Heavy-tailed regression with a generalized median-of-means

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University

Sivan Sabato SIVAN.SABATO@MICROSOFT.COM
Microsoft Research New England, 1 Memorial Drive, Cambridge, MA 02446

Abstract

This work proposes a simple and computationally efficient estimator for linear regression, and other smooth and strongly convex loss minimization problems. We prove loss approximation guarantees that hold for general distributions, including those with heavy tails. All prior results only hold for estimators which either assume bounded or subgaussian distributions, require prior knowledge of distributional properties, or are not known to be computationally tractable. In the special case of linear regression with possibly heavy-tailed responses and with bounded and well-conditioned covariates in d-dimensions, we show that a random sample of size $\tilde{O}(d \log(1/\delta))$ suffices to obtain a constant factor approximation to the optimal loss with probability $1-\delta$, a minimax optimal sample complexity up to log factors. The core technique used in the proposed estimator is a new generalization of the median-of-means estimator to arbitrary metric spaces.

1. Introduction

Many standard methods for estimation and statistical learning are designed for optimal behavior in expectation, yet they may be suboptimal for high-probability guarantees. For instance, the population mean of a random variable can be estimated by the empirical mean, which is minimax-optimal with respect to the expected squared error. However, the deviations of this estimator from the true mean may be large with constant probability unless higher-order moments are controlled in some way, such as a subguassianity assumption (Catoni, 2012); similar issues arise in multivariate and high-dimensional estimation problems, such as linear regression and convex loss minimization. In many practical applications, distributions are *heavy-tailed* and thus are not subgaussian—they may not even have finite high-order moments. Thus, standard techniques such as empirical averages may be inappropriate, in spite of their optimality guarantees under restrictive assumptions.

A case in point is the classical problem of linear regression, where the goal is to estimate a linear function of a random vector $\boldsymbol{X}$ (the covariate) that predicts the response (label) Y with low mean squared error. The common approach for this problem is to use ordinary least squares or ridge regression, which minimize the loss on a finite labeled sample (with regularization in the case of ridge regression). The analyses of Srebro et al. (2010) and Hsu et al. (2012) for these estimators give sharp rates of convergence of the mean squared error of the resulting predictor to the optimal attainable loss, but only under assumptions of boundedness. Audibert & Catoni also analyze these estimators using PAC-Bayesian techniques, and manage to remove the boundedness assumptions, but they only provide asymptotic guarantees or guarantees which hold only if $n \geq \Omega(1/\delta)$. The failure of these estimators for general unbounded distributions may not be surprising given their inherent non-robustness to heavy-tailed distributions as discussed later in this work.

To overcome the issues raised above, we propose simple and computationally efficient estimators for linear regression and other convex loss minimization problems. The estimators have near-optimal approximation guarantees, even when the data distributions are heavy-tailed. Our estimator for the linear regression of a response Y on a d-dimensional covariate vector $\boldsymbol{X}$ converges to the optimal loss at an optimal rate with high probability, with only an assumption of bounded constant-order moments for $\boldsymbol{X}$ and Y (see Theorem 1). For comparison, the only previous result with a comparable guarantee is based on an estimator which requires prior knowledge about the response distribution and which is not known to be computationally tractable (Audibert & Catoni, 2011). Furthermore, in the

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

case where $\boldsymbol{X}$ is bounded and well-conditioned (but the distribution of Y may still be heavy-tailed), our estimator achieves, with probability $\geq 1-\delta$, a multiplicative constant approximation of the optimal squared loss, with a sample size of $n \geq O(d\log(d)\cdot\log(1/\delta))$ (see Theorem 2). This improves on the previous work of Mahdavi & Jin (2013), whose estimator, based on stochastic gradient descent, requires under the same conditions a sample size of $n \geq O(d^5\log(1/(\delta L_\star^{\mathrm{sq}})))$, where $L_\star^{\mathrm{sq}}$ is the optimal squared loss. We also prove an approximation guarantee in the case where $\boldsymbol{X}$ has a bounded distribution in an infinite-dimensional Hilbert space, as well as general results for other loss minimization problems with smooth and strongly-convex losses.

Our estimation technique is a new generalization of the median-of-means estimator used by Alon et al. (1999) and many others (see, for instance, Nemirovsky & Yudin, 1983, p. 243). The basic 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 with an appropriate criterion. If an estimator from one group is good with better-than-fair chance, then the selected estimator will be good with probability exponentially close to one. Our generalization provides a new simple selection criterion which yields the aforementioned improved guarantees. We believe that our new generalization of this basic technique will be applicable to many other problems with heavy-tailed distributions. Indeed, the full version of this paper (Hsu & Sabato, 2013) reports additional applications to sparse linear regression and low-rank matrix approximation. In an independent work, Minsker (2013) considers other variations of the original median-of-means estimator.

We begin by stating and discussing the main results for linear regression in Section 2. We then explain the core technique in Section 3. The application of the technique for smooth and convex losses is analyzed in Section 4. Section 5 provides the derivations of our main results for regression.

2. Main results

In this section we state our main results for linear regression, which are specializations of more general results given in Section 4. Unlike standard high-probability bounds for regression, the bounds below make no assumption on the range or the tails of the response distribution other than a trivial requirement that the optimal squared loss be finite. We give different bounds depending on conditions on the covariate distributions.

Let $[n] := \{1,2,\ldots,n\}$ for any natural number $n \in \mathbb{N}$. Let $\mathcal{Z}$ be a data space, $\mathbb{X}$ a parameter space, $\mathcal{D}$ a distribution over $\mathcal{Z}$, and Z 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 finite sample $T\subset\mathcal{Z}$ (where T is a multiset), $L_T(\boldsymbol{w}) := |T|^{-1}\sum_{z\in T}\ell(z,\boldsymbol{w})$. Let Id be the identity operator on $\mathbb{X}$, and $L_\star := \min_{\boldsymbol{w}} L(\boldsymbol{w})$. Set $\boldsymbol{w}_\star$ such that $L_\star = L(\boldsymbol{w}_\star)$.

For regression, we assume the parameter space $\mathbb{X}$ is a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathbb{X}}$, and $\mathcal{Z} := \mathbb{X}\times\mathbb{R}$. The loss is the squared loss $\ell = \ell^{\mathrm{sq}}$, 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 $\ell^\lambda((\boldsymbol{x},y),\boldsymbol{w}) := \frac{1}{2}(\langle\boldsymbol{x},\boldsymbol{w}\rangle_{\mathbb{X}}-y)^2 + \frac{1}{2}\lambda\langle\boldsymbol{w},\boldsymbol{w}\rangle_{\mathbb{X}}$; note that $\ell^0 = \ell^{\mathrm{sq}}$. We analogously define L^{sq}, L_T^{sq}, $L_\star^{\mathrm{sq}}$, L^λ, *etc.* as above.

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 $\boldsymbol{a}\mapsto m^{-1}\sum_{i=1}^m \boldsymbol{X}_i\langle\boldsymbol{X}_i,\boldsymbol{a}\rangle_{\mathbb{X}}$.

The proposed algorithm for regression (Algorithm 1) is as follows. 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 several independent estimations of the covariance matrix Σ from i.i.d. samples to select a single regressor from the k regressors at hand. The variant in Step 5 may be used to obtain tighter bounds in some cases discussed below.

Algorithm 1 Regression for heavy-tails

input $\lambda\geq 0$, sample sizes n, 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 $S_1,\ldots,S_k$ from D, each of size $\lfloor n/k\rfloor$.
3: For each $i\in[k]$, let $\boldsymbol{w}_i\in\operatorname{argmin}_{\boldsymbol{w}\in\mathbb{X}} L_{S_i}^\lambda(\boldsymbol{w})$.
4: Draw a random i.i.d sample T of size n', and split it to k samples $\{T_j\}_{j\in[k]}$ of equal size.
5: For each $i\in[k]$, let r_i be the median of the values in $\{\langle\boldsymbol{w}_i-\boldsymbol{w}_j,(\Sigma_{T_j}+\lambda\,\mathrm{Id})(\boldsymbol{w}_i-\boldsymbol{w}_j)\rangle \mid j\in[k]\setminus\{i\}\}$. [**Variant**: Use Σ_T instead of Σ_{T_j}.]
6: Set $i_\star := \arg\min_{i\in[k]} r_i$.
7: Return $\hat{\boldsymbol{w}} := \boldsymbol{w}_{i_\star}$.

First, consider the finite-dimensional case, where $\mathbb{X}=\mathbb{R}^d$, and assume Σ is not singular. 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.[1]

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

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

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

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

Theorem 1. *Assume Σ is not singular. If $\boldsymbol{X}$ satisfies Condition 1 with parameters c and η, then there is a constant $C = C(c, \eta)$ such that Algorithm 1 with $\lambda = 0$, $n \geq Cd\log(1/\delta)$, and $n' \geq C\log(1/\delta)$, with probability at least $1 - \delta$,*

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

Define the following finite fourth-moment conditions:

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

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 & Catoni (2011) proposes an estimator for linear regression based on optimization of a robust loss function (see also Catoni, 2012) which achieves essentially the same guarantee as Theorem 1 (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.

[1] As shown by Srivastava & 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.

Theorem 1 can be specialized for other 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 (and, here, we do not assume Condition 1). Then, the following result can be derived using Algorithm 1, with the variant of Step 5 for slightly tighter guarantees.

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

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

Note that $\mathbb{E}(\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{X}) = \mathbb{E}\operatorname{tr}(\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{X}) = \operatorname{tr}(\mathrm{Id}) = d$, 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 (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 $O(d\log(1/\delta))$.

In recent work of Mahdavi & 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 & 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 2 shows that $\tilde{O}(R^2\log(1/\delta))$ suffice when using our estimator.

It is interesting to note that here we achieve a constant factor approximation to $L_\star$ with a sample complexity that does not depend on the value of $L_\star$. This contrasts with other parametric learning settings, such as classification, where constant approximation requires $\Omega(1/L_\star)$ samples, and even active learning can only improve the dependence to $\Omega(\log(1/L_\star))$ (see, *e.g.*, Balcan et al., 2006).

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 3. *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 1 with $\lambda > 0$. With probability at least $1 - \delta$, as soon as $n \geq O((V^2/\lambda)\log(V/\sqrt{\lambda})\log(1/\delta))$ and $n' \geq O((V^2/\lambda)\log(V/(\delta\sqrt{\lambda}))$,*

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

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 $\lambda = \Theta(\sqrt{L_\star^{\mathrm{sq}} V^2 \log(1/\delta)/(B^2 n)})$ yields that as soon as $n \geq \tilde{O}(B^2 V^2 \log(1/\delta)/L_\star^{\mathrm{sq}})$ and $n' \geq \tilde{O}(B^2 V^2 \log(1/\delta)/L_\star^{\mathrm{sq}})$,

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

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

3. The core technique

In this section we present the core technique from which Algorithm 1 is derived. We first demonstrate the underlying principle via the *median-of-means* estimator, and then explain the generalization to arbitrary metric spaces.

3.1. Warm-up: median-of-means estimator

Algorithm 2 Median-of-means estimator

input Sample $S \subset \mathbb{R}$ of size n, number of groups $k \in \mathbb{N}$ which divides n.

output Population mean estimate $\hat{\mu} \in \mathbb{R}$.

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

We first motivate our procedure for approximate loss minimization by considering the special case of estimating a scalar population mean using a *median-of-means* estimator, given in Algorithm 2. This estimator, heavily used in the streaming algorithm literature (Alon et al., 1999, though a similar technique also appears in the textbook by Nemirovsky & Yudin, 1983 as noted by Levin, 2005), partitions a sample into k equal-size groups, and returns the median of the sample means of each group. The input parameter k is a constant determined by the desired confidence level (*i.e.*, $k = \log(1/\delta)$ for confidence $\delta \in (0, 1)$). The following result is well known.

Proposition 1. *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 divides n. With probability at least $1 - e^{-k/4.5}$, the estimate $\hat{\mu}$ returned by Algorithm 2 on input (S, k) satisfies $|\hat{\mu} - \mu| \leq \sigma\sqrt{6k/n}$.*

Proof. 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 $b_i := \mathbb{1}\{|\mu_i - \mu| \leq \sqrt{6\sigma^2 k/n}\}$. 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 $\{\sum_{i=1}^k b_i > k/2\}$, at least half of the μ_i are within $\sqrt{6\sigma^2 k/n}$ of μ, so the same holds for the median of the μ_i. □

Remark 1. It is remarkable that the estimator has $O(\sigma/\sqrt{n})$ convergence with exponential probability tails, even though the random variable x may have heavy-tails (*e.g.*, no bounded moments beyond the variance). Catoni (2012) also presents mean estimators with these properties *and also asymptotically optimal constants*, although the estimators require σ as a parameter.

Remark 2. Catoni (2012) shows that the empirical mean cannot provide a qualitatively similar guarantee: 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

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

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 generalization of the median-of-means estimator for arbitrary metric spaces, with a metric that can only be crudely estimated. 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). 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 $\boldsymbol{w}_0$.

The first abstraction captures the generation of candidate solutions obtained from independent subsamples. We assume there is an oracle $\mathrm{APPROX}_{\rho,\varepsilon}$ which, upon querying, returns a random $\boldsymbol{w} \in \mathbb{X}$ satisfying

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

We assume that the responses of $\mathrm{APPROX}_{\rho,\varepsilon}$ are generated independently. Note that the $2/3$ could be replaced by another constant larger than half; we have not made any attempt to optimize constants.

To second abstraction captures the limitations in calculating the metric. We assume there is an oracle DIST_ρ which, if queried with any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$, returns a random number $\mathrm{DIST}_\rho(\boldsymbol{x}, \boldsymbol{y})$ satisfying

$$\Pr\Big[\rho(\boldsymbol{x}, \boldsymbol{y})/2 \leq \mathrm{DIST}_\rho(\boldsymbol{x}, \boldsymbol{y}) \leq 2\rho(\boldsymbol{x}, \boldsymbol{y})\Big] \geq 8/9. \quad (4)$$

Algorithm 3 Robust approximation with random distances

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

1: For each $i \in [k]$, let $\boldsymbol{w}_i$ be the response from querying $\mathrm{APPROX}_{\rho,\varepsilon}$; set $W := \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$.
2: For each $i \in [k]$, let $r_i := \mathrm{median}\{\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) : j \in [k]\}$; set $i_\star := \arg\min_{i\in[k]} r_i$.
3: Return $\hat{\boldsymbol{w}} := \boldsymbol{w}_{i_\star}$.

We assume that the responses of DIST_ρ are generated independently (and independent of $\mathrm{APPROX}_{\rho,\varepsilon}$). Note that the responses need not correspond to a metric. Moreover, we will only query DIST_ρ for the pairwise distances of k fixed points (the candidate parameters $W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$), and it will suffice for the responses within each set $\{\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j)\}_{j\in[k]\setminus\{i\}}$ for any fixed i to be mutually independent.

The proposed procedure, given in Algorithm 3, generates k candidate solutions by querying $\mathrm{APPROX}_{\rho,\varepsilon}$ k times, and then selects a single candidate using a randomized generalization of the median. Specifically, for each $i \in [k]$, the radius of smallest ball centered at $\boldsymbol{w}_i$ that contains more than half of $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$ is approximated using calls to DIST_ρ; the $\boldsymbol{w}_i$ with the smallest such approximation is returned. Again, the number of candidates k determines the resulting confidence level. The following theorem provides a guarantee for Algorithm 3. The idea of the proof is illustrated in Figure 1. A similar technique was proposed by Nemirovsky & Yudin (1983), however their formulation relies on knowledge of ϵ and the metric.

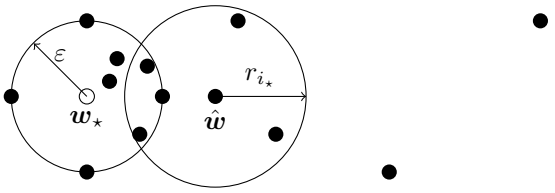

Figure 1. The main argument in the proof of Theorem 4, illustrated on the Euclidean plane. With probability at least $1-\delta$, at least $3k/5$ of the $\boldsymbol{w}_i$ (depicted by full circles) are within ε of $\boldsymbol{w}_\star$ (the empty circle). Therefore, with high probability, $\hat{\boldsymbol{w}}$ is within $\varepsilon + r_{i_\star} \le 9\varepsilon$ of $\boldsymbol{w}_\star$.

Theorem 4. *With probability at least* $1-(k+1)e^{-k/45}$, *Algorithm 3 returns* $\hat{\boldsymbol{w}} \in \mathbb{X}$ *satisfying* $\rho(\boldsymbol{w}_\star, \hat{\boldsymbol{w}}) \le 9\varepsilon$.

Proof. For each $i \in [k]$, let $b_i := \mathbb{1}\{\rho(\boldsymbol{w}_\star, \boldsymbol{w}_i) \le \varepsilon\}$. Note that the b_i are independent indicator random variables, each with $\mathbb{E}(b_i) \ge 2/3$. By Hoeffding's inequality, $\Pr[\sum_{i=1}^k b_i > 3k/5] \ge 1 - e^{-k/45}$. Henceforth condition on the event $\sum_{i=1}^k b_i > 3k/5$, *i.e.*, that more than $3/5$ of the $\boldsymbol{w}_i$ are contained in $B_\rho(\boldsymbol{w}_\star, \varepsilon)$.

Suppose $\boldsymbol{w}_i \in B_\rho(\boldsymbol{w}_\star, \varepsilon)$, and let $y_{i,j} := \mathbb{1}\{\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \le 4\varepsilon\}$. Observe that for every $\boldsymbol{w}_j \in B_\rho(\boldsymbol{w}_\star, \varepsilon)$, $\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \le 2\varepsilon$ by the triangle inequality, and thus

$$\Pr\Big[\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \le 4\varepsilon\Big] \ge \Pr\Big[\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \le 2\rho(\boldsymbol{w}_i, \boldsymbol{w}_j)\Big] \ge 8/9$$

for such $\boldsymbol{w}_j$, *i.e.*, $\mathbb{E}(y_{i,j}) \ge 8/9$. Therefore $\mathbb{E}(\sum_{j=1}^k y_{i,j}) \ge \sum_{j\in[k]:\boldsymbol{w}_j\in B_\rho(\boldsymbol{w}_\star,\varepsilon)} \mathbb{E}y_{i,j} \ge 8k/15 > k/2$. By Hoeffding's inequality, $\Pr[\sum_{i=1}^k y_{i,j} \le k/2] \le e^{-k/45}$. Thus, with probability at least $1-e^{-k/45}$, $r_i = \mathrm{median}\{\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) : j \in [k]\} \le 4\varepsilon$.

Now suppose $\boldsymbol{w}_i \notin B_\rho(\boldsymbol{w}_\star, 9\varepsilon)$. Let $z_{i,j} := \mathbb{1}\{\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) > 4\varepsilon\}$. Observe that for every $\boldsymbol{w}_j \in B_\rho(\boldsymbol{w}_\star, \varepsilon)$, $\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \ge \rho(\boldsymbol{w}_\star, \boldsymbol{w}_i) - \rho(\boldsymbol{w}_\star, \boldsymbol{w}_j) > 8\varepsilon$ by the triangle inequality, and thus

$$\Pr\Big[\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) > 4\varepsilon\Big] \ge \Pr\Big[\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) \ge (1/2)\rho(\boldsymbol{w}_i, \boldsymbol{w}_j)\Big] \ge 8/9$$

for such $\boldsymbol{w}_j$, *i.e.*, $\mathbb{E}(z_{i,j}) \ge 8/9$. Therefore, as before $\mathbb{E}(\sum_{j=1}^k z_{i,j}) \ge 8k/15 > k/2$. By Hoeffding's inequality, with probability at least $1-e^{-k/45}$, $r_i = \mathrm{median}\{\mathrm{DIST}_\rho(\boldsymbol{w}_i, \boldsymbol{w}_j) : j \in [k]\} > 4\varepsilon$.

Now take a union bound over the up to k events described above (at most one for each $\boldsymbol{w}_i \in W$) to conclude that with probability at least $1-(k+1)e^{-k/45}$, (i) $|W \cap B_\rho(\boldsymbol{w}_\star, \varepsilon)| \ge 3k/5 > 0$, (ii) $r_i \le 4\varepsilon$ for all $\boldsymbol{w}_i \in W \cap B_\rho(\boldsymbol{w}_\star, \varepsilon)$, and (iii) $r_i > 4\varepsilon$ 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 must satisfy $\boldsymbol{w}_i \in B_\rho(\boldsymbol{w}_\star, 9\varepsilon)$. □

4. Minimizing strongly convex losses

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

We employ the definitions for a general loss $\ell\colon \mathcal{Z} \times \mathbb{X} \to \mathbb{R}_+$ given in Section 2. 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 $\boldsymbol{w}_\star := \arg\min_{\boldsymbol{w}\in\mathbb{X}} L(\boldsymbol{w})$.[2]

Suppose $(\mathbb{X}, \|\cdot\|)$ is a Banach space. Denote by $\|\cdot\|_*$ the dual norm, so $\|\boldsymbol{y}\|_* = \sup\{\langle \boldsymbol{y}, \boldsymbol{x}\rangle\colon \boldsymbol{x} \in \mathbb{X}, \|\boldsymbol{x}\| \le 1\}$ for

[2] This holds, for instance, if L is strongly convex.

$\boldsymbol{y} \in \mathbb{X}^*$. Also, denote by $B_{\|\cdot\|}(\boldsymbol{c}, r) := \{\boldsymbol{x} \in \mathbb{X} \colon \|\boldsymbol{x}-\boldsymbol{c}\| \le r\}$ the ball of radius $r \ge 0$ around $\boldsymbol{c} \in \mathbb{X}$.

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

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

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

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

We say $\|\cdot\|$ is *γ-smooth* if $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|^2$ is γ-smooth with respect to $\|\cdot\|$.

Fix a norm $\|\cdot\|$ on $\mathbb{X}$ with a dual norm $\|\cdot\|_*$. The metric ρ used by Algorithm 3 is defined by $\rho(\boldsymbol{w}_1, \boldsymbol{w}_2) = \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$. We denote ρ by $\|\cdot\|$ as well. 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 equal-size groups $S_1, S_2, \ldots, S_k$, and let the response to the i-th query to $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ be the loss minimizer on S_i, *i.e.*, $\arg\min_{\boldsymbol{w} \in \mathbb{X}} L_{S_i}(\boldsymbol{w})$. We call this implementation *subsampled empirical loss minimization*. We further assume that there exists some sample size n_k that allows $\mathrm{DIST}_{\|\cdot\|}$ to be correctly implemented using any i.i.d. sample of size $n' \ge n_k$. Clearly, if S is an i.i.d. sample from $\mathcal{D}$, and $\mathrm{DIST}_{\|\cdot\|}$ is approximated using a separate sample, then the queries to $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ are independent from each other and from $\mathrm{DIST}_{\|\cdot\|}$. Thus, to apply Theorem 4, it suffices to show that Eq. (3) holds.

We assume $\|\cdot\|_*$ is γ-smooth for some $\gamma > 0$. Let n_α denote the smallest sample size such that the following holds: With probability $\ge 5/6$ over the choice of an i.i.d. sample T of size $|T| \ge n_\alpha$ from $\mathcal{D}$, for all $\boldsymbol{w} \in \mathbb{X}$,

$$L_T(\boldsymbol{w}) \ge L_T(\boldsymbol{w}_\star) + \langle \nabla L_T(\boldsymbol{w}_\star), \boldsymbol{w} - \boldsymbol{w}_\star \rangle + \frac{\alpha}{2}\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2. \tag{5}$$

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

The following lemma proves that Eq. (3) holds under these assumptions with

$$\varepsilon := 2\sqrt{\frac{6\gamma k \mathbb{E}\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2}{n\alpha^2}}. \tag{6}$$

Lemma 1. *Assume k divides n, and that S is an i.i.d. sample from $\mathcal{D}$ of size $n \ge k \cdot n_\alpha$. Then subsampled empirical loss minimization using the sample S is a correct implementation of $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$ for up to k queries.*

Proof. 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, since $\|\cdot\|$ is γ-smooth, $\mathbb{E}\|\nabla L_{S_i}(\boldsymbol{w}_\star)\|_*^2 \le \gamma(k/n)\mathbb{E}\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2$ (see Juditsky & Nemirovski, 2008). By Markov's inequality,

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

Moreover, the assumption that $n/k \ge n_\alpha$ implies that with probability at least $5/6$, Eq. (5) 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 $\boldsymbol{w}_i := \arg\min_{\boldsymbol{w} \in \mathbb{X}} L_{S_i}(\boldsymbol{w})$,

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

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 $\Pr[\|\boldsymbol{w}_i - \boldsymbol{w}_\star\| \le \varepsilon] \ge 2/3$. □

Combining Lemma 1 and Theorem 4 gives the following theorem.

Theorem 5. *Assume $k := C\lceil \log(1/\delta) \rceil$ (for some universal constant $C > 0$) divides n, S is an i.i.d. sample from $\mathcal{D}$ of size $n \ge k \cdot n_\alpha$, and S' is an i.i.d. sample from $\mathcal{D}$ of size $n' \ge n_k$. Further, assume Algorithm 3 uses the subsampled empirical loss minimization to implement $\mathrm{APPROX}_{\|\cdot\|,\varepsilon}$, where ε is as in Eq. (6), as well as implementation of $\mathrm{DIST}_{\|\cdot\|}$ using S'. Then with probability at least $1 - \delta$, the parameter $\hat{\boldsymbol{w}}$ returned by Algorithm 3 satisfies, (for some universal constant C)*

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

We give an easy corollary of Theorem 5 for the case where ℓ is smooth.

Corollary 1. *Assume the same conditions as Theorem 5, and also that: (i) $\boldsymbol{w} \mapsto \ell(z, \boldsymbol{w})$ is β-smooth with respect to $\|\cdot\|$ for all $z \in \mathcal{Z}$, and (ii) $\boldsymbol{w} \mapsto L(\boldsymbol{w})$ is $\bar{\beta}$-smooth with respect to $\|\cdot\|$. Then with probability at least $1 - \delta$, (for some universal constant $C > 0$)*

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

Proof. Due to the smoothness assumption on ℓ, $\|\nabla \ell(z, \boldsymbol{w}_\star)\|_*^2 \le 4\beta\ell(z, \boldsymbol{w}_\star)$ for all $z \in \mathcal{Z}$ (Srebro et al., 2010, Lemma 2.1). Thus, $\mathbb{E}[\|\nabla \ell(Z, \boldsymbol{w}_\star)\|_*^2] \le 4\beta L(\boldsymbol{w}_\star)$. The result follows using Theorem 5 and since $L(\hat{\boldsymbol{w}}) - L(\boldsymbol{w}_\star) \le \frac{\bar{\beta}}{2}\|\hat{\boldsymbol{w}} - \boldsymbol{w}_\star\|^2$, due to the strong smoothness of L and the optimality of $L(\boldsymbol{w}_\star)$. □

Corollary 1 implies that for smooth losses, Algorithm 3 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 $k = O(\log(1/\delta))$.

Remark 3. 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 3 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 2 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 1. 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.

5. Least squares linear regression

We now show how to apply our analysis for squared loss minimization using an appropriate norm and an upper bound on n_α. Assume $\mathbb{X}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{X}}$, and that L_T is twice-differentiable (which is the case for square 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

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

for any sample $T \subseteq \mathcal{Z}$. Therefore, to establish a bound on n_α, it suffices to find a size of T such that for an i.i.d. sample T from $\mathcal{D}$,

$$\Pr\left[\inf_{\boldsymbol{\delta} \in \mathbb{X} \setminus \{\boldsymbol{0}\}, \tilde{\boldsymbol{w}} \in \mathbb{R}^d} \frac{\langle \boldsymbol{\delta}, \nabla^2 L_T(\tilde{\boldsymbol{w}}) \boldsymbol{\delta} \rangle_{\mathbb{X}}}{\|\boldsymbol{\delta}\|^2} \geq \alpha\right] \geq 5/6. \quad (7)$$

For ease of exposition, we start with analysis for the case where Y is allowed to be heavy-tailed, but $\boldsymbol{X}$ is assumed to be light-tailed. The analysis is provided in Section 5.1 and Section 5.2. The analysis for the case where $\boldsymbol{X}$ can also be heavy tailed is provided in Section 5.3.

Recall that for a sample $T := \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m\}$ of m independent copies of a random vector $\boldsymbol{X} \in \mathbb{X}$, Σ_T is the empirical second-moment operator based on T. The following result bounds the spectral norm deviation of Σ_T from the population second moment operator Σ under a boundedness assumption on $\boldsymbol{X}$.

Lemma 2 (Specialization of Lemma 1 in 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 $\boldsymbol{a} \in \mathbb{X}$,*

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

We use the boundedness assumption on $\boldsymbol{X}$ 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, as in Litvak et al. 2005).

5.1. Finite-dimensional ordinary least squares

Consider first ordinary least squares in the finite-dimensional case. In this case $\mathbb{X} = \mathbb{R}^d$ and Algorithm 1 can be used with $\lambda = 0$. It is easy to see that Algorithm 1 is a specialization of Algorithm 3 with subsampled empirical loss minimization when $\ell = \ell^{\mathrm{sq}}$. We now prove Theorem 2. Recall that in this theorem we assume the variant of Algorithm 1, in which step 5 uses the covariance matrix of the entire T sample, Σ_T, instead of separate matrices $\Sigma_{T_{i,j}}$. Thus the norm we use in Algorithm 3 is $\|\cdot\|_T$, defined as $\|\boldsymbol{a}\|_T = \sqrt{\boldsymbol{a}^\top \Sigma_T \boldsymbol{a}}$, with the oracle $\mathrm{DIST}_{\|\cdot\|} = \mathrm{DIST}_{\|\cdot\|_T}$ that always provides the correct distance.

Proof of Theorem 2. The proof is derived from Corollary 1 as follows. First, it is easy to check that the dual of $\|\cdot\|_T$ is 1-smooth. Let the norm $\|\cdot\|_\Sigma$ be defined by $\|\boldsymbol{a}\|_\Sigma = \sqrt{\boldsymbol{a}^\top \Sigma \boldsymbol{a}}$. By Lemma 2, if $n' \geq O(R^2 \log(R/\delta))$, with probability at least $1-\delta$, $(1/2)\|a\|_\Sigma^2 \leq \|a\|_T^2 \leq 2\|a\|_\Sigma^2$ for all $\boldsymbol{a} \in \mathbb{R}^d$. Denote this event $\mathcal{E}$ and assume for the rest of the proof that $\mathcal{E}$ occurs. Since ℓ^{sq} is R^2-smooth with respect to $\|\cdot\|_\Sigma$, and L^{sq} is 1-smooth with respect to $\|\cdot\|_\Sigma$, the same holds, up to constant factors, for $\|\cdot\|_T$. Moreover, for any sample S,

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

By Lemma 2 with $\lambda = 0$, if $|S| \geq 80R^2 \log(24|S|^2)$ then with probability at least $5/6$, $\forall \boldsymbol{\delta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, $\boldsymbol{\delta}^\top \Sigma_S \boldsymbol{\delta} / \boldsymbol{\delta}^\top \Sigma \boldsymbol{\delta} \geq 1/2$. Therefore Eq. (7) holds for $\|\cdot\|_T$ with $\alpha = 1/4$ and $n_{1/4} = O(R^2 \log R)$. We can thus apply Corollary 1 with $\alpha = 1/4$, $\beta = 4R^2$, $\bar{\beta} = 4$, $\gamma = 1$, and $n_{1/4} = O(R^2 \log R)$, so with probability at least $1 - \delta$, the parameter $\hat{\boldsymbol{w}}$ returned by Algorithm 1 (with the variant) satisfies

$$L(\hat{\boldsymbol{w}}) \leq \left(1 + O\left(\frac{R^2 \log(1/\delta)}{n}\right)\right) L(\boldsymbol{w}_\star), \tag{8}$$

as soon as $n \geq O(R^2 \log(R) \log(1/\delta))$. A union bound over the probability that $\mathcal{E}$ also occurs finishes the proof. □

5.2. Ridge regression

In a general, possibly infinite-dimensional, Hilbert space $\mathbb{X}$, the variant of Algorithm 1 can be used with $\lambda > 0$. In this case, the algorithm is a specialization of Algorithm 3 with subsampled empirical loss minimization when $\ell = \ell^\lambda$, with the norm defined by $\|\boldsymbol{a}\|_{T,\lambda} = \sqrt{\boldsymbol{a}^\top (\Sigma_T + \lambda \operatorname{Id}) \boldsymbol{a}}$.

Proof of Theorem 3. First, it is easy to check that the dual of $\|\cdot\|_{T,\lambda}$ is 1-smooth. As in the proof of Theorem 2, by Lemma 2 if $n' \geq O((V^2/\lambda) \log(V/(\delta\sqrt{\lambda})))$ then with probability $1 - \delta$ the norm $\|\boldsymbol{a}\|_{T,\lambda}$ is equivalent to the norm $\|\cdot\|_{\Sigma,\lambda} = \sqrt{\boldsymbol{a}^\top (\Sigma + \lambda \operatorname{Id}) \boldsymbol{a}}$ up to constant factors. 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 $\|\cdot\|_{\Sigma,\lambda}$. In addition, L^λ is 1-smooth with respect to $\|\cdot\|_{\Sigma,\lambda}$. Using Lemma 2 with $r_\lambda = V/\lambda$, we have, similarly to the proof of Theorem 2, $n_{1/4} = O((V^2/\lambda) \log(V/\sqrt{\lambda}))$. Setting $\alpha = 1/4$, $\beta = 4(1 + V^2/\lambda)$, $\bar{\beta} = 4$, $\gamma = 1$, and $n_{1/4}$ as above, to match the actual norm $\|\cdot\|_{T,\lambda}$, we have 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 $n \geq O((V^2/\lambda) \log(V/\sqrt{\lambda}) \log(1/\delta))$.

We are generally interested in comparing to the minimum square loss $L^{\mathrm{sq}}_\star := \inf_{\boldsymbol{w} \in \mathbb{X}} L^{\mathrm{sq}}(\boldsymbol{w})$, 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 $\lambda = \Theta(\sqrt{L^{\mathrm{sq}}_\star V^2 \log(1/\delta)/(B^2 n)})$ yields

$$L^{\mathrm{sq}}(\hat{\boldsymbol{w}}) + \lambda \langle \hat{\boldsymbol{w}}, \hat{\boldsymbol{w}} \rangle_{\mathbb{X}} \leq L^{\mathrm{sq}}_\star + O\left(\sqrt{\frac{L^{\mathrm{sq}}_\star B^2 V^2 \log(1/\delta)}{n}} + \frac{(L^{\mathrm{sq}}_\star + B^2 V^2) \log(1/\delta)}{n}\right)$$

as soon as $n \geq \tilde{O}(B^2 V^2 \log(1/\delta)/L^{\mathrm{sq}}_\star)$. □

5.3. Heavy-tailed covariates

In this section we prove Theorem 1. When the regression covariates are not bounded or subgaussian as in the two previous sections, the empirical second-moment matrix may deviate significantly from its population counterpart with non-negligible probability. In this case we use Algorithm 1 with the original step 5 so that for any $i \in [k]$, the responses $\{\mathrm{DIST}_{\|\cdot\|}(\boldsymbol{w}_i, \boldsymbol{w}_j)\}_{j \in [k] \setminus \{i\}}$ are mutually independent.

For simplicity, we work in finite-dimensional Euclidean space $\mathbb{X} := \mathbb{R}^d$ and consider $\lambda = 0$. The analysis shows that Algorithm 1 is an instance of subsampled empirical loss minimization for ℓ^{sq} with the norm $\|\boldsymbol{a}\|_\Sigma = \sqrt{\boldsymbol{a}^\top \Sigma \boldsymbol{a}}$. Recall that we assume Condition 1 given in Section 2. The following lemma shows that under this condition, $O(d)$ samples suffice so that the expected spectral norm distance between the empirical second-moment matrix and Σ is bounded.

Lemma 3 (Corollary 1.2 from Srivastava & Vershynin 2013, essentially)**.** *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 fixed $\eta, c > 0$, there is a constant θ, such that for any $\epsilon \in (0, 1)$, if $n \geq \theta \epsilon^{-2-2/\eta} d$, then $\mathbb{E} \|\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - \operatorname{Id}\|_2 \leq \epsilon$.*

Lemma 3 implies that for the norm $\|\cdot\|_\Sigma$, $n_{1/2} = O(c'_\eta d)$ where $c'_\eta = \theta \cdot 2^{O(1+1/\eta)}$. Therefore, for $k = O(\log(1/\delta))$, subsampled empirical loss minimization requires $n \geq k \cdot n_{1/2} = O(c'_\eta d \log(1/\delta))$ samples to correctly implement $\mathrm{APPROX}_{\|\cdot\|_\Sigma, \varepsilon}$ for ε as in Eq. (6).

Step 5 in Algorithm 1 implements $\mathrm{DIST}_{\|\cdot\|_\Sigma}$ such that for every i, $\{\mathrm{DIST}_{\|\cdot\|_\Sigma}(\boldsymbol{w}_i, \boldsymbol{w}_j)\}_{j \in [k] \setminus \{i\}}$ are estimated using independent samples T_j. We now need to show that this implementation satisfies Eq. (4). By Lemma 3, for every $i, j \in [k]$ an i.i.d. sample T_j of size $O(c'_\eta)$ suffices so that with probability at least $8/9$,

$$\begin{aligned}(1/2)\|\Sigma^{1/2}(\boldsymbol{w}_i - \boldsymbol{w}_j)\|_2 &\leq \|\Sigma^{1/2}_{T_j}(\boldsymbol{w}_i - \boldsymbol{w}_j)\|_2 \\ &\leq 2\|\Sigma^{1/2}(\boldsymbol{w}_i - \boldsymbol{w}_j)\|_2.\end{aligned}$$

Thus for $k = O(\log(1/\delta))$, the total size of the sample T in Algorithm 1 needs to be $n' = O(c'_\eta \log(1/\delta))$. Setting $\alpha = 1/2$, $\gamma = 1$ and $n_\alpha = O(c'_\eta d)$, Theorem 1 is now derived from Theorem 5, by applying the identity

$$\|\nabla \ell^{\mathrm{sq}}((\boldsymbol{X}, Y), \boldsymbol{w}_\star)\|_{\Sigma,*} = 2\|\Sigma^{-1/2} \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_\star - Y)\|_2.$$

References

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

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

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Twenty-Third International Conference on Machine Learning*, 2006.

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

Hsu, Daniel and Sabato, Sivan. Approximate loss minimization with heavy tails. *ArXiv e-prints*, arXiv:1307.1827, 2013. Arxiv preprint.

Hsu, Daniel, Kakade, Sham M., and Zhang, Tong. Random design analysis of ridge regression. In *Twenty-Fifth Conference on Learning Theory*, 2012.

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

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

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

Litvak, Alexander E., Pajor, Alain, Rudelson, Mark, and Tomczak-Jaegermann, Nicole. 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.

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

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

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

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

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

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

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