption for sake of simplicity; it is possible to remove the boundedness assumption, and the logarithmic dependence on the cardinality of T, under different conditions on $\boldsymbol{X}$ (*e.g.*, assuming $\Sigma^{-1/2}\boldsymbol{X}$ has subgaussian projections, see Litvak et al. 2005). We now prove Theorem 20, Theorem 21 and Theorem 19.

5.2.1 Ordinary Least Squares in Finite Dimensions

Consider first ordinary least squares in the finite-dimensional case. In this case $\mathbb{X} = \mathbb{R}^d$, the inner product $\langle \boldsymbol{a}, \boldsymbol{b}\rangle_{\mathbb{X}} = \boldsymbol{a}^\top \boldsymbol{b}$ is the usual coordinate dot product, and the second-moment operator is $\Sigma = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)$. We assume that Σ is non-singular, so L has a unique minimizer. Here Algorithm 4 can be used with $\lambda = 0$. It is easy to see that Algorithm 4 with the **variant** steps is a specialization of Algorithm 2 with subsampled empirical loss minimization when $\ell = \ell^{\text{sq}}$, with the norm defined by $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\top \Sigma_S \boldsymbol{a}}$. We now prove the guarantee for finite dimensional regression.

Proof [of Theorem 20] The proof is derived from Corollary 16 as follows. First, suppose for simplicity that $\Sigma_s = \Sigma$, so that $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\top \Sigma \boldsymbol{a}}$. It is easy to check that $\|\cdot\|_*$ is 1-smooth, ℓ is R^2-smooth with respect to $\|\cdot\|$, and L^{sq} is 1-smooth with respect to $\|\cdot\|$. Moreover, consider a random sample T. By definition

$$\frac{\boldsymbol{\delta}^\top \nabla^2 L_T(\tilde{\boldsymbol{w}})\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|^2} = \frac{\boldsymbol{\delta}^\top \Sigma_T \boldsymbol{\delta}}{\boldsymbol{\delta}^\top \Sigma \boldsymbol{\delta}}.$$

By Lemma 22 with $\lambda = 0$, $\Pr[\inf\{\boldsymbol{\delta}^\top \Sigma_T \boldsymbol{\delta}/(\boldsymbol{\delta}^\top \Sigma \boldsymbol{\delta}) : \boldsymbol{\delta} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}\} \geq 1/2] \geq 5/6$, provided that $|T| \geq 80R^2 \log(24|S|^2)$. Therefore $n_{0.5} = O(R^2 \log R)$. We can thus apply Corollary 16 with $\alpha = 0.5$, $\beta = R^2$, $\bar{\beta} = 1$, $\gamma = 1$, and $n_{0.5} = O(R^2 \log R)$, so with probability at least $1-\delta$, the parameter $\hat{\boldsymbol{w}}$ returned by Algorithm 4 satisfies

$$L(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{R^2 \log(1/\delta)}{n}\right)\right) L(\boldsymbol{w}_\star), \tag{9}$$

as soon as $n \geq O(R^2 \log(R)\log(1/\delta))$.

Now, by Lemma 22, if $n \geq O(R^2 \log(R/\delta))$, with probability at least $1-\delta$, the norm induced by Σ_S satisfies $(1/2)\boldsymbol{a}^\top \Sigma \boldsymbol{a} \leq \boldsymbol{a}^\top \Sigma_S \boldsymbol{a} \leq 2\boldsymbol{a}^\top \Sigma \boldsymbol{a}$ for all $\boldsymbol{a} \in \mathbb{R}^d$. Therefore, by a union bound, the norm used by the algorithm is equivalent to the norm induced by the true Σ up to constant factors, and thus leads to the same guarantee as given above (where the constant factors are absorbed into the big-O notation). ■

The rate achieved in Eq. (9) is well-known to be optimal up to logarithmic factors (Nussbaum, 1999). A standard argument for this, which we reference in the sequel, is as follows. Consider a distribution over $\mathbb{R}^d \times \mathbb{R}$ where $\boldsymbol{X} \in \mathbb{R}^d$ is distributed uniformly over some orthonormal basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_d$, and $Y := \boldsymbol{X}^\top \boldsymbol{w}_\star + Z$ for $Z \sim \mathcal{N}(0, \sigma^2)$ independent of $\boldsymbol{X}$. Here, $\boldsymbol{w}_\star$ is an arbitrary vector in $\mathbb{R}^d$, $R = \sqrt{d}$, and the optimal square loss is $L(\boldsymbol{w}_\star) = \sigma^2$. Among n independent copies of $(\boldsymbol{X}, Y)$, let n_i be the number of copies with $\boldsymbol{X} = \boldsymbol{e}_i$, so $\sum_{i=1}^d n_i = n$. Estimating $\boldsymbol{w}_\star$ is equivalent to d Gaussian mean estimation problems, with a minimax loss of

$$\begin{aligned}\inf_{\hat{\boldsymbol{w}}} \sup_{\boldsymbol{w}_\star} \mathbb{E}\big(L(\hat{\boldsymbol{w}})\big) - L(\boldsymbol{w}_\star) &= \inf_{\hat{\boldsymbol{w}}} \sup_{\boldsymbol{w}_\star} \mathbb{E}\left(\frac{1}{d}\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_2^2\right) \\ &= \frac{1}{d}\sum_{i=1}^d \frac{\sigma^2}{n_i} \geq \frac{d\sigma^2}{n} = \frac{dL(\boldsymbol{w}_\star)}{n}. \end{aligned} \tag{10}$$

Note that this also implies a lower bound for any estimator with exponential concentration. That is, for any estimator $\hat{\boldsymbol{w}}$, if there is some $A > 0$ such that for any $\delta \in (0,1)$, $\mathbb{P}[L(\hat{\boldsymbol{w}}) > L(\boldsymbol{w}_\star) + A\log(1/\delta)] < \delta$, then $A \geq \mathbb{E}(L(\hat{\boldsymbol{w}}) - L(\boldsymbol{w}_\star)) \geq dL(\boldsymbol{w}_\star)/n$.

5.2.2 Ridge Regression

In a general, possibly infinite-dimensional, Hilbert space $\mathbb{X}$, Algorithm 4 can be used with $\lambda > 0$. In this case, Algorithm 4 with the **variant** steps is again a specialization of Algorithm 2 with subsampled empirical loss minimization when $\ell = \ell^\lambda$, with the norm defined by $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\top(\Sigma_S + \lambda\,\mathrm{Id})\boldsymbol{a}}$.

Proof [of Theorem 21] As in the finite-dimensional case, assume first that $\Sigma_S = \Sigma$, and consider the norm $\|\cdot\|$ defined by $\|\boldsymbol{a}\| := \sqrt{\langle \boldsymbol{a}, (\Sigma + \lambda\,\mathrm{Id})\boldsymbol{a}\rangle_{\mathbb{X}}}$. It is easy to check that $\|\cdot\|_*$ is 1-smooth. Moreover, since we assume that $\Pr[\langle \boldsymbol{X}, \boldsymbol{X}\rangle_{\mathbb{X}} \leq V^2] = 1$, we have $\langle \boldsymbol{x}, (\Sigma + \lambda I)^{-1}\boldsymbol{x}\rangle_{\mathbb{X}} \leq \langle \boldsymbol{x}, \boldsymbol{x}\rangle_{\mathbb{X}}/\lambda$ for all $\boldsymbol{x} \in \mathbb{X}$, so $\Pr[\langle \boldsymbol{X}, (\Sigma + \lambda I)^{-1}\boldsymbol{X}\rangle_{\mathbb{X}} \leq V^2/\lambda] = 1$. Therefore ℓ^λ is $(1 + V^2/\lambda)$-smooth with respect to $\|\cdot\|$. In addition, L^λ is 1-smooth with respect to $\|\cdot\|$. Using Lemma 22 with $r_\lambda = V/\lambda$, we have, similarly to the proof of Theorem 20, $n_{0.5} = O((V^2/\lambda)\log(V/\sqrt{\lambda}))$. Setting $\alpha = 0.5$, $\beta = 1 + V^2/\lambda$, $\bar{\beta} = 1$, $\gamma = 1$, and $n_{0.5}$ as above, we conclude that with probability $1 - \delta$,

$$L^\lambda(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{(1 + V^2/\lambda)\log(1/\delta)}{n}\right)\right) L^\lambda(\boldsymbol{w}_\star),$$

as soon as $n \geq O((V^2/\lambda)\log(V/\sqrt{\lambda})\log(1/\delta))$. Again as in the proof of Theorem 20, by Lemma 22 Algorithm 4 may use the observable norm $\boldsymbol{a} \mapsto \langle \boldsymbol{a}, (\Sigma_S + \lambda I)\boldsymbol{a}\rangle_{\mathbb{X}}^{1/2}$ instead of the unobservable norm $\boldsymbol{a} \mapsto \langle \boldsymbol{a}, (\Sigma + \lambda I)\boldsymbol{a}\rangle_{\mathbb{X}}^{1/2}$ by applying a union bound, if $n \geq O((V^2/\lambda)\log(2V/(\delta\sqrt{\lambda})))$, losing only constant factors, .

We are generally interested in comparing to the minimum square loss $L_\star^{\mathrm{sq}} := \inf_{\boldsymbol{w} \in \mathbb{X}} L^{\mathrm{sq}}(\boldsymbol{w})$, rather than the minimum regularized square loss $\inf_{\boldsymbol{w} \in \mathbb{X}} L^\lambda(\boldsymbol{w})$. Assuming the minimizer is achieved by some $\bar{\boldsymbol{w}} \in \mathbb{X}$ with $\langle \bar{\boldsymbol{w}}, \bar{\boldsymbol{w}}\rangle_{\mathbb{X}} \leq B^2$, the choice $\lambda = \Theta(\sqrt{L_\star^{\mathrm{sq}} V^2 \log(2/\delta)/(B^2 n)})$ yields

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) + \lambda\langle \hat{\boldsymbol{w}}, \hat{\boldsymbol{w}}\rangle_{\mathbb{X}} \leq L_\star^{\mathrm{sq}} + O\left(\sqrt{\frac{L_\star^{\mathrm{sq}} B^2 V^2 \log(2/\delta)}{n}} + \frac{(L_\star^{\mathrm{sq}} + B^2V^2)\log(2/\delta)}{n}\right)$$

as soon as $n \geq \tilde{O}(B^2V^2\log(2/\delta)/L_\star^{\mathrm{sq}})$. ∎

By this analysis, a constant factor approximation for $L_\star^{\mathrm{sq}}$ is achieved with a sample of size $\tilde{O}(B^2V^2\log(1/\delta)/L_\star^{\mathrm{sq}})$. As in the finite-dimensional setting, this rate is known to be optimal up to logarithmic factors (Nussbaum, 1999). Indeed, a similar construction to that from Section 5.2.1 implies

$$\inf_{\hat{\boldsymbol{w}}}\sup_{\boldsymbol{w}_\star}\mathbb{E}\big(L(\hat{\boldsymbol{w}})-L(\boldsymbol{w}_\star)\big) \;\geq\; \Omega\left(\frac{1}{d}\cdot\frac{L_\star B^2V^2\sum_{i=1}^d n_i^{-1}}{B^2V^2+L_\star\sum_{i=1}^d n_i^{-1}}\right) \;\geq\; \Omega\left(\frac{1}{d}\cdot\frac{L_\star B^2V^2 d^2/n}{B^2V^2+L_\star d^2/n}\right) \tag{11}$$

(here, $\boldsymbol{X}\in\{V\boldsymbol{e}_i : i\in[d]\}$ has Euclidean length V almost surely, and B is a bound on the Euclidean length of $\boldsymbol{w}_\star$). For $d=\sqrt{B^2V^2n/\sigma^2}$, the bound becomes

$$\inf_{\hat{\boldsymbol{w}}}\sup_{\boldsymbol{w}_\star}\mathbb{E}\big(L(\hat{\boldsymbol{w}})-L(\boldsymbol{w}_\star)\big) \;\geq\; \Omega\left(\sqrt{\frac{L_\star B^2V^2}{n}}\right).$$

As before, this minimax bound also implies a lower bound on any estimator with exponential concentration.

5.2.3 Heavy-tail Covariates

When the covariates are not bounded or subgaussian, the empirical second-moment matrix may deviate significantly from its population counterpart with non-negligible probability. In this case it is not possible to approximate the norm $\|\boldsymbol{a}\|=\sqrt{\boldsymbol{a}^\top(\Sigma+\lambda\,\mathrm{Id})\boldsymbol{a}}$ in Step 2 of Algorithm 2 using a single small sample (as discussed in Section 5.2.1 and Section 5.2.2). However, we may use Algorithm 3 instead of Algorithm 2, which only requires the stochastic distance measurements to be relatively accurate with some constant probability. The full version of Algorithm 4 is exactly such an implementation.

We now prove Theorem 19. Define $c_\eta := 512(48c)^{2+2/\eta}(6+6/\eta)^{1+4/\eta}$ (which is C_{main} from Srivastava and Vershynin, 2013). The following lemma shows that $O(d)$ samples suffice so that the expected spectral norm distance between the empirical second-moment matrix and Σ is bounded.

Lemma 23 (Implication of Corollary 1.2 from Srivastava and Vershynin, 2013) *Let $\boldsymbol{X}$ satisfy Condition 1, and let $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$ be independent copies of $\boldsymbol{X}$. Let $\widehat{\Sigma}:=\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_i\boldsymbol{X}_i^\top$. For any $\epsilon\in(0,1)$, if $n\geq c_\eta\epsilon^{-2-2/\eta}d$, then*

$$\mathbb{E}\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}-\mathrm{Id}\,\|_2 \;\leq\; \epsilon.$$

Lemma 23 implies that $n_{0.5}=O(c'_\eta d)$ where $c'_\eta=c_\eta\cdot 2^{O(1+1/\eta)}$. Therefore, for $k=O(\log(1/\delta))$, subsampled empirical loss minimization requires $n\geq k\cdot n_{0.5}=O(c'_\eta d\log(1/\delta))$ samples to correctly implement $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$, for ε as in Eq. (5).

Step 5 in Algorithm 4 implements $\mathrm{DIST}^j_{\|\cdot\|}$ as returning f_j such that $f_j(\boldsymbol{v}):=\|\Sigma_{S_j}^{1/2}(\boldsymbol{v}-\boldsymbol{w}_j)\|_2$. First, we show that Assumption 3 holds. By Lemma 23, an i.i.d. sample T of size $O(c'_\eta d)$ suffices so that with probability at least 8/9, for every $\boldsymbol{v}\in\mathbb{R}^d$,

$$(1/2)\|\Sigma^{1/2}(\boldsymbol{v}-\boldsymbol{w}_j)\|_2\leq\|\Sigma_T^{1/2}(\boldsymbol{v}-\boldsymbol{w}_j)\|_2\leq 2\|\Sigma^{1/2}(\boldsymbol{v}-\boldsymbol{w}_j)\|_2.$$

In particular, this holds for $T = S_j$, as long as $|S_j| \geq O(c'_\eta d)$. Thus, for $k = O(\log(1/\delta))$, Assumption 3 holds if $n \geq O(c'_\eta d \log(1/\delta))$. Assumption 4 (independence) also holds, since f_j depends only on S_j, and $S_1, \ldots, S_k$ are statistically independent.

Putting everything together, we have (as in Section 5.2.1) $\alpha = 0.5$ and $\gamma = 1$. We obtain the final bound from Theorem 17 as follows: if $n \geq O(c'_\eta d \log(1/\delta))$, then with probability at least $1 - \delta$,

$$L(\hat{\boldsymbol{w}}) - L(\boldsymbol{w}_\star) = \|\Sigma^{1/2}(\hat{\boldsymbol{w}} - \boldsymbol{w}_\star)\|_2^2 \leq O\left(\frac{\mathbb{E}\|\Sigma^{-1/2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)\|_2^2 \log(1/\delta)}{n}\right). \tag{12}$$

6. Other Applications

In this section we show how the core techniques we discuss can be used for other applications, namely Lasso and low-rank matrix approximation.

6.1 Sparse Parameter Estimation with Lasso

In this section we consider L^1-regularized linear least squared regression (Lasso) (Tibshirani, 1996) with a random subgaussian design, and show that Algorithm 2 achieves the same fast convergence rates for sparse parameter estimation as Lasso, even when the noise is heavy-tailed.

Let $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$ and $\boldsymbol{w}_\star \in \mathbb{R}^d$. Let D be a distribution over $\mathcal{Z}$, such that for $(\boldsymbol{X}, Y) \sim D$, we have $Y = \boldsymbol{X}^\top \boldsymbol{w}_\star + \varepsilon$ where ε is an independent random variable with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] \leq \sigma^2$. We assume that $\boldsymbol{w}_\star$ is sparse: Denote the support of a vector $\boldsymbol{w}$ by $\operatorname{supp}(\boldsymbol{w}) := \{j \in [d] : \boldsymbol{w}_j \neq 0\}$. Then $s := |\operatorname{supp}(\boldsymbol{w}_\star)|$ is assumed to be small compared to d. The *design matrix* for a sample $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is an $l \times d$ matrix with the rows $\boldsymbol{x}_i^\top$.

For $\lambda > 0$, consider the Lasso loss $\ell((\boldsymbol{x}, y), \boldsymbol{w}) = \frac{1}{2}(\boldsymbol{x}^\top \boldsymbol{w} - y)^2 + \lambda\|\boldsymbol{w}\|_1$. Let $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^d$. A random vector $\boldsymbol{X}$ in $\mathbb{R}^d$ is *subgaussian* (with moment 1) if for every vector $\boldsymbol{u} \in \mathbb{R}^d$, $\mathbb{E}[\exp(\boldsymbol{X}^\top \boldsymbol{u})] \leq \exp(\|\boldsymbol{u}\|_2^2/2)$.

The following theorem shows that when Algorithm 2 is used with subsampled empirical loss minimization over the Lasso loss, and D generates a subgaussian random design, then $\boldsymbol{w}$ can be estimated for any type of noise ε, including heavy-tailed noise.

In order to obtain guarantees for Lasso the design matrix must satisfy some regularity conditions. We use the *Restricted Eigenvalue condition* (RE) proposed in Bickel et al. (2009), which we presently define. For $\boldsymbol{w} \in \mathbb{R}^d$ and $J \subseteq [d]$, let $[\boldsymbol{w}]_J$ be the $|J|$-dimensional vector which is equal to $\boldsymbol{w}$ on the coordinates in J. Denote by $\boldsymbol{w}_{[s]}$ the s-dimensional vector with coordinates equal to the s largest coordinates (in absolute value) of $\boldsymbol{w}$. Let $\boldsymbol{w}_{[s]^C}$ be the $(d-s)$-dimensional vector which includes the coordinates not in $\boldsymbol{w}_{[s]}$. Define the set $E_s = \{\boldsymbol{u} \in \mathbb{R}^d \setminus \{0\} \mid \|\boldsymbol{u}_{[s]^C}\|_1 \leq 3\|\boldsymbol{u}_{[s]}\|_1\}$. For an $l \times d$ matrix Ψ (for some integer l), let $\gamma(\Psi, s) = \min_{\boldsymbol{u} \in E_s} \frac{\|\Psi \boldsymbol{u}\|_2}{\|\boldsymbol{u}_{[s]}\|_2}$. The RE condition for Ψ with sparsity s requires that $\gamma(\Psi, s) > 0$. We further denote $\eta(\Psi, s) = \max_{\boldsymbol{u} \in \mathbb{R}^d \setminus \{0\} : |\operatorname{supp}(\boldsymbol{u})| \leq s} \frac{\|\Psi \boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2}$.

Theorem 24 *Let $C, c > 0$ be universal constants. Let $\Sigma \in \mathbb{R}^{d \times d}$ be a positive semi definite matrix. Denote $\eta := \eta(\Sigma^{\frac{1}{2}}, s)$ and $\gamma := \gamma(\Sigma^{\frac{1}{2}}, s)$. Assume the random design setting* *defined above, with* $\boldsymbol{X} = \Sigma^{\frac{1}{2}}\boldsymbol{Z}$*, where* $\boldsymbol{Z}$ *is a subgaussian random vector. Suppose Algorithm 2 uses subsampled empirical loss minimization with the empirical Lasso loss, with* $\lambda = 2\sqrt{\sigma^2\eta^2\log(2d)\log(1/\delta)/n}$*. If* $n \geq cs\frac{\eta^2}{\gamma^2}\log(d)\log(1/\delta)$*, then with probability* $1-\delta$*, The vector* $\hat{\boldsymbol{w}}$ *returned by Algorithm 2 satisfies*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_2 \leq \frac{C\sigma\eta}{\gamma^2}\sqrt{\frac{s\log(2d)\log(1/\delta)}{n}}.$$

For the proof of Theorem 24, we use the following theorem, adapted from Bickel et al. (2009) and Zhang (2009). The proof is provided in Appendix A for completeness.

Theorem 25 (Bickel et al. (2009); Zhang (2009)) *Let* $\Psi = [\Psi_1|\Psi_2|\dots|\Psi_d] \in \mathbb{R}^{n\times d}$ *and* $\boldsymbol{\varepsilon} \in \mathbb{R}^n$*. Let* $y = \Psi\boldsymbol{w}_\star + \boldsymbol{\varepsilon}$ *and* $\hat{\boldsymbol{w}} \in \operatorname{argmin}_{\boldsymbol{w}} \frac{1}{2}\|\Psi\boldsymbol{w} - y\|_2^2 + \lambda\|\boldsymbol{w}\|_1$*. Assume that* $|\operatorname{supp}(\boldsymbol{w}_\star)| = s$ *and that* $\gamma(\Psi, s) > 0$*. If* $\|\Psi^\top\boldsymbol{\varepsilon}\|_\infty \leq \lambda/2$*, then*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_2 \leq \frac{12\lambda\sqrt{s}}{\gamma^2(\Psi, s)}.$$

Proof [of Theorem 24] Fix $i \in [k]$, and let $n_i = n/k$. Let $\Psi \in \mathbb{R}^{n_i\times d}$ be the design matrix for S_i and let $\boldsymbol{w}_i$ be the vector returned by the algorithm in round i, $\boldsymbol{w}_i \in \operatorname{argmin} \frac{1}{2n}\|\Psi\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1$. It is shown in Zhou (2009) that if $n_i \geq C\frac{\eta^2}{\gamma^2}s\log(d)$ for a universal constant C, then with probability 5/6, $\min_{\boldsymbol{u}\in E_s}\frac{\|\Psi\boldsymbol{u}\|_2}{\|\Sigma^{\frac{1}{2}}\boldsymbol{u}\|_2} \geq \sqrt{n_i}/2$. Call this event $\mathcal{E}$. By the definition of γ, we have that under $\mathcal{E}$,

$$\gamma(\Psi, s) = \min_{\boldsymbol{u}\in E_s}\frac{\|\Psi\boldsymbol{u}\|_2}{\|\boldsymbol{u}_{[s]}\|_2} = \min_{\boldsymbol{u}\in E_s}\frac{\|\Psi\boldsymbol{u}\|_2}{\|\Sigma^{\frac{1}{2}}\boldsymbol{u}\|_2}\frac{\|\Sigma^{\frac{1}{2}}\boldsymbol{u}\|_2}{\|\boldsymbol{u}_{[s]}\|_2} \geq \sqrt{n}\,\gamma/2.$$

If $\mathcal{E}$ holds and $\|\Psi^\top\boldsymbol{\varepsilon}\|_\infty \leq n\lambda/2$, then we can apply Theorem 25 (with $n\lambda$ instead of λ). We now show that this inequality holds with a constant probability. Fix the noise vector $\boldsymbol{\varepsilon} = \boldsymbol{y} - \Psi\boldsymbol{w}_\star$. For $l \in [d]$, since the coordinates of $\boldsymbol{\varepsilon}$ are independent and each row of Ψ is an independent copy of the vector $\boldsymbol{X} = \Sigma^{\frac{1}{2}}\boldsymbol{Z}$, we have

$$\mathbb{E}[\exp([\Psi^\top\boldsymbol{\varepsilon}]_l) \mid \boldsymbol{\varepsilon}] = \prod_{j\in[n]}\mathbb{E}[\exp(\Psi_{j,l}\boldsymbol{\varepsilon}_j) \mid \boldsymbol{\varepsilon}] = \prod_{j\in[n]}\mathbb{E}[\exp(\boldsymbol{Z}(\boldsymbol{\varepsilon}_j\Sigma^{\frac{1}{2}}\boldsymbol{e}_l)) \mid \boldsymbol{\varepsilon}].$$

Since $\|\boldsymbol{\varepsilon}_j\Sigma^{\frac{1}{2}}\boldsymbol{e}_l\|_2 \leq \boldsymbol{\varepsilon}_j\eta$, we conclude that

$$\mathbb{E}[\exp([\Psi^\top\boldsymbol{\varepsilon}]_l) \mid \boldsymbol{\varepsilon}] \leq \prod_{j\in[n]}\exp(\boldsymbol{\varepsilon}_j^2/2) = \exp(\eta^2\|\boldsymbol{\varepsilon}\|_2^2/2).$$

Therefore, for $\xi > 0$

$$\begin{aligned}
\xi\mathbb{E}[\|\Psi^\top\boldsymbol{\varepsilon}\|_\infty \mid \boldsymbol{\varepsilon}] &= \mathbb{E}[\max_l(\xi|[\Psi^\top\boldsymbol{\varepsilon}]_l|) \mid \boldsymbol{\varepsilon}] = \mathbb{E}[\log\max_l\exp(\xi|[\Psi^\top\boldsymbol{\varepsilon}]_l|) \mid \boldsymbol{\varepsilon}] \\
&\leq \mathbb{E}[\log\left(\sum_l\exp(\xi[\Psi^\top\boldsymbol{\varepsilon}]_l) + \exp(-\xi[\Psi^\top\boldsymbol{\varepsilon}]_i)\right) \mid \boldsymbol{\varepsilon}] \\
&\leq \log\left(\sum_l\mathbb{E}[\exp(\xi[\Psi^\top\boldsymbol{\varepsilon}]_l) \mid \boldsymbol{\varepsilon}] + \mathbb{E}[\exp(-\xi[\Psi^\top\boldsymbol{\varepsilon}]_l) \mid \boldsymbol{\varepsilon}]\right) \\
&\leq \log(2d) + \xi^2\eta^2\|\boldsymbol{\varepsilon}\|_2^2/2.
\end{aligned}$$

Since $\mathbb{E}[\boldsymbol{\varepsilon}_j^2] \leq \sigma^2$ for all j, we have $\mathbb{E}[\|\boldsymbol{\varepsilon}\|^2] \leq n_i\sigma^2/2$. Therefore

$$\mathbb{E}[\|\Psi^\top \boldsymbol{\varepsilon}\|_\infty] \leq \frac{\log(2d)}{\xi} + \xi n_i \eta^2 \sigma^2/2.$$

Minimizing over $\xi > 0$ we get $\mathbb{E}[\|\Psi^\top \boldsymbol{\varepsilon}\|_\infty] \leq 2\sqrt{\sigma^2\eta^2 \log(2d) n_i/2}$. therefore by Markov's inequality, with probability at least $5/6$, $\frac{1}{n_i}\|\Psi^\top \boldsymbol{\varepsilon}\|_\infty \leq 2\sqrt{\sigma^2\eta^2\log(2d)/n_i} = \lambda$. With probability at least $2/3$ this holds together with $\mathcal{E}$.

In this case, by Theorem 25,

$$\|\boldsymbol{w}_i - \boldsymbol{w}_\star\|_2 \leq \frac{12\lambda\sqrt{s}}{\gamma^2(\Psi, s)} \leq \frac{24}{\gamma^2}\sqrt{\frac{s\sigma^2\eta^2\log(2d)}{n_i}}.$$

Therefore $\text{APPROX}_{\|\cdot\|,\epsilon}$ satisfies Assumption 1 with ϵ as in the right hand side above. The statement of the theorem now follows by applying Proposition 9 with $k = O(\log(1/\delta)$, and noting that $n_i = O(n/\log(1/\delta))$. ■

It is worth mentioning that we can apply our technique to the fixed design setting, where design matrix $X \in \mathbb{R}^{n\times d}$ is fixed and not assumed to come from any distribution. If X satisfies the RE condition, as well as a certain low-leverage condition—specifically, that the *statistical leverage scores* (Chatterjee and Hadi, 1986) of any $n \times O(s)$ submatrix of X be roughly $O(1/(ks\log d))$—then Algorithm 2 can be used with the subsampled empirical loss minimization implementation of $\text{APPROX}_{\|\cdot\|,\varepsilon}$ to obtain similar guarantees as in the random subgaussian design setting.

We note that while standard analyses of sparse estimation with mean-zero noise assume light-tailed noise (Zhang, 2009; Bickel et al., 2009), there are several works that analyze sparse estimation with heavy-tailed noise under various assumptions. For example, several works assume that the median of the noise is zero (*e.g.*, Wang 2013; Belloni and Chernozhukov 2011; Zou and Yuan 2008; Wu and Liu 2009; Wang et al. 2007; Fan et al. 2012). van de Geer and Müller (2012) analyze a class of optimization functions that includes the Lasso and show polynomial convergence under fourth-moment bounds on the noise. Chatterjee and Lahiri (2013) study a two-phase sparse estimator for mean-zero noise termed the Adaptive Lasso, proposed in Zou (2006), and show asymptotic convergence results under mild moment assumptions on the noise.

6.2 Low-rank Matrix Approximation

The proposed technique can be easily applied also to low-rank covariance matrix approximation for heavy tailed distributions. Let $\mathcal{D}$ be a distribution over $\mathcal{Z} = \mathbb{R}^d$ and suppose our goal is to estimate $\varSigma = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ to high accuracy, assuming that $\varSigma$ is (approximately) low rank. Here $\mathbb{X}$ is the space of $\mathbb{R}^{d\times d}$ matrices, and $\|\cdot\|$ is the spectral norm. Denote the Frobenius norm by $\|\cdot\|_F$ and the trace norm by $\|\cdot\|_{\text{tr}}$. For $S = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \subseteq \mathbb{R}^d$, define the empirical covariance matrix $\varSigma_S = \frac{1}{n}\sum_{i\in[n]} \boldsymbol{X}_i\boldsymbol{X}_i^\top$. We have the following result for low-rank estimation:

Lemma 26 (Koltchinskii et al. 2011) *Let $\hat{\Sigma} \in \mathbb{R}^{d\times d}$. Assume $\lambda \geq \|\hat{\Sigma} - \Sigma\|$, and let*

$$\Sigma_\lambda \in \operatorname*{argmin}_{A\in\mathbb{R}^{d\times d}} \frac{1}{2}\|\hat{\Sigma} - A\|_F^2 + \lambda\|A\|_{\mathrm{tr}}, \tag{13}$$

If $\lambda \geq \|\hat{\Sigma} - \Sigma\|$, then

$$\frac{1}{2}\|\hat{\Sigma}_\lambda - \Sigma\|_F^2 \leq \inf_{A\in\mathbb{R}^{d\times d}} \left\{ \frac{1}{2}\|A - \Sigma\|_F^2 + \frac{1}{2}(\sqrt{2}+1)^2\lambda^2 \operatorname{rank}(A) \right\}.$$

Now, assume condition 1 holds for $\mathcal{X} \sim \mathcal{D}$, and suppose for simplicity that $\|\Sigma\| \leq 1$. In this case, by Lemma 23, A random sample S of size $n' = c'_\eta \epsilon^{-2-2/\eta} d$, where $c'_\eta = c_\eta (3/2)^{2+2/\eta}$ suffices to get an empirical covariance matrix Σ_S such that $\|\Sigma_S - \Sigma\| \leq \epsilon$ with probability at least $2/3$.

Given a sample of size n from $\mathcal{D}$, We can thus implement $\text{APPROX}_{\|\cdot\|,\varepsilon}$ that simply returns the empirical covariance matrix of a sub-sample of size $n' = n/k$, so that Assumption 1 holds for an appropriate ε. By Proposition 9, Algorithm 2 returns $\hat{\Sigma}$ such that with probability at least $1 - \exp(-k/18)$, $\|\hat{\Sigma} - A\| \leq 3\epsilon$. The resulting $\hat{\Sigma}$ can be used to minimize Eq. (13) with $\lambda = 3\epsilon := O\left((c'_\eta d \log(1/\delta)/n)^{1/2(1+1/\eta)}\right)$. The output matrix Σ_λ satisfies, with probability at least $1 - \delta$,

$$\frac{1}{2}\|\Sigma_\lambda - \Sigma\|_F^2 \leq \inf_{A\in\mathbb{R}^{d\times d}} \left\{ \frac{1}{2}\|A - \Sigma\|_F^2 + O\left((c'_\eta d \log(1/\delta)/n)^{1/(1+1/\eta)}\right) \cdot \operatorname{rank}(A) \right\}.$$

7. A Comparison of Robust Distance Approximation Methods

The approach described in Section 3 for selecting a single $\boldsymbol{w}_i$ out of the set $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$, gives one Robust Distance Approximation procedure (see Def. 1), in which the $\boldsymbol{w}_i$ with the lowest median distance from all others is selected. In this section we consider other Robust Distance Approximation procedures and their properties. We distinguish between procedures that return $y \in W$, which we term *set-based*, and procedures that might return any $y \in \mathbb{X}$, which we term *space-based*.

Recall that we consider a metric space $(\mathbb{X}, \rho)$, with $W \subseteq \mathbb{X}$ a (multi)set of size k and $w_\star$ a distinguished element. Let $W_+ := W \cup \{w_\star\}$. In this formalization, the procedure used in Algorithm 2 is to simply select $y \in \operatorname{argmin}_{w\in W} \Delta_W(w, 0)$, a set-based procedure. A natural variation of this is the space-based procedure: select $y \in \operatorname{argmin}_{w\in\mathbb{X}} \Delta_W(w, 0)$.[6] A different approach, proposed by Minsker (2013), is to select $y \in \operatorname{argmin}_{w\in\mathbb{X}} \sum_{\bar{w}\in W} \rho(w, \bar{w})$, that is to minimize the geometric median over the space. Minsker analyzes this approach for Banach and Hilbert spaces. We show that minimizing the geometric median also achieves similar guarantees in general metric spaces.

In the following, we provide detailed guarantees for the approximation factor C_α of the two types of procedures, for general metric spaces as well as for Banach and Hilbert spaces, and for set-based and sample-based procedures. We further provide lower bounds for specific procedures, as well as lower bounds that hold for any procedure. In Section 7.4

6. The space-based median distance approach might not always be computationally feasible; see discussion in Section 7.4.

we summarize the results and compare the guarantees of the two procedures and the lower bounds. For a more useful comparison, we take into account the fact that the value of α usually affects not only the approximation factor, but also the upper bound obtained for $\Delta_W(w_\star, \alpha)$.

7.1 Minimizing the Median Distance

Minimizing the median distance over the set of input points was shown in Proposition 8 to achieve an approximation factor of 3. In this section we show that this upper bound on the approximation factor is tight for this procedure, even in a Hilbert space. Here and below, we say that an approximation factor upper bound is *tight* if for any constant smaller than this upper bound, there are a suitable space and a set of points in that space, such that the procedure achieves for this input a larger approximation factor than said constant.

The approximation factor can be improved to 2 for a sample-based procedure. This factor is tight as well, even assuming a Hilbert space. The following theorem summarizes these facts.

Theorem 27 *Let $k \geq 2$, and suppose that $\Delta_W(w_\star, \gamma) \leq \epsilon$ for some $\gamma > 0$. Let $y \in \operatorname{argmin}_{w \in W} \Delta_W(w, 0)$. Further, suppose that $W_+ \subseteq \mathbb{X}$, and let $\bar{y} \in \operatorname{argmin}_{w \in \mathbb{X}} \Delta_W(w, 0)$. Then*

- *For any metric space, $\rho(w_\star, y) \leq 3\epsilon$;*
- *For any metric space, $\rho(w_\star, \bar{y}) \leq 2\epsilon$;*
- *There exists a set on the real line such that $\rho(w_\star, y) = 3\epsilon$, where ρ is the distance induced by the inner product;*
- *There exists a set on the real line such that $\rho(w_\star, \bar{y}) = 2\epsilon$, where ρ is the distance induced by the inner product.*

Proof First, we prove the two upper bounds. Since $\Delta_W(w_\star, \gamma) \leq \epsilon$, we have $|B(w_\star, \epsilon) \cap W| > k/2$. Let $w \in |B(w_\star, \epsilon) \cap W|$. Then by the triangle inequality, $B(w, 2\epsilon) \supseteq B(w_\star, \epsilon)$. Therefore $\Delta_W(w, 0) \leq 2\epsilon$. It follows that $\Delta_W(y, 0) \leq 2\epsilon$, hence $|B(y, 2\epsilon) \cap W| \geq k/2$. By the pigeon hole principle, $|B(w_\star, \epsilon) \cap B(y, 2\epsilon)| > 0$, therefore $\rho(w_\star, y) \leq 3\epsilon$.

As for $\bar{y}$, since this is a minimizer over the entire space $\mathbb{X}$ which includes $w_\star$, we have $\Delta_W(y, \gamma) \leq \Delta_W(w_\star, \gamma) \leq \epsilon$. Therefore, similarly to the argument for y, we have $\rho(w_\star, y) \leq 2\epsilon$.

To see that these bounds are tight, we construct simple examples on the real line. For y, suppose $w_\star = \epsilon$, and consider W with k points as follows: $k/2 - 1$ points at 0, 2 points at 2ϵ, and $k/2 - 1$ points at 4ϵ. The points at 4ϵ are clearly in $\operatorname{argmin}_{w \in W} \Delta_W(w, 0)$, therefore $\rho(w_\star, y) = 3\epsilon$.

For $\bar{y}$, suppose $w_\star = \epsilon$, and consider W with k points as follows: 2 points at 0, $k/2 - 1$ points at 2ϵ, and $k/2 - 1$ points at 3ϵ. The points at 3ϵ are clearly in $\operatorname{argmin}_{w \in W_+} \Delta_W(w, 0)$, therefore $\rho(w_\star, \bar{y}) = 2\epsilon$. ∎

The non-uniqueness of the median distance minimizer is exploited in the lower bounds in Theorem 27. This suggests that some kind of aggregation of the median distance minimizers may provide a smaller bound at least in certain scenarios.

7.2 The Geometric Median

For $w \in \mathbb{X}$, denote the sum of distances from points in the input set by $\mathrm{sumd}(w) := \sum_{v \in W} \rho(w, v)$. Minsker (2013) suggests to minimize the sum of distances over the entire space, that is, to select the geometric median. Minsker shows that when this procedure is applied in a Hilbert space, $C_\alpha \leq \frac{\frac{1}{2}+\alpha}{\sqrt{2\alpha}}$, and for a Banach space $C_\alpha \leq 1 + \frac{1}{2\alpha}$. Here we show that in fact $C_\alpha \leq 1 + \frac{1}{2\alpha}$ for general metric spaces. The proof holds, in particular, for Banach spaces, and thus this provide a more direct argument that does not require the special properties of Banach spaces. We further show that for general metric spaces, this upper bound on the approximation factor is tight.

Minimizing over the entire space is a computationally intensive procedure, involving convex approximation. Moreover, if the only access to the metric is via estimated distances based on samples, as in Algorithm 3, then there are additional statistical challenges. It is thus of interest to also consider the simpler set-based procedure, and we provide approximation guarantees for this procedure as well. We show that an approximation factor of $2 + \frac{1}{2\alpha}$ can be guaranteed for set-based procedures in general metric spaces, and this is also tight, even for Banach spaces.

The following theorem provides a bound that holds in several of these settings.

Theorem 28 *Let $k \geq 2$. Let $y \in \mathrm{argmin}_{w \in W} \mathrm{sumd}(w)$, and let $\bar{y} \in \mathrm{argmin}_{w \in W_+} \mathrm{sumd}(w)$. Then*

1. *For any metric space $(\mathbb{X}, \rho)$ and W, W_+,*
$$\rho(w_\star, y) \leq \left(2 + \frac{1}{2\alpha}\right) \Delta_W(w_\star, \alpha).$$
2. *For any constant $C < (2 + \frac{1}{2\alpha})$, there exists a problem in a Banach space such that $\rho(w_\star, y) > C \cdot \Delta_W(w_\star, \alpha)$. Thus the upper bound above is tight.*
3. *For any metric space $(\mathbb{X}, \rho)$ and W, W_+,*
$$\rho(w_\star, \bar{y}) \leq \left(1 + \frac{1}{2\alpha}\right) \Delta_W(w_\star, \alpha).$$
4. *For any constant $C < (1 + \frac{1}{2\alpha})$, there exists a problem in a metric space such that $\rho(w_\star, \bar{y}) > C \cdot \Delta_W(w_\star, \alpha)$. Thus the upper bound above is tight for general metric spaces.*

Proof Let $w \in \mathrm{argmin}_{w \in B(w_\star, \epsilon) \cap W} \rho(w, y)$. Let $Z \subset B(w_\star, \epsilon) \cap W$ such that $|Z| = k(\frac{1}{2} + \alpha)$ (we assume for simplicity that $k(\frac{1}{2} + \alpha)$ is an integer; the proof can be easily modified to accommodate the general case). For $v \in Z$, $\rho(w,v) \le \rho(w,w_\star) + \rho(w_\star,v)$. For $v \in W \setminus Z$, $\rho(w,v) \le \rho(w,y) + \rho(y,v)$. Therefore

$$\mathrm{sumd}(w) \le \sum_{v \in Z} (\rho(w,w_\star) + \rho(w_\star,v)) + \sum_{v \in W \setminus Z} (\rho(w,y) + \rho(y,v)).$$

By the definition of w as a minimizer, for $v \in Z$, $\rho(y,v) \ge \rho(y,w)$. Thus

$$\mathrm{sumd}(y) \ge \sum_{v \in Z} \rho(y,w) + \sum_{v \in W \setminus Z} \rho(y,v).$$

Since $\mathrm{sumd}(y) \le \mathrm{sumd}(w)$, we get

$$\sum_{v \in Z} \rho(y,w) + \sum_{v \in W \setminus Z} \rho(y,v) \le \sum_{v \in Z} (\rho(w,w_\star) + \rho(w_\star,v)) + \sum_{v \in W \setminus Z} (\rho(w,y) + \rho(y,v)).$$

Hence, since $\rho(v,w_\star) \le \epsilon$ for $v \in Z$,

$$(|Z| - |W \setminus Z|)\rho(w,y) \le 2|Z|\epsilon.$$

Since $|Z| = k(\frac{1}{2} + \alpha)$ it follows that $\rho(w,y) \le (1 + \frac{1}{2\alpha})\epsilon$. In addition,

$$\rho(w_\star,y) \le \rho(w,w_\star) + \rho(w,y) \le \epsilon + \rho(w,y),$$

therefore

$$\rho(w_\star,y) \le \left(2 + \frac{1}{2\alpha}\right)\epsilon.$$

This shows that for any metric space, the set-based geometric median gives an approximation factor of $2 + \frac{1}{2\alpha}$, proving item 1.

For the space-based geometric median, consider $\bar{w} \in \operatorname{argmin}_{w \in B(w_\star,\epsilon)\cap W} \rho(w,\bar{y})$. We have $\mathrm{sumd}(\bar{y}) \le \mathrm{sumd}(w_\star)$. In addition,

$$\mathrm{sumd}(w_\star) \le \sum_{v \in Z} \rho(w_\star,v) + \sum_{v \in W \setminus Z} (\rho(w_\star,\bar{w}) + \rho(\bar{w},\bar{y}) + \rho(\bar{y},v)).$$

Therefore,

$$\sum_{v \in Z} \rho(\bar{y},\bar{w}) + \sum_{v \in W \setminus Z} \rho(\bar{y},v) \le \sum_{v \in Z} \rho(w_\star,v) + \sum_{v \in W \setminus Z} (\rho(w_\star,w) + \rho(w,\bar{y}) + \rho(\bar{y},v)).$$

Since $\rho(w_\star,v) \le \epsilon$ for $v \in Z$, and $\rho(w_\star,\bar{w}) \le \epsilon$, it follows

$$(|Z| - |W \setminus Z|)\rho(\bar{w},\bar{y}) \le k\epsilon.$$

Therefore $\rho(\bar{w},\bar{y}) \le \frac{1}{2\alpha}\epsilon$, hence

$$\rho(w_\star,\bar{y}) \le \rho(w_\star,\bar{w}) + \rho(\bar{w},\bar{y}) \le \left(1 + \frac{1}{2\alpha}\right)\epsilon.$$

This gives an approximation factor of $1 + \frac{1}{2\alpha}$ for space-based geometric median, proving item 3.

To see that both of these bounds are tight, let $n = k(\frac{1}{2} + \alpha)$, and let $\mathbb{X} = W_+ = \{v_1, \ldots, v_n, y_1, \ldots, y_{k-n}, w_\star\}$. Define $\rho(\cdot, \cdot)$ as follows (for all pairs $i \neq j$, $l \neq t$):

$$\begin{aligned}
&\rho(w_\star, v_i) = \epsilon\\
&\rho(w_\star, y_l) = \beta\\
&\rho(v_i, v_j) = 2\epsilon\\
&\rho(v_i, y_t) = \beta - \epsilon\\
&\rho(y_t, y_l) = 0.
\end{aligned}$$

One can verify that for any $\beta \leq (2 + \frac{1}{2\alpha} - \frac{1}{k\alpha})\epsilon$, $\mathrm{sumd}(y_l) \leq \mathrm{sumd}(v_i)$ for all l, i. Therefore, the approximation factor for set-based geometric median in a general metric space is lower-bounded by $2 + \frac{1}{2\alpha}$ for general k. This holds also for Banach spaces as well, Since any metric space can be embedded into a Banach space (Kuratowski, 1935). This proves item 2.

For space-based geometric median, note that if $\beta \leq (1 + \frac{1}{2\alpha})\epsilon$, then $\mathrm{sumd}(w_\star) \geq \mathrm{sumd}(y_l)$. Therefore the space-based upper bound is tight for a general metric space. This proves item 4. ■

Since $\alpha \in (0, \frac{1}{2})$, the guarantee for the geometric median in these settings is always worse than the guarantee for minimizing the median distance. Factoring in the dependence on α, the difference is even more pronounced. The full comparison is given in Section 7.4 below.

7.3 Optimal Approximation Factor

In this section we give lower bounds that hold for any robust distance approximation procedure. A lower bound of $C > 0$ for a category of metric spaces and a type of procedure indicates that if a procedure of this type guarantees a distance approximation C_α for all metric spaces of the given category, then necessarily $C_\alpha \geq C$. As shown below, in many cases the lower bounds provided here match the upper bounds obtained by either the median distance or the geometric median.

The following theorem gives a lower bound of 3 for the achievable approximation factor of set-based procedures in Banach spaces (and so, also in general metric spaces). This factor is achieved by the median distance minimizer, as shown in Theorem 27.

Theorem 29 *Consider set-based robust distance approximation procedures. For any $\alpha \in (0, \frac{1}{2})$, and for any such procedure, there exists a problem in a Banach space for which the approximation factor of the procedure is at least* 3.

Proof Fix α, and let $n = \lceil \frac{1}{\frac{1}{2} - \alpha} \rceil$. Define the metric space $\mathbb{X} = \{a_1, \ldots, a_n, b_1, \ldots, b_n\}$ with the metric $\rho(\cdot, \cdot)$ defined as follows: For all $i \neq j$, $\rho(a_i, a_j) = 2$, $\rho(a_i, b_j) = 1$, $\rho(b_i, b_j) = 2$. For all i, $\rho(a_i, b_i) = 3$. See Figure 2 for illustration.

Consider the multi-set W with k/n elements at every b_i. It is easy to check that for every a_i, $\Delta_W(a_i, \alpha) \leq \Delta_W(a_i, 1/2 - 1/n) = 1$. On the other hand, since the problem is

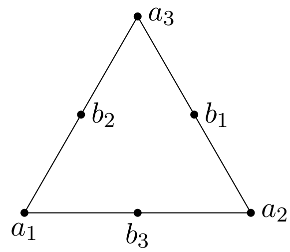

Figure 2: The metric defined in Theorem 29 for $n = 3$. The distances are shortest paths on the underlying undirected graph, where all edges are the same length.

symmetric for permutations of the indices $1, \ldots, n$, no procedure can distinguish the cases $w_\star = a_i$ for different $i \in [n]$. For any choice $y = b_i \in W$, if $w_\star = a_i$ then $\rho(w_\star, y) = 3$. Therefore the approximation factor of any procedure is at least 3. Since any metric space can be embedded into a Banach space (Kuratowski, 1935) this result holds also for Banach spaces. ■

Next, we give a lower bound of 2 for space-based procedures over general metric spaces. Theorem 27 shows that this factor is also achieved by minimizing the median distance.

Theorem 30 *Consider robust space-based distance approximation procedures. For any $\alpha \in (0, \frac{1}{2})$, and for any such procedure, there exists a problem for which the approximation factor of the procedure is at least* 2*.*

Proof Fix α, and let $n = \lceil \frac{1}{\frac{1}{2} - \alpha} \rceil$. Define the metric space $\mathbb{X} = \{a_1, \ldots, a_n, b_1, \ldots, b_n\}$ with the metric $\rho(\cdot, \cdot)$ defined as follows: For all $i \neq j$, $\rho(a_i, a_j) = 2$, $\rho(a_i, b_j) = 1$, $\rho(b_i, b_j) = 1$. For all i, $\rho(a_i, b_i) = 2$. See Figure 3 for illustration.

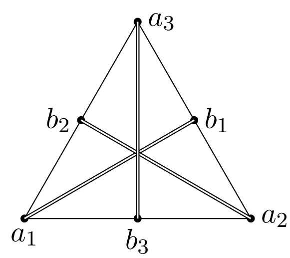

Figure 3: The metric defined in Theorem 29 for $n = 3$. The distances are shortest paths on the underlying undirected graph. The full lines are edges of length 1, the double lines from a_i to b_i are edges of length 2.

Consider the multi-set W with k/n points at every b_i. It is easy to check that for every a_i, $\Delta_W(a_i, \alpha) \leq \Delta_W(a_i, 1/2 - 1/n) = 1$. On the other hand, since the problem is symmetric for permutations of the indices $1, \ldots, n$, no procedure can distinguish the cases $w_\star = a_i$ for different $i \in [n]$. Moreover, any point y in the space has $\rho(a_i, y) = 2$ for at least one $i \in [n]$. Therefore the approximation factor of any procedure is at least 2. ∎

For lower bounds on Hilbert spaces and Banach spaces, we require the following lemma, which gives the radius of the ball inscribing the regular simplex in a p-normed space.

Lemma 31 *Consider $\mathbb{R}^n$ with the p-norm for $p > 1$. Let $e_1, \ldots, e_n$ be the standard basis vectors, and let $r_{n,p}$ be the minimal number for which there exists an $x \in \mathbb{R}^n$ such that $B(x, r) \supseteq \{e_1, \ldots, e_n\}$. Then $r_{n,p} = ((1+(n-1)^{-1/(p-1)})^{-p}+(n-1)(1+(n-1)^{1/(p-1)})^{-p})^{1/p}$. This radius is obtained with the center x such that for all i, $x_i = (1 + (n-1)^{1/(p-1)})^{-1}$.*

Proof It is easy to see that due to symmetry, $x = (a, a, \ldots, a)$ for some real number a. Thus $r_{n,p} = \inf_{a\in\mathbb{R}} \|e_1 - (a, \ldots, a)\|_p$. We have $\|e_1 - (a, \ldots, a)\|_p^p = |1 - a|^p + (n-1)|a|^p$. Minimizing over a gives $a = (1 + (n-1)^{1/(p-1)})^{-1}$, and

$$r_{n,p}^p = |1 - a|^p + (n-1)|a|^p = (1 + (n-1)^{-1/(p-1)})^{-p} + (n-1)(1 + (n-1)^{1/(p-1)})^{-p}.$$

∎

We now prove a lower bound for robust distance approximation in Hilbert spaces. Unlike the previous lower bounds, this lower bound depends on the value of α.

Theorem 32 *Consider robust distance approximation procedures for $(\mathbb{X}, \rho)$ a Hilbert space. For any $\alpha \in (0, \frac{1}{2})$, the following holds:*

- *For any set-based procedure, there exists a problem such that the procedure achieves an approximation factor at least*

$$\sqrt{1 + \frac{2}{\left\lceil \frac{1}{\frac{1}{2}-\alpha} \right\rceil - 2}}.$$

- *For any space-based procedure, there exists a problem such that the procedure achieves an approximation factor at least*

$$\sqrt{1 + \frac{1}{\left\lceil \frac{1}{\frac{1}{2}-\alpha} \right\rceil^2 - 2\left\lceil \frac{1}{\frac{1}{2}-\alpha} \right\rceil}}.$$

The space-based bound given in Theorem 32 is tight for $\alpha \to 1/2$. This can be seen by noting that the limit of the space-based lower bound for $\alpha \to 1/2$ is $(\frac{1}{2} + \alpha)/\sqrt{2\alpha}$, which is exactly the guarantee provided in Minsker (2013) for the space-based geometric median procedure. For smaller α, there is a gap between the guarantee of Minsker for the geometric median and our lower bound.

Proof Fix α, and let $n = \lceil \frac{1}{\frac{1}{2}-\alpha} \rceil$. Consider the Euclidean space $\mathbb{R}^n$ with $\rho(x, y) = \|x - y\|$. Let $e_1, \ldots, e_n$ be the standard basis vectors. These are the vertices of a regular simplex with side length $\|e_i - e_j\| = \sqrt{2}$. Let $b_1, \ldots, b_n$ such that b_i is the center of the hyperface of the simplex opposing e_i. Then $\|b_i - e_j\| = r_{n-1,2}$ for all $j \neq i$, where $r_{n,2} = \sqrt{\frac{n-1}{n}}$ is as defined in Lemma 31. (see Figure 4).

Consider W with k/n points at each of $b_1, \ldots, b_n$. Then $\Delta_W(e_i, \alpha) \leq \Delta_W(e_i, 1 - \frac{1}{n}) = \|e_i - b_j\| = r_{n-1,2}$ for any $j \neq i$. Any set-based procedure must select b_i for some i. if $w_\star = e_i$, the resulting approximation factor is $\|e_i - b_i\|/r_{n-1,2} = \sqrt{\frac{n-2}{n-1}}\|e_i - b_i\|$. For $\|b_i - e_i\|$, consider for instance b_1 and e_1. We have $b_1 = (0, \frac{1}{n-1}, \ldots, \frac{1}{n-1})$, therefore $\|b_1 - e_1\| = \sqrt{\frac{n}{n-1}}$. The approximation factor of the procedure is thus at least $\sqrt{\frac{n}{n-2}}$.

For a set-based procedure, whatever y it returns, there exists at least one i such that $\|y - a_i\| \geq r_{n,2}$. Therefore the approximation factor is at least $r_{n,2}/r_{n-1,2} = \sqrt{\frac{n-1}{n}}/\sqrt{\frac{n-2}{n-1}} = \sqrt{1 + \frac{1}{n^2 - 2n}}$.

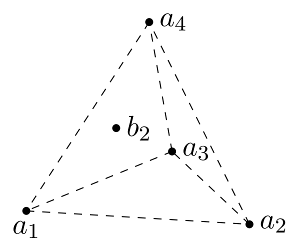

Figure 4: The regular simplex in $\mathbb{R}^3$, $n = 4$. a_i is a vertex, b_i is the center of the face opposite a_i.

■

For space-based procedures, we have seen that while there exists a lower bound of 2 for general metric spaces, in a Hilbert space better approximation factors can be achieved. Is it possible that in Banach spaces the same approximation factor can also be achieved? The following theorem shows that the answer is no.

Theorem 33 *Let $\alpha = 1/6$. There exists a Banach space for which an approximation factor of $(\frac{1}{2} + \alpha)/\sqrt{2\alpha}$ cannot be achieved.*

Proof Consider the space $\mathbb{R}^n$ with the distance defined by a p-norm. Let $n = 1/(\frac{1}{2} - \alpha) = 3$. Construct W as in the proof of Theorem 32, with k/n points in each of $b_1, \ldots, b_n$, where b_i is the center (in the p-norm) of the hyperface opposing the basis vector e_i. As in the proof of Theorem 32, the approximation factor for any space-based procedure for this problem is at least $r_{n,p}/r_{n-1,p}$. For $p = 3/2$, we have $r_{n,p}/r_{n-1,p} = \frac{2}{5^{1/3}} > \frac{2}{\sqrt{3}} = \frac{\frac{1}{2}+\alpha}{\sqrt{2\alpha}}$. ■

	General Metric	Banach	Hilbert
Set-based			
Optimal	$= 3$	$= 3$	$\geq \sqrt{1 + \frac{2}{\left\lceil \frac{1}{\frac{1}{2}-\alpha} \right\rceil - 2}} \xrightarrow{\alpha \to 1/2} 1/\sqrt{2\alpha}$
Median distance	$= 3$	$= 3$	$= 3$
Geometric median	$= 2 + 1/(2\alpha)$	$= 2 + 1/(2\alpha)$	Open
Space-based			
Optimal	$= 2$	Strictly larger than for Hilbert spaces	$\geq \sqrt{1 + \frac{1}{\left\lceil \frac{1}{\frac{1}{2}-\alpha} \right\rceil^2 - 2\left\lceil \frac{1}{\frac{1}{2}-\alpha} \right\rceil}} \xrightarrow{\alpha \to 1/2} \frac{\frac{1}{2}+\alpha}{\sqrt{2\alpha}}$
Median distance	$= 2$	$= 2$	$= 2$
Geometric median	$= 1 + 1/(2\alpha)$	$\leq 1 + 1/(2\alpha)$ $(\star)$	$\leq (\frac{1}{2} + \alpha)/\sqrt{2\alpha}$ $(\star)$

Table 1: Approximation factors for $\alpha \in (0, 1/2)$, based on type of procedure and type of space. Results marked with $(\star)$ are due to Minsker (2013). Equality indicates matching upper and lower bounds.

	General Metric	Banach	Hilbert
Set-based			
Optimal	$= 6$	$= 6$	≥ 3.46
Median distance	$= 6$	$= 6$	$= 6$
Geometric median	$= 14.92$	$= 14.92$	Open
Space-based			
Optimal	$= 4$	Open	≥ 2.31
Median distance	$= 4$	$= 4$	$= 4$
Geometric median	$= 11.65$	≤ 11.65	≤ 3.33

Table 2: Optimal normalized approximation factors based on the values of C_α given in Table 1. The value in each case is $\inf_{\alpha \in (0,\frac{1}{2})} \frac{C_\alpha}{(\frac{1}{2}-\alpha)}$ for the corresponding C_α. All non-integers are rounded to 2 decimal places.

7.4 Comparison of Selection Procedures

The results provided above are summarized in Table 1. When comparing different procedures for different values of α, it is useful to compare not only the respective approximation factors but also the upper bound that can be obtained for $\Delta_W(w_\star, \alpha)$. Typically, as in the proof of Proposition 9, this upper bound will stem from first bounding $\mathbb{E}[\rho(w_\star, w)] \leq \epsilon$, where the expectation is taken over random i.i.d. draws of w, and then applying Markov's inequality to obtain $\mathbb{P}[\rho(w_\star, w) \leq \frac{\epsilon}{\frac{1}{2}-\alpha}] \geq \frac{1}{2} + \alpha$. In the final step Hoeffding's inequality guarantees that if k is large enough, $|B(w_\star, \epsilon/(\frac{1}{2} - \alpha)) \cap W|$ approaches $k(\frac{1}{2} + \alpha)$. Therefore, for a large k and a procedure for α with an approximation factor C_α, the guarantee approaches $\rho(y, w_\star) \leq \frac{C_\alpha}{(\frac{1}{2}-\alpha)} \cdot \epsilon$. For a procedure with an approximation factor C_α, we call $\frac{C_\alpha}{(\frac{1}{2}-\alpha)}$ the *normalized* approximation factor of the procedure. This is the approximation factor with respect to $\mathbb{E}[\rho(w_\star, \alpha)]$. When the procedure supports a range of α, the optimal normalized factor can be found by minimizing $\frac{C_\alpha}{(\frac{1}{2}-\alpha)}$ over $\alpha \in (0, \frac{1}{2})$. If $C_\alpha = C$ is a constant, the optimal normalized approximation factor is $2C$, achieved when $\alpha = 0$. The optimal normalized approximation factors, based on the known approximation factors as a function of α, are given in Table 2.

We observe that for set-based procedures, the median distance is superior to the geometric median for general metric spaces as well as for general Banach spaces. It is an open question whether better results can be achieved for Hilbert spaces using set-based procedures.

For space-based procedures, the median distance is again superior, except in the case of a Hilbert space, where the geometric median is superior. The case of a Hilbert space is arguably the most useful in common applications such as linear regression. Nevertheless, gaps still remain and it would be interesting to develop optimal methods.

Implementing the geometric median procedure in a space-based formulation is computationally efficient for Hilbert spaces when accurate distances are available Minsker (2013). However, it is unknown whether and how the procedure can be implemented when only unreliable distance estimations are available, as in Section 3.3. A useful implementation should be both computationally feasible and statistically efficient, while degrading the approximation factors as little as possible.

8. Predicting Without a Metric on Predictors

The core technique presented above allows selecting a good candidate out of a set that includes mostly good candidates, in the presence of a metric between candidates. If the final goal is prediction of a scalar label, good prediction can still be achieved without access to a metric between candidates, using the following simple procedure: For every input data point, calculate the prediction of every candidate, and output the median of the predictions. This is a straight-forward generalization of voting techniques for classification such as when using bagging (Breiman, 1996).[7] The following lemma shows that this approach leads to guarantees similar to those achieved by Proposition 9.

7. Note, however, that the usual implementation of bagging for regression involves averaging over the outputs of the classifiers, and not taking the median.

Lemma 34 *Let D, $\ell : \mathcal{Z} \times \mathbb{X} \to \mathbb{R}_+$ and $L : \mathbb{X} \to \mathbb{R}_+$ be defined as in Section 4. Assume that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and there are functions $f : \mathcal{X} \times \mathbb{X} \to \mathbb{R}$ (the prediction function) and $g : \mathbb{R} \times \mathbb{R}$ (the link function) such that $\ell((\boldsymbol{x}, y), \boldsymbol{w}) = g(f(\boldsymbol{x}, \boldsymbol{w}), y)$. Assume that g is convex its first argument. Suppose that we have k predictors $w_1, \ldots, w_k$ such that for at least $(\frac{1}{2} + \gamma)k$ of them, $L(\boldsymbol{w}) \leq \bar{\ell}$. For $x \in \mathcal{X}, y \in \mathcal{Y}$, let $\hat{y}(\boldsymbol{x})$ be the median of $f(\boldsymbol{x}, \boldsymbol{w}_1), \ldots, f(\boldsymbol{x}, \boldsymbol{w}_k)$, and let $\hat{\ell}(\boldsymbol{x}, y) = g(\hat{y}(\boldsymbol{x}), y)$. Let $\hat{L} := \mathbb{E}[\hat{\ell}(\hat{y}(\boldsymbol{x}))]$. Then*

$$\hat{L} \leq \left(\frac{1}{2\gamma} + 1\right) \bar{\ell}.$$

Proof Let $I = \{i : L(\boldsymbol{w}_i) \leq \bar{\ell}\}$. Assume without loss of generality that for $i \in [k-1]$, $f(\boldsymbol{x}, \boldsymbol{w}_i) \leq f(\boldsymbol{x}, \boldsymbol{w}_{i+1})$. Let $t \in [k]$ such that $\hat{y}(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{w}_t)$. By the convexity of g, at least one of $g(f(\boldsymbol{x}, \boldsymbol{w}_t), y) \leq g(f(\boldsymbol{x}, \boldsymbol{w}_{t-1}, y))$ and $g(f(\boldsymbol{x}, \boldsymbol{w}_t), y) \leq g(f(\boldsymbol{x}, \boldsymbol{w}_{t+1}, y))$ holds. assume without loss of generality that the first inequality holds. It follows that for all $i \in [t]$, $g(f(\boldsymbol{x}, \boldsymbol{w}_i), y) \geq g(f(\boldsymbol{x}, \boldsymbol{w}_t, y))$. Therefore,

$$\begin{aligned}
\hat{\ell}(\boldsymbol{x}, y) = g(f(\boldsymbol{x}, \boldsymbol{w}_t), y)) &\leq \frac{1}{|I \cap [t]|} \sum_{i \in I \cap [t]} g(f(\boldsymbol{x}, \boldsymbol{w}_i), y) \\
&\leq \frac{1}{|I \cap [t]|} \sum_{i \in I} g(f(\boldsymbol{x}, \boldsymbol{w}_i), y) = \frac{1}{|I \cap [t]|} \sum_{i \in I} \ell((\boldsymbol{x}, y), \boldsymbol{w}_i).
\end{aligned}$$

Taking expectation over $(\boldsymbol{x}, y)$,

$$\hat{L} \leq \frac{1}{|I \cap [t]|} \sum_{i \in I} L(\boldsymbol{w}_i) \leq \frac{|I|}{|I \cap [t]|} \bar{\ell} \leq \frac{\frac{1}{2} + \gamma}{\gamma} \bar{\ell},$$

where the last inequality follows from the assumption that $|I| \geq (\frac{1}{2} + \gamma)k$. ■

A downside of this approach is that each prediction requires many applications of a predictor. If there is also access to unlimited unlabeled data, a possible approach to circumvent this issue is to generate predictions for a large set of random unlabeled data points based on the aggregate predictor, and then use the resulting labeled pairs as a training set to find a single predictor with a loss that approaches the loss of the aggregate predictor. A similar approach for derandomizing randomized classifiers was suggested by Kääriäinen (2005).

9. Conclusion

In this paper we show several applications of a generalized median-of-means approach to estimation. In particular, for linear regression we establish convergence rates for heavy-tailed distributions that match the min-max rates up to logarithmic factors. We further show conditions that allow parameter estimation using the Lasso under heavy-tailed noise, and cases under which low-rank covariance matrix approximation is possible for heavy-tailed distributions.

The core technique is based on performing independent estimates on separate random samples, and then combining these estimates. Other works have considered approaches which resemble this general scheme but provide other types of guarantees. For instance, in Zhang et al. (2013), faster parallel kernel ridge regression is achieved by performing loss minimizations on independent samples and then averaging the resulting estimators. In Rakhlin et al. (2013), faster rates of convergence for regression for some classes of estimators are achieved, using linear combinations of risk minimizers over subsets of the class of estimators. These works, together with ours, demonstrate that empirical risk minimization can be used as a black box to generate new algorithms with improved statistical performance.

Acknowledgments

Part of this work was completed while the authors were at Microsoft Research New England. Daniel Hsu was supported by a Yahoo Academic Career Enhancement Award. Sivan Sabato is supported by the Lynne and William Frankel Center for Computer Science.

Appendix A. Proof of Theorem 25

From the definition of $\hat{\boldsymbol{w}}$ as a minimizer we have

$$\|\Psi(\boldsymbol{w}_\star - \hat{\boldsymbol{w}})\|_2^2 + 2\lambda\|\hat{\boldsymbol{w}}\|_1 \leq 2\lambda\|\boldsymbol{w}_\star\|_1 + 2\varepsilon^\top\Psi(\hat{\boldsymbol{w}} - \boldsymbol{w}_\star). \tag{14}$$

By Hölder's inequality the assumptions of the theorem, $2\varepsilon^\top\Psi(\hat{\boldsymbol{w}} - \boldsymbol{w}_\star) \leq 2\|\varepsilon^\top\Psi\|_\infty\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_1 \leq \lambda\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_1$. Combining this with Eq. (14) gives

$$\|\Psi(\boldsymbol{w}_\star - \hat{\boldsymbol{w}})\|_2^2 \leq 2\lambda\|\boldsymbol{w}_\star\|_1 - 2\lambda\|\hat{\boldsymbol{w}}\|_1 + \lambda\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_1.$$

Adding $\lambda\|(\hat{\boldsymbol{w}} - \boldsymbol{w})\|_1$ to both sides we get

$$\begin{aligned}
\|\Psi(\boldsymbol{w}_\star - \hat{\boldsymbol{w}})\|_2^2 + \lambda\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_1 &\leq 2\lambda\Big(\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|_1 + \|\boldsymbol{w}_\star\|_1 - \|\hat{\boldsymbol{w}}\|_1\Big) \\
&= 2\lambda\sum_{j=1}^{d}\Big(|\hat{\boldsymbol{w}}[j] - \boldsymbol{w}_\star[j]| + |\boldsymbol{w}_\star[j]| - |\hat{\boldsymbol{w}}[j]|\Big) \\
&= 2\lambda\sum_{j\in\mathrm{supp}(\boldsymbol{w})}\Big(|\hat{\boldsymbol{w}}[j] - \boldsymbol{w}_\star[j]| + |\boldsymbol{w}_\star[j]| - |\hat{\boldsymbol{w}}[j]|\Big) \\
&\leq 4\lambda\sum_{j\in\mathrm{supp}(\boldsymbol{w})}|\hat{\boldsymbol{w}}[j] - \boldsymbol{w}_\star[j]| \\
&= 4\lambda\|[\hat{\boldsymbol{w}} - \boldsymbol{w}_\star]_{\mathrm{supp}(\boldsymbol{w})}\|_1.
\end{aligned}$$

It follows that

$$\|[\hat{\boldsymbol{w}} - \boldsymbol{w}_\star]_{\mathrm{supp}(\boldsymbol{w}_\star)^C}\|_1 \leq 3\|[\hat{\boldsymbol{w}} - \boldsymbol{w}_\star]_{\mathrm{supp}(\boldsymbol{w}_\star)}\|,$$

therefore $\hat{\boldsymbol{w}} - \boldsymbol{w}_\star \in E_s$. Denote $\boldsymbol{\delta} = \hat{\boldsymbol{w}} - \boldsymbol{w}$. The above derivation also implies

$$\|\Psi\boldsymbol{\delta}\|_2^2 \leq 3\lambda\|[\boldsymbol{\delta}]_{\mathrm{supp}(\boldsymbol{w}_\star)}\|_1 \leq 3\lambda\|\boldsymbol{\delta}_{[s]}\|_1 \leq 3\lambda\sqrt{s}\|\boldsymbol{\delta}_{[s]}\|_2.$$

Denote for brevity $\gamma = \gamma(\Psi, s)$. From the definition of γ,

$$\|\boldsymbol{\delta}_{[s]}\|_2^2 \leq \frac{1}{\gamma^2}\|\Psi\boldsymbol{\delta}\|_2^2 \leq \frac{3\lambda\sqrt{s}\|\boldsymbol{\delta}_{[s]}\|_2}{\gamma^2},$$

Therefore $\|\boldsymbol{\delta}_{[s]}\|_2 \leq \frac{3\lambda\sqrt{s}}{\gamma^2}$. Now,

$$\|\boldsymbol{\delta}\|_2 = \|\boldsymbol{\delta}_{[s]^C}\|_2 + \|\boldsymbol{\delta}_{[s]}\|_2 \leq \sqrt{\|\boldsymbol{\delta}_{[s]^C}\|_1 \|\boldsymbol{\delta}_{[s]^C}\|_\infty} + \|\boldsymbol{\delta}_{[s]}\|_2.$$

From $\boldsymbol{\delta} \in E_s$ we get $\|\boldsymbol{\delta}_{[s]^C}\|_1 \leq 3\|\boldsymbol{\delta}_{[s]}\|_1$. In addition, since $\boldsymbol{\delta}_{[s]}$ spans the largest coordinates of $\boldsymbol{\delta}$ in absolute value, $\|\boldsymbol{\delta}_{[s]^C}\|_\infty \leq \|\boldsymbol{\delta}_{[s]}\|_1/s$. Combining these with the inequality above we get

$$\|\boldsymbol{\delta}\|_2 \leq 3\|\boldsymbol{\delta}_{[s]}\|_1/\sqrt{s} + \|\boldsymbol{\delta}_{[s]}\|_2 \leq 4\|\boldsymbol{\delta}_{[s]}\|_2 \leq \frac{12\lambda\sqrt{s}}{\gamma^2}.$$

■

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 1999.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Stat.*, 39(5):2766–2794, 2011.

Alexandre Belloni and Victor Chernozhukov. l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *ArXiv e-prints*, June 2014.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59:7711–7717, 2013.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. H. Poincar Probab. Statist.*, 48(4):1148–1185, 2012.

A Chatterjee and SN Lahiri. Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259, 2013.

Samprit Chatterjee and Ali S Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Jianqing Fan, Yingying Fan, and Emre Barut. Adaptive robust variable selection. *arXiv preprint arXiv:1205.4795*, 2012.

Daniel Hsu and Sivan Sabato. Approximate loss minimization with heavy tails. *CoRR*, abs/1307.1827, 2013. URL `http://arxiv.org/abs/1307.1827`.

Daniel Hsu and Sivan Sabato. Heavy-tailed regression with a generalized median-of-means. In *Thirty-First International Conference on Machine Learning*, 2014.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

P. J. Huber. *Robust Statistics*. Wiley, 1981.

Anatoli Juditsky and Arkadii S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *ArXiv e-prints*, 0809.0813, 2008.

Matti Kääriäinen. Generalization error bounds using unlabeled data. In *Learning Theory*, pages 127–142. Springer, 2005.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

Casimir Kuratowski. Quelques problèmes concernant les espaces métriques non-séparables. *Fundamenta Mathematicae*, 25(1):534–545, 1935.

O. V. Lepski. Asymptotically minimax adaptive estimation I: Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36(4):682–697, 1991.

M. Lerasle and R. I. Oliveira. Robust empirical mean Estimators. *ArXiv e-prints*, December 2011.

Leonid A. Levin. Notes for miscellaneous lectures. *CoRR*, abs/cs/0503039, 2005.

Alexander E. Litvak, Alain Pajor, Mark Rudelson, and Nicole Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195(2):491–523, 2005. ISSN 0001-8708. doi: 10.1016/j.aim.2004.08.004. URL `http://dx.doi.org/10.1016/j.aim.2004.08.004`.

Mehrdad Mahdavi and Rong Jin. Passive learning with target risk. In *Twenty-Sixth Conference on Learning Theory*, 2013.

S. Mendelson. Learning without Concentration. *ArXiv e-prints*, January 2014.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *arXiv preprint arXiv:1308.1334*, 2013.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

M. Nussbaum. Minimax risk: Pinsker bound. In S. Kotz, editor, *Encyclopedia of Statistical Sciences, Update Volume 3*, pages 451–460. Wiley, New York, 1999.

Roberto Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15(19):203–212, 2010.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *arXiv preprint arXiv:1308.1147*, 2013.

O. Shamir. The Sample Complexity of Learning Linear Predictors with the Squared Loss. *ArXiv e-prints*, June 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, 2010.

N. Srivastava and R. Vershynin. Covariance estimation for distributions with $2+\epsilon$ moments. *Annals of Probability*, 41:3081–3111, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Sara van de Geer and Patric Müller. Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science*, 27(4):469–480, 2012.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

Lie Wang. L1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 2013.

J. Wolfowitz. Minimax estimates of the mean of a normal distribution with known variance. *The Annals of Mathematical Statistics*, 21:218–230, 1950.

Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, 19 (2):801, 2009.

Tong Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *arXiv preprint arXiv:1305.5029*, 2013.

Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126, 2008.