Proceedings of Machine Learning Research vol 99:1–15, 2019 32nd Annual Conference on Learning Theory

On Mean Estimation for General Norms with Statistical Queries

Jerry Li JERRL@MICROSOFT.COM
Microsoft Research

Aleksandar Nikolov ANIKOLOV@CS.TORONTO.EDU
University of Toronto

Ilya Razenshteyn ILYARAZ@MICROSOFT.COM
Microsoft Research

Erik Waingarten EAW@CS.COLUMBIA.EDU
Columbia University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the problem of mean estimation for high-dimensional distributions given access to a statistical query oracle. For a normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ and a distribution supported on vectors $x \in \mathbb{R}^d$ with $\|x\|_X \leq 1$, the task is to output an estimate $\widehat{\mu} \in \mathbb{R}^d$ which is ε-close in the distance induced by $\|\cdot\|_X$ to the true mean of the distribution. We obtain sharp upper and lower bounds for the statistical query complexity of this problem when the the underlying norm is *symmetric* as well as for Schatten-p norms, answering two questions raised by Feldman, Guzmán, and Vempala (SODA 2017).

Keywords: statistical queries, mean estimation, normed spaces

1. Introduction

Let D be a distribution over $\mathbb{R}^d$. Informally speaking, in the statistical query model (SQ), one learns about D as follows. Given a query $h\colon \mathbb{R}^d \to [-1; 1]$, the SQ oracle with tolerance $\tau > 0$ reports $\mathbf{E}_{x\sim D}[h(x)]$ perturbed by error of scale roughly τ. The SQ model was introduced in Kearns (1998) as a way to capture "learning algorithms that construct a hypothesis based on statistical properties of large samples rather than on the idiosyncrasies of a particular sample."

The original motivation for the SQ framework was to provide an evidence of computational hardness of various learning problems (beyond sample complexity) by proving lower bounds on their SQ complexity. Indeed, many learning algorithms (see Feldman (2016b) for an overview) can be captured by the SQ framework, and, furthermore, the only known technique that gives a polynomial-time algorithm for a learning problem with exponential SQ complexity Kearns (1998) is *Gaussian elimination over finite fields*, whose utility for learning is currently extremely limited. This reasoning suggests the following heuristic:

If solving a learning problem to accuracy $\varepsilon > 0$ requires $d^{\omega(1)}$ SQ queries with tolerance $\varepsilon^{O(1)}/d^{O(1)}$, then it is unlikely to be doable in time $d^{O(1)}$ using any *algorithm.*

This heuristic together with the respective SQ lower bounds provided strong evidence of hardness of many problems such as: learning parity with noise Kearns (1998), learning intersection of half-spaces Klivans and Sherstov (2007), the planted clique problem Feldman et al. (2013b), robust estimation of high-dimensional Gaussians and non-Gaussian component analysis Diakonikolas et al. (2017), learning a small neural network Song et al. (2017), adversarial learning Bubeck et al. (2018), robust linear regression Diakonikolas et al. (2019), among others.

However, over time, the SQ model has generated significant *intrinsic* interest Feldman (2016a), in part due to the connections to distributed learning Steinhardt et al. (2016) and local differential privacy Kasiviswanathan et al. (2011). In particular, the new goal is to understand the trade-off between the number and the tolerance of SQ queries, and the accuracy of the resulting solution for various learning problems, which is *more nuanced* than what is necessary for the above "crude" heuristic. In a paper by Feldman, Guzman, and Vempala Feldman et al. (2017), this was done for perhaps the most basic learning problem, *mean estimation*, which is formulated as follows.

Problem 1 (Mean estimation using statistical queries) *Let D be a distribution over the* unit ball *B_X of a normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$, and suppose we are allowed $d^{O(1)}$ statistical queries with tolerance $\varepsilon > 0$. What is the smallest $\varepsilon' > 0$, for which we can always recover a point $\widehat{x}$ such that $\|\widehat{x} - \mathbf{E}_{\boldsymbol{x}\sim D}[\boldsymbol{x}]\| \leq \varepsilon'$ holds with high probability over the randomness of the estimation algorithm.*

Clearly, $\varepsilon' \geq \varepsilon$, and, as Feldman et al. (2017) showed, $\varepsilon' \leq O(\varepsilon\sqrt{d})$ for every norm. The goal of this paper is to understand which properties of the underlying normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ admit efficient SQ algorithms with ε'/ε as small as possible. Thus, we will refer to a norm $\|\cdot\|$ over $\mathbb{R}^d$ as *tractable* if one can achieve $\varepsilon' \leq \varepsilon \cdot \mathrm{poly}(\log d, \log(1/\varepsilon))$ (with $\mathrm{poly}(d)$ queries of tolerance ε). The main result of Feldman et al. (2017) can be stated as follows.

Theorem 1 (Feldman et al. (2017)) *The ℓ_p norm over $\mathbb{R}^d$ is tractable if and only if $p \geq 2$.*

The fact that the ℓ_∞ norm is tractable is trivial, since we can estimate each coordinate of the mean separately. However, the corresponding algorithm for ℓ_p norms for $2 \leq p < \infty$ is more delicate and is based on random rotations, while the naïve coordinate-by-coordinate estimator merely gives $\varepsilon' = \varepsilon d^{\Theta_p(1)}$. Feldman et al. (2017) raise several intriguing open problems, among them the following two: **(1)** Characterize tractable norms beyond ℓ_p; **(2)** Solve Problem 1 for the spectral norm and other Schatten-p norms of matrices. In this paper, we make progress towards solving the first problem and completely resolve the second one.

1.1. Our results

Symmetric norms. Our first result gives a complete characterization of *symmetric* tractable norms. A norm is symmetric if it is invariant under all permutations of coordinates and sign flips (for many examples beyond ℓ_p norms, see Andoni et al. (2017)). Recently there has been substantial progress in understanding various algorithmic tasks for general symmetric norms Błasiok et al. (2017); Andoni et al. (2017); Song et al. (2018); Andoni et al. (2018). In this paper, we significantly extend Theorem 1 to all the symmetric norms. To formulate our result, we need to define the *type*-2 *constant* of a normed space, which is one of the standard bi-Lipschitz invariants (Wojtaszczyk (1996)).

Definition 2 *For a normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$, the type-2 constant of X, denoted by $T_2(X)$, is defined as the smallest $T > 0$ such that the following holds. For every sequence of vectors $x_1, x_2, \ldots, x_n \in X$ and for uniformly random $\boldsymbol{\varepsilon} \sim \{-1,1\}^n$, one has:*

$$\left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{-1,1\}^n} \left\|\sum_{i=1}^n \boldsymbol{\varepsilon}_i x_i\right\|_X^2\right)^{1/2} \leq T \cdot \left(\sum_{i=1}^n \|x_i\|_X^2\right)^{1/2}. \tag{1}$$

We are now ready to state our result.

Theorem 3 *A symmetric normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ is tractable iff $T_2(X) \leq \mathrm{poly}(\log d)$.*

Theorem 3 easily implies Theorem 1, since for $1 \leq p < 2$, $T_2(\ell_p) = d^{\Omega_p(1)}$, while for $2 \leq p < \infty$ one has $T_2(\ell_p) \leq \sqrt{p-1}$ and $T_2(\ell_\infty) \leq O(\sqrt{\log d})$ (Ball et al. (1994)). For a quantitative version of Theorem 3, see Theorem 12 and Theorem 13. We note in passing that type-p is also known to characterize the performance of the empirical mean estimator for infinite-dimensional Banach spaces, see Chapter 9 of Ledoux and Talagrand (2011).

Schatten-p norms. Recall that for a matrix M, the Schatten-p norm of M is the ℓ_p norm of the singular values of M. In particular, the Schatten-∞ norm of M is simply the spectral norm of M, and the Schatten-2 norm corresponds to the Frobenius norm. Such norms are very well-studied and arise naturally in many applications in learning and probability theory. Our second main result settles the tractability of Schatten-p norms, resolving a question of Feldman et al. (2017).

Theorem 4 *The Schatten-p norm is tractable iff $p = 2$.*

For a quantitative version of Theorem 4, see Theorem 20. Theorem 4 shows that one cannot remove "symmetric" from Theorem 3, since type-2 constants of Schatten-p spaces are essentially the same as for the corresponding ℓ_p spaces (Ball et al. (1994)). Specifically, for $p > 2$, Schatten-p spaces have small type-2 constant, but are intractable. In particular, we show that the best mean estimation algorithm for Schatten-p can be obtained by embedding the space into ℓ_2 (via the identity map) and then using the ℓ_2 estimation algorithm from Feldman et al. (2017).

1.2. Techniques

The main technical tool underlying the algorithm for mean estimation in symmetric norms is the following geometric statement. For any symmetric norm $(\mathbb{R}^d, \|\cdot\|_X)$, consider the set $R_j \subset B_X$ consisting of the *level-j ring*, i.e., all points $x \in B_X$ whose non-zero coordinates have absolute value between $2^{-(j+1)}$ and 2^{-j}, and consider the smallest radius $r > 0$ where $R_j \subset rB_{\ell_2}$. Then,

$$R_j \subset rB_{\ell_2} \cap 2^{-j} B_{\ell_\infty} \subset (5T_2(X)\log_2 d) B_X. \tag{2}$$

Given the above geometric statement, which generalizes the similar statement for ℓ_p norms from Feldman et al. (2017), we generalize the algorithm from Feldman et al. (2017) to the symmetric norms setting. The resulting algorithm partitions vectors into levels and uses ℓ_2 and ℓ_∞ subroutines from Feldman et al. (2017).

The lower bound for norms with large type-2 constants is a generalization of the result in Feldman et al. (2017); in particular, the hard distributions for ℓ_p from Feldman et al. (2017) are supported on basis vectors, which are exactly those achieving $T_2(\ell_p)$ in (1). For general norms X, we consider the analogous distributions supported on an arbitrary set of vectors achieving $T_2(X)$ in (1); however, the fact that we have much less control on the vectors necessitates additional care.

The Schatten-p norms, for $p > 2$, do satisfy $T_2(S_p) \leq \sqrt{\log d}$, so new ideas are required in proving the lower bound. We show the lower bound for carefully crafted hard distributions, using hypercontractivity to show concentration of the result of an arbitrary statistical query.

2. Preliminaries

Here we introduce some basic notions about normed spaces and statistical algorithms. We will use boldfaced letters for random variables, and the notation $\boldsymbol{\varepsilon} \sim \{-1,1\}^n$ will mean that $\boldsymbol{\varepsilon}$ is a random vector chosen uniformly from $\{-1,1\}^n$.

Definition 5 *For any vector $x \in \mathbb{R}^d$, we let $|x|$ be the vector x with each coordinate replaced by its absolute value, and let $x^* = P|x|$ be the vector obtained by applying the permutation matrix P to $|x|$ which sorts coordinates of $|x|$ by order of non-increasing value. A normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ is symmetric if $\|x\|_X = \|x^*\|_X$ holds for every $x \in \mathbb{R}^d$.*

We recall that ℓ_p^d is the normed space over $\mathbb{R}^d$ with the norm of a vector x given by $\|x\|_p = (|x_1|^p + \ldots + |x_d|^p)^{1/p}$. The Schatten-$p$ space $S_p^d = (\mathbb{R}^{d^2}, \|\cdot\|_{S_p})$ is defined over $d \times d$ matrices with real entries, and the norm of a matrix is defined as the ℓ_p^d norm of its singular values. We omit the superscript d and just write ℓ_p and S_p when this does not cause confusion.

For a normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$, let $B_X = \{x \in \mathbb{R}^d : \|x\|_X \le 1\}$ be the unit ball of the norm X. Furthermore, for $p \in [1, \infty)$, we let $L_p(X) = (\mathbb{R}^{dn}, \|\cdot\|_{L_p(X)})$ be the normed space over sequences of vectors $x = (x_1, \ldots, x_n) \in \mathbb{R}^{dn}$ where $\|x\|_{L_p(X)} = (\sum_{i=1}^n \|x_i\|_X^p)^{1/p}$.

Next we define the type of a normed space.

Definition 6 *Let $X = (\mathbb{R}^d, \|\cdot\|_X)$ be a normed space, $n \in \mathbb{N}$, and $p \in [1,2]$. Let $T_p(X,n)$ be the infimum over $T > 0$ such that:*

$$\left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{-1,1\}^n} \left[\left\| \sum_{i=1}^n \boldsymbol{\varepsilon}_i x_i \right\|_X^2 \right] \right)^{1/2} \le T \left(\sum_{i=1}^n \|x_i\|_X^p \right)^{1/p},$$

for all $x_1, \ldots, x_n \in \mathbb{R}^d$. We let $T_p(X) = \sup_{n \in \mathbb{N}} T_p(X,n)$, and say X has type *p with constant $T_p(X)$.*

Note that, by the parallelogram identity, the Euclidean space $(\mathbb{R}^d, \|\cdot\|_2)$ has type 2 with constant 1, and in fact the inequality becomes an equality. Together with John's theorem (John (1948)), this implies that any d-dimensional normed space has type 2 with constant at most $\sqrt{d}$. However, we are typically interested in spaces that have type p with constant independent of dimension. It follows from the results in Ball et al. (1994) that for $p \ge 2$, ℓ_p^d has type 2 with constant $\sqrt{p-1}$, and for $1 \le p < 2$, ℓ_p^d has type p with constant 1; at the same time, considering the standard basis of $\mathbb{R}^d$ shows that for $1 \le p < q \le 2$, the type q constant of ℓ_p^d goes to infinity with the dimension d. Moreover, these results also hold for Schatten-p spaces.

Finally, we define formally statistical algorithms and the STAT and VSTAT oracles. We follow the definitions from Feldman et al. (2013a).

Definition 7 *Let D be a distribution supported on Ω. For a tolerance parameter $\tau > 0$, the oracle $\mathrm{STAT}(\tau)$ takes a query function $h \colon \Omega \to [-1,1]$, and returns some value $v \in \mathbb{R}$ satisfying $|v - \mathbf{E}_{\boldsymbol{x} \sim D}[h(\boldsymbol{x})]| \le \tau$. For a sample size parameter $t > 0$, the $\mathrm{VSTAT}(t)$ oracle takes a query function $h \colon \Omega \to [0,1]$ and returns some value $v \in \mathbb{R}$ such that $|v - p| \le \tau$, for $p = \mathbf{E}_{\boldsymbol{x} \sim D}[h(\boldsymbol{x})]$, and $\tau = \max\{1/t, \sqrt{p(1-p)/t}\}$.*

We call an algorithm that accesses the distribution D only via one of the above oracles a statistical algorithm.

Clearly, $\mathrm{VSTAT}(t)$ is at least as strong as $\mathrm{STAT}(1/2\sqrt{t})$ and no stronger than $\mathrm{STAT}(1/t)$. The lower bounds presented will follow the framework of Feldman et al. (2018).

Definition 8 *The discrimination norm $\kappa_2(D, \mathcal{D})$ for a distribution D supported on Ω and a set $\mathcal{D}$ of distributions supported on Ω is given by:*

$$\kappa_2(D, \mathcal{D}) = \max_{\substack{h\colon \Omega \to \mathbb{R} \\ \|h\|_D = 1}} \left\{ \mathop{\mathbf{E}}_{\mathbf{D} \sim \mathcal{D}} \left[\left| \mathop{\mathbf{E}}_{\boldsymbol{x} \sim D}[h(\boldsymbol{x})] - \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathbf{D}}[h(\boldsymbol{x})] \right| \right] \right\},$$

where $\mathbf{D} \sim \mathcal{D}$ is sampled uniformly at random, and $\|h\|_D^2 = \mathbf{E}_{\boldsymbol{y} \sim D}[h(\boldsymbol{y})^2]$. The decision problem $\mathcal{B}(D, \mathcal{D})$ is the problem of distinguishing whether an unknown distribution $\mathbf{H} = D$ or is sampled uniformly from $\mathcal{D}$. The statistical dimension with discrimination norm κ, $\mathrm{SDN}(\mathcal{B}(D, \mathcal{D}), \kappa)$, is the largest integer t such that for a finite subset $\mathcal{D}_D \subset \mathcal{D}$, any subset $\mathcal{D}' \subset \mathcal{D}_D$ of size at least $|\mathcal{D}_D|/t$ satisfies $\kappa_2(\mathcal{D}', D) \leq \kappa$.

Theorem 9 (Theorem 7.1 in Feldman et al. (2018)) *For $\kappa > 0$, let $t = \mathrm{SDN}(\mathcal{B}(D, \mathcal{D}), \kappa)$ for a distribution D and set of distributions $\mathcal{D}$ supported on a domain Ω. Any randomized statistical algorithm that solves $\mathcal{B}(D, \mathcal{D})$ with probability at least $2/3$ requires $t/3$ calls to $\mathrm{VSTAT}(1/(3\kappa^2))$.*

3. Symmetric norms

3.1. Mean estimation using SQ for type-2 symmetric norms

Definition 10 *Let $X = (\mathbb{R}^d, \|\cdot\|_X)$ be any symmetric norm with $\|e_1\|_X = 1$. Let $\ell_X\colon (0, 1] \to \{0, 1, \ldots, d\}$ be the maximum number of nonzero coordinates set to t in a vector within the unit ball of X, i.e.,*

$$\ell_X(t) = \max \left\{ k : \|(\underbrace{t, \ldots, t}_{k}, 0, \ldots, 0)\|_X \leq 1 \right\},$$

and $m_X\colon (0, 1] \to \mathbb{R}^{\geq 0}$ be the maximum ℓ_2 norm of a vector within the unit ball of X with nonzero coordinates set to t, i.e.,

$$m_X(t) = \max \left\{ \|x\|_2 : x = (t, \ldots, t, 0, \ldots, 0) \in B_X \right\}.$$

The following is the main lemma needed for the statistical query algorithm for type-2 symmetric norms. The lemma is a generalization of Lemma 3.12 from Feldman et al. (2017) from ℓ_p norms (with $p > 2$) to arbitrary type-2 symmetric norms. The lemma bounds the norm in X of an arbitrary vector x, given corresponding bounds on the $\|x\|_\infty$ and $\|x\|_2$.

Lemma 11 *Let $d \in \mathbb{N}$ be large enough, and $X = (\mathbb{R}^d, \|\cdot\|_X)$ be a symmetric norm with type-2 constant $T_2(X) \in [1, \infty)$. Fix any $t \in (0, 1]$, and let $x \in \mathbb{R}^d$ satisfy $\|x\|_\infty \leq t$ and $\|x\|_2 \leq m_X(t)$. Then, $\|x\|_X \leq T_2(X) \cdot 5 \log_2 d$.*

Proof Given the vector $x \in \mathbb{R}^d$, consider the sets $B_j(x) \subset [d]$ for $j \in \{0, \ldots, 2\log_2(d)\}$ given by

$$B_j(x) = \left\{ i \in [d] : t \cdot 2^{-j-1} < |x_i| \leq t \cdot 2^{-j} \right\},$$

and let $x^{(j)} \in \mathbb{R}^d$ be the vector given by letting the first $|B_j(x)|$ coordinates be $t \cdot 2^{-j}$, and the remaining coordinates be 0. Because X is symmetric with respect to changing the sign of any coordinate of x, the triangle inequality easily implies that $\|x\|_X$ is monotone with respect to $|x_i|$ for any $i \in [d]$. Then, by the triangle inequality and the fact that X is symmetric with $\|e_1\|_X = 1$, $\|x\|_X \le \sum_{j=0}^{2\log_2 d} \|x^{(j)}\|_X + t/d$; thus, it remains to bound $\|x^{(j)}\|_X$ for every $j \in \{0, \dots, 2\log_2(d)\}$.

We then have $\sqrt{|B_j(x)|} \cdot t \cdot 2^{-j} = \|x^{(j)}\|_2 \le 2m_X(t) \le 2t\sqrt{\ell_X(t)}$, where, in the first inequality, we used the fact that $\|x^{(j)}\|_2 \le 2\|x\|_2 \le 2m_X(t)$, and, in the second inequality, we used the definition of $\ell_X(t)$. As a result, we have $|B_j(x)| \le 4\ell_X(t) \cdot 2^{2j}$. Consider partitioning the non-zero coordinates of $x^{(j)}$ into at most $s = 4 \cdot 2^{2j}$ groups, each of size at most $\ell_X(t)$, and let $v_1, \dots, v_s \in \mathbb{R}^d$ be the coordinate projections of $x^{(j)}$ onto each respective part, so that $x^{(j)} = \sum_{i=1}^{s} v_s$. We have

$$\|x^{(j)}\|_X^2 = \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{-1,1\}^s} \left[\left\| \sum_{i=1}^{s} \boldsymbol{\varepsilon}_i v_i \right\|_X^2 \right] \overset{(a)}{\le} T_2(X)^2 \sum_{i=1}^{s} \|v_i\|_X^2 \overset{(b)}{\le} 4T_2(X)^2,$$

where the equality uses the symmetry of X with respect to changing signs of coordinates, (a) uses the definition of type constants, and (b) follows from the definition of $\ell_X(t)$. We obtain the desired lemma by summing over all $\|x^{(j)}\|_X$, for $j \in \{0, \dots, 2\log_2(d)\}$. ■

With this structural result, we now show:

Theorem 12 *Let $X = (\mathbb{R}^d, \|\cdot\|)$ be a symmetric norm with type-2 constant $T_2(X) \in [1, \infty)$ normalized so $\|e_1\|_X = 1$. There exists an algorithm for mean estimation over X making $O(d \log_2(d/\varepsilon))$ queries to $\mathrm{STAT}(\alpha)$, where the accuracy α satisfies $\alpha = \Omega\left(\frac{\varepsilon}{T_2(X) \cdot \log d \cdot \log(d/\varepsilon))}\right)$.*

Proof For $j \in \{0, \dots, 2\log_2(d/\varepsilon)\}$, and $w \in \mathbb{R}^d$, let $R_j(w)$ be the level j vector of w, i.e., $R_j(w) = \sum_{i=1}^{d} e_i w_i \mathbf{1}\{w_i \in (2^{-j-1}, 2^{-j}]\}$. For any fixed distribution $\mathcal{D}$ supported on the unit ball of X, we may consider the distribution $\mathcal{D}_j$ of $R_j(\boldsymbol{x})$ where $\boldsymbol{x} \sim \mathcal{D}$. Denote $\mu = \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}]$ and $\mu_j = \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}_j}[\boldsymbol{x}]$, so that distributions $\mathcal{D}_j$ satisfy $\|\mu - \sum_j \mu_j\|_X \le \varepsilon^2/d$. As a result, the sum of $\varepsilon/(3\log_2(d/\varepsilon))$-approximations of μ_j would result in an ε-approximation of μ.

The algorithm proceeds by estimating the mean of each distribution $\mathcal{D}_j$ and then taking the sum of all estimates:

1. For each $j \in \{0, \dots, 2\log_2(d/\varepsilon)\}$, we consider $\mathcal{H}_\infty^{(j)}$ as the distribution given by $\boldsymbol{x}/2^{-j}$ where $\boldsymbol{x} \sim \mathcal{D}_j$, and $\mathcal{H}_2^{(j)}$ as the distribution given by $\boldsymbol{x}/(2m_X(2^{-j}))$. Note that $\mathcal{H}_\infty^{(j)}$ is supported on B_{ℓ_∞}, and $\mathcal{H}_2^{(j)}$ is supported on B_{ℓ_2}.
   - Perform the mean estimation algorithms for $\mathcal{H}_\infty^{(j)}$ and $\mathcal{H}_2^{(j)}$ as given in Theorem 3.4 and 3.9 of Feldman et al. (2017) (which makes d queries and $2d$ queries, respectively) with error parameter $\varepsilon\gamma$ where[1] $\gamma \gtrsim 1/(T_2(X) \log d \log(d/\varepsilon))$ to obtain vectors $v_\infty^{(j)}, v_2^{(j)} \in \mathbb{R}^d$, and let $w_\infty^{(j)} = 2^{-j} v_\infty^{(j)}$ and $w_2^{(j)} = 2m_X(2^{-j}) v_2^{(j)}$, where

$$\left\| \mu_j - w_\infty^{(j)} \right\|_\infty \le \varepsilon\gamma \cdot 2^{-j} \qquad \text{and} \qquad \left\| \mu_j - w_2^{(j)} \right\|_2 \le 2\varepsilon\gamma \cdot m_X(2^{-j}). \tag{3}$$

1. Here and in the rest of the paper we use $A \gtrsim B$ to mean that there exists an absolute constant $C > 0$, independent of all other parameters, such that $A \ge B/C$, and, analogously, $A \lesssim B$ to mean $A \le CB$

- Find a vector $w^{(j)} \in \mathbb{R}^d$ for which $\|w^{(j)} - w_\infty^{(j)}\|_\infty \le \varepsilon\gamma 2^{-j}$ and $\|w^{(j)} - w_2^{(j)}\|_2 \le 2\varepsilon\gamma m_X(2^{-j})$, and return $w^{(j)}$ as an estimate for μ_j.

2. Given estimates $w^{(j)} \in \mathbb{R}^d$ for all $j \in \{0, \dots, 2\log_2(d/\varepsilon)\}$, output $\sum_j w^{(j)}$.

We note that the inequalities in (3) follow from the fact that $v_\infty^{(j)}$ and $v_2^{(j)}$ are $\varepsilon\gamma$-approximations for $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{H}_\infty^{(j)}}[\boldsymbol{x}]$ (in ℓ_∞) and $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{H}_2^{(j)}}[\boldsymbol{x}]$ (in ℓ_2), respectively, and that

$$2^{-j} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{H}_\infty^{(j)}}[\boldsymbol{x}] = 2m_X(2^{-j}) \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{H}_2^{(2)}}[\boldsymbol{x}] = \mu_j.$$

In order to see that $w^{(j)}$ is a good estimate for μ_j, let $y_j = \mu_j - w^{(j)}$ be the error vector in the approximation. From the triangle inequality, and the definition of $w^{(j)}$, we have $\|y\|_\infty \le 2\varepsilon\gamma \cdot 2^{-j}$ and $\|y\|_2 \le 4\varepsilon\gamma \cdot m_X(2^{-j})$, so that Lemma 11 implies $\|y\|_X \le 20\varepsilon\gamma \cdot T_2(X)\log_2 d \le \varepsilon/(3\log_2(d/\varepsilon))$, for small enough γ and large enough d. ∎

3.2. Lower bounds for normed spaces with large type-2 constants

We now give a lower bound for normed spaces which have large type-2 constant.

Theorem 13 *Let $X = (\mathbb{R}^d, \|\cdot\|_X)$ be a normed space with type-2 constant $T_2(X) \in [1, \infty)$. There exists an $\varepsilon > 0$ such that any statistical algorithm for mean estimation in X with error ε making queries to $\mathrm{VSTAT}(1/(3\kappa^2))$ must make $\exp\left(\Omega\left(\frac{T_2(X)^2\kappa^2}{\varepsilon^2\log d}\right)\right)$ such queries.*

The immediate corollary of Theorem 13 shows the upper bound from Theorem 12 is tight up to poly-logarithmic factors.

Corollary 14 *Let $X = (\mathbb{R}^d, \|\cdot\|_X)$ be a normed space with type-2 constant $T_2(X) \in [1, \infty)$. Any algorithm for mean estimation in X making $d^{O(1)}$-queries to $\mathrm{VSTAT}(\alpha)$ must have $\alpha = O\left(\frac{\varepsilon\log d}{T_2(X)}\right)$.*

We set up some notation and basic observations leading to a proof of Theorem 13. The proof of the next lemma, which shows that we may assume the vectors that certify a large type-2 constant to be of almost equal size, appears in the appendix.

Lemma 15 *Let $X = (\mathbb{R}^d, \|\cdot\|_X)$ be a normed space with type-2 constant $T_2(X) \in [1, \infty)$. Then, for any $t < T_2(X)$, there exists some $n \in \mathbb{N}$, as well as a sequence of vectors $x = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$, where $1 \le \|x_i\|_X \le 2$ for every $i \in [n]$, and*

$$\left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{-1,1\}^n}\left[\left\|\sum_{i=1}^n \boldsymbol{\varepsilon}_i x_i\right\|_X^2\right]\right)^{1/2} \ge t_2(x)\left(\sum_{i=1}^n \|x_i\|_X^2\right)^{1/2} \tag{4}$$

with $t_2(x) > t/C$ for an absolute constant C.

Description of the lower bound instance In this section we describe the instance which achieves the lower bound in Theorem 13.

Fix a sequence $x = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ satisfying (4) guaranteed to exists by Lemma 15, and let the sequence $\widehat{x} = (\widehat{x}_1, \dots, \widehat{x}_n) \in (B_X)^n$ be defined by $\widehat{x}_i = x_i/\|x_i\|_X$. In the language of Feldman et al. (2017), let D be the *reference* distribution supported on B_X given by sampling $\boldsymbol{y} \sim D$ where for all $i \in [n]$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim D}[\boldsymbol{y} = \widehat{x}_i] = \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim D}[\boldsymbol{y} = -\widehat{x}_i] = \frac{1}{2} \cdot \frac{\|x_i\|_X}{\|x\|_{L_1(X)}}, \tag{5}$$

so that $\mu_0 = \mathbf{E}_{\boldsymbol{y}\sim\mathcal{D}}[\boldsymbol{y}] = 0 \in \mathbb{R}^d$. We will let ε_0 be so that $\varepsilon_0 \le t_2(x) \cdot \|x\|_{L_2(X)}/\|x\|_{L_1(X)}$. For $z \in \{-1,1\}^n$, let D_z be the distribution supported on B_X given by sampling $\boldsymbol{y} \sim D_z$ where for all $i \in [n]$,

$$\begin{aligned} \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim D_z}[\boldsymbol{y} = \widehat{x}_i] &= \frac{\|x_i\|_X}{\|x\|_{L_1(X)}} \cdot \left(\frac{1}{2} + \frac{z_i\varepsilon_0}{2 \cdot t_2(x)} \cdot \frac{\|x\|_{L_1(X)}}{\|x\|_{L_2(X)}}\right) \\ \mathop{\mathbf{Pr}}_{\boldsymbol{y}\sim D_z}[\boldsymbol{y} = -\widehat{x}_i] &= \frac{\|x_i\|_X}{\|x\|_{L_1(X)}} \cdot \left(\frac{1}{2} - \frac{z_i\varepsilon_0}{2 \cdot t_2(x)} \cdot \frac{\|x\|_{L_1(X)}}{\|x\|_{L_2(X)}}\right). \end{aligned} \tag{6}$$

Then,

$$\mu_z \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\boldsymbol{y}\sim D_z}[\boldsymbol{y}] = \frac{\varepsilon_0}{t_2(x)\|x\|_{L_2(X)}} \sum_{i=1}^n z_i x_i. \tag{7}$$

Consider the distribution $\mathcal{D}$ on distributions which is uniform over all D_z where $z \in \{-1,1\}^n$. Then, we have:

$$\begin{aligned} \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\{-1,1\}^n}[\|\mu_{\boldsymbol{z}}\|_X] &= \frac{\varepsilon_0}{t_2(x)\|x\|_{L_2(X)}} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{-1,1\}^n}\left[\left\|\sum_{i=1}^n \boldsymbol{\varepsilon}_i x_i\right\|_X\right] &\quad (8)\\ &\gtrsim \frac{\varepsilon_0}{t_2(x)\|x\|_{L_2(X)}} \left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{-1,1\}^n}\left[\left\|\sum_{i=1}^n \boldsymbol{\varepsilon}_i x_i\right\|_X^2\right]\right)^{1/2} = \varepsilon_0, &\quad (9)\\ \mathop{\mathbf{E}}_{\boldsymbol{z}\sim\{-1,1\}^n}\left[\|\mu_{\boldsymbol{z}}\|_X^2\right] = \varepsilon_0^2. & &\quad (10) \end{aligned}$$

where (9) follows from the Khintchine-Kahane inequalities and the definition of $t_2(x)$. By the Paley-Zygmund inequality, $\mathbf{Pr}_{\boldsymbol{z}\sim\{-1,1\}^n}[\|\mu_{\boldsymbol{z}}\|_X \ge \varepsilon] = \Omega(1)$, for some $\varepsilon = \Omega(\varepsilon_0)$. We thus conclude the following lemma, which follows from the preceding discussion.

Lemma 16 *Suppose there exists a statistical algorithm for mean estimation over X with error ε making $q(\varepsilon)$ queries to $\mathrm{VSTAT}(\alpha(\varepsilon))$, then for distribution D as in (5) and set $\mathcal{D}$ as in (6), $\mathcal{B}(D,\mathcal{D})$ has a statistical algorithm making $q(\varepsilon)$ queries of accuracy $\mathrm{VSTAT}(\alpha(\varepsilon))$ which succeeds with constant probability.*

We now turn to computing the statistical dimension of $\mathcal{B}(D,\mathcal{D})$, as described in Definition 8.

Lemma 17 *Let D and $\mathcal{D}$ be the distribution and the set over B_X defined in (5) and (6). For $\kappa > 0$,* $\mathrm{SDN}(\mathcal{B}(D,\mathcal{D}),\kappa) \ge \exp(\Omega(\frac{\kappa^2 t_2(x)^2}{\varepsilon^2}))$.

Proof Let $h\colon B_X \to \mathbb{R}$ be any function with $\|h\|_D = 1$. Note that

$$\mathop{\mathbf{E}}_{\boldsymbol{y}\sim D_z}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\boldsymbol{y}\sim D}[h(\boldsymbol{y})] = \frac{\varepsilon_0}{2t_2(x)\cdot \|x\|_{L_2(X)}} \sum_{i=1}^{n} z_i \|x_i\|_X \left(h(\widehat{x}_i) - h(-\widehat{x}_i)\right),$$

so that by the Hoeffding inequality, any $\alpha > 0$ satisfies

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\boldsymbol{z}\sim\{-1,1\}^n}\left[\left|\mathop{\mathbf{E}}_{\boldsymbol{y}\sim D_{\boldsymbol{z}}}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\boldsymbol{y}\sim D}[h(\boldsymbol{y})]\right| \geq \alpha\right] &\leq \exp\left(-\frac{2\alpha^2 t_2(x)^2 \|x\|_{L_2(X)}^2}{\varepsilon_0^2 \sum_{i=1}^n \|x_i\|_X^2 (h(\widehat{x}_i) - h(-\widehat{x}_i))^2}\right). \\ &\leq \exp\left(-\Omega\left(\frac{\alpha^2 t_2(x)^2}{\varepsilon_0^2}\right)\right),\end{aligned}$$

where we used the fact that $1 \leq \|x_i\| \leq 2$, as well as the fact that $\|h\|_D = 1$ to say that $\|x\|_{L_2(X)}^2 \gtrsim \sum_{i=1}^n \|x_i\|_X^2 (h(\widehat{x}_i) - h(-\widehat{x}_i))^2$. Let $Z \subset \{-1,1\}^n$ be any subset of size $|Z| \geq 2^d/r$, and let $\mathcal{D}_Z = \{D_z : z \in Z\} \subset \mathcal{D}$ be the corresponding set of distributions, and so, similarly to the proof of Lemma 3.21 in Feldman et al. (2017),

$$\mathop{\mathbf{Pr}}_{\boldsymbol{z}\sim Z}\left[\left|\mathop{\mathbf{E}}_{\boldsymbol{y}\sim D}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\boldsymbol{y}\sim D_{\boldsymbol{z}}}[h(\boldsymbol{y})]\right| \geq \alpha\right] \leq r\exp\left(-\Omega\left(\frac{\alpha^2 t_2(x)^2}{\varepsilon_0^2}\right)\right),$$

which implies $\mathbf{E}_{\boldsymbol{z}\sim Z}\left[|\mathbf{E}_{\boldsymbol{y}\sim D}[h(\boldsymbol{y})] - \mathbf{E}_{\boldsymbol{y}\sim D_{\boldsymbol{z}}}[h(\boldsymbol{y})]|\right] \lesssim \frac{\varepsilon_0\sqrt{\ln r}}{t_2(x)}$. Then, for any $\varepsilon \leq \varepsilon_0$, any subset of $\mathcal{D}$ containing at least $\exp(-O(\kappa^2 t_2(x)^2/\varepsilon^2))$-fraction of distributions will have expectation within κ of $\mathbf{E}_{\boldsymbol{y}\sim D}[h(\boldsymbol{y})]$. ∎

Combining Lemma 17, Lemma 16, and Theorem 9, we obtain a proof of Theorem 13.

4. Lower bounds for Schatten-p norms

For the remainder of the section, $S_p = (\mathbb{R}^{d\times d}, \|\cdot\|_{S_p})$ is the Schatten-p normed space, defined over the vector space of $d\times d$ matrices, and $\|x\|_{S_p} = (\sum_{i=1}^d |\sigma_i(x)|^p)^{1/p}$ where $\sigma_i(x)$ is the i-th singular value of x. By a straightforward calculation, the following upper bound holds by embedding into $\ell_2^{d\times d}$ via the identity map, and applying SQ mean estimation algorithm for ℓ_2:

Corollary 18 *There exists a statistical algorithm for mean estimation in S_p making $d^{O(1)}$-queries to $\mathrm{STAT}(\alpha)$ with $\alpha = \Omega\left(\frac{\varepsilon}{d^{1/2-1/p}}\right)$.*

The rest of this section is dedicated to showing the following lower bound, which yields the corresponding lower bound to Corollary 18.

Theorem 19 *There exists an $\varepsilon > 0$ such that any SQ algorithm for mean estimation in S_p with error ε making queries to $\mathrm{VSTAT}(1/(3\kappa^2))$ must make $\exp(\Omega(\min\{\frac{\kappa^2 d^{1-2/p}}{\varepsilon^2}, d+\log\kappa\}))$ queries.*

Similarly to Theorem 13, we obtain the following, which shows that Corollary 18 is optimal.

Corollary 20 *Any statistical algorithm for mean estimation in S_p making $d^{O(1)}$-queries to $\mathrm{STAT}(\alpha)$ must have $\alpha = O\left(\frac{\varepsilon}{d^{1/2-1/p}}\right)$.*

Description of the lower bound instance We now describe the instance which achieves the lower bound in Theorem 19. Consider the distribution D supported on $d \times d$ matrices generated by the following process: 1) let $\boldsymbol{\pi} \sim \mathcal{S}_d$ be a uniformly random permutation on $[d]$, 2) independently sample $\boldsymbol{z} \sim \{-1,1\}^d$, and output the matrix $\boldsymbol{y} = y(\boldsymbol{\pi}, \boldsymbol{z}) = (y(\boldsymbol{\pi}, \boldsymbol{z})_{ij}) \in \mathbb{R}^{d \times d}$ where

$$y(\boldsymbol{\pi}, \boldsymbol{z})_{ij} = \begin{cases} \boldsymbol{z}_i / d^{1/p} & j = \boldsymbol{\pi}(i) \\ 0 & \text{o.w} \end{cases} .$$

Note that $\boldsymbol{y} \sim D$ always satisfies $|\sigma_1(\boldsymbol{y})| = \cdots = |\sigma_d(\boldsymbol{y})| = 1/d^{1/p}$, so that $\|\boldsymbol{y}\|_{S_p} = 1$, and that $\mathbf{E}_{\boldsymbol{y} \sim D}[\boldsymbol{y}] = 0$.

Let $0 < \varepsilon \le \gamma d^{1/p}$ be a parameter for a sufficiently small constant $\gamma > 0$. For $a, b \in \{-1,1\}^d$, let $D_{a,b}$ be the distribution supported on $d \times d$ matrices generated by the following process: 1) let $\boldsymbol{\pi} \sim \mathcal{S}_d$ be a uniformly random permutation on $[d]$, 2) sample $\boldsymbol{z} \sim \{-1,1\}^d$ where each $i \in [d]$ is independently distributed with $\mathbf{Pr}[\boldsymbol{z}_i = a_i b_{\boldsymbol{\pi}(i)}] = \frac{1}{2} + \frac{\varepsilon d^{1/p}}{2}$, and output the matrix $\boldsymbol{y} = y(\boldsymbol{\pi}, \boldsymbol{z})$. Similarly to the case with D, $\boldsymbol{y} \sim D_{a,b}$ always satisfies $|\sigma_1(\boldsymbol{y})| = \cdots = |\sigma_d(\boldsymbol{y})| = 1/d^{1/p}$, so that $\|\boldsymbol{y}\|_{S_p} = 1$. Furthermore, in this case, we have $\mu_{a,b} = \mathbf{E}_{\boldsymbol{y} \sim D_{a,b}}[\boldsymbol{y}] = \frac{\varepsilon}{d} \cdot ab^{\mathsf{T}}$, and $\|\mu_{a,b}\|_{S_p} = \varepsilon$. Finally, we let $\mathcal{D}$ be the set of distributions given by $D_{a,b}$ where $a, b \in \{-1,1\}^d$. Since every distribution in $\mathcal{D}$ has mean with S_p norm at least ε, we obtain the following lemma.

Lemma 21 *Suppose there is a statistical algorithm for mean estimation with error ε for making $q(\varepsilon)$ queries of accuracy $\alpha(\varepsilon)$, then $\mathcal{B}(D, \mathcal{D})$ has a randomized statistical algorithm making $q(\varepsilon)$ queries of accuracy $\alpha(\varepsilon)$ succeeding with the same probability.*

Similarly to the case in Section 3.2, we obtain lower bounds on algorithms using statistical queries by giving a lower bound on the statistical dimension of $\mathcal{B}(D, \mathcal{D})$. Theorem 19 is implied by Lemma 22 below.

Lemma 22 *Let D and $\mathcal{D}$ be the distribution and the set over B_{S_p} defined above. For $\kappa > 0$,* $\mathrm{SDN}(\mathcal{B}(D, \mathcal{D}), \kappa) \ge \exp(\Omega(\min\{\frac{\kappa^2 d^{1-2/p}}{\varepsilon^2}, d + \log \kappa\}))$.

Proof Let $h \colon B_{S_p} \to \mathbb{R}$ be any function with $\|h\|_D = 1$, and denote the Boolean function $H_h \colon \{-1,1\}^d \times \{-1,1\}^d \to \mathbb{R}$ by:

$$\begin{aligned} H_h(a,b) &= \mathop{\mathbf{E}}_{\boldsymbol{y} \sim D_{a,b}}[h(\boldsymbol{y})] - \mathop{\mathbf{E}}_{\boldsymbol{y} \sim D}[h(\boldsymbol{y})] \\ &= \frac{1}{d!} \sum_{\pi \in \mathcal{S}_d} \frac{1}{2^d} \sum_{z \in \{-1,1\}^d} h(y(\pi, z)) \left(\prod_{i=1}^d (1 + \varepsilon d^{1/p} z_i a_i b_{\pi(i)}) - 1 \right) \\ &= \frac{1}{d!} \sum_{\pi \in \mathcal{S}_d} \sum_{S \subset [d]} (\varepsilon d^{1/p})^{|S|} \cdot \chi_S(ab_\pi) \cdot \widehat{h_\pi}(S), \end{aligned} \tag{11}$$

where we write $h_\pi \colon \{-1,1\}^d \to [0,1]$ to denote $h_\pi(z) = h(y(\pi, z))$, for $S \subset [d]$, $\chi_S \colon \{-1,1\}^d \to \{-1,1\}$ is given by $\chi_S(z) = \prod_{i \in S} z_i$, and $ab_\pi \in \{-1,1\}^d$ denotes the vector where $(ab_\pi)_i = a_i b_{\pi(i)}$. Further consolidating terms, we can write

$$H_h(a,b) = \frac{1}{d!} \sum_{t=1}^d (\varepsilon d^{1/p})^t \sum_{\substack{S,T \subseteq [d] \\ |S|=|T|=t}} \Gamma_{S,T} \cdot \chi_S(a) \chi_T(b) \qquad \text{where} \qquad \Gamma_{S,T} = \sum_{\substack{\pi \in \mathcal{S}_d: \\ \pi(S)=T}} \widehat{h_\pi}(S). \tag{12}$$

Similarly to the proof of Lemma 17, we will use a concentration bound on $H(\boldsymbol{a}, \boldsymbol{b})$ when $\boldsymbol{a}, \boldsymbol{b} \sim \{-1,1\}^d$ to derive a bound on the statistical dimension. Specifically, Lemma 23 (which we state and prove next), as well as a union bound, implies that for any $2 \le q \le d/(2e)$, and any set of pairs $Z \subset \{-1,1\}^d \times \{-1,1\}^d$ of size at least $2^{2d}/r$, and $\mathcal{D}_Z = \{D_{a,b} : (a,b) \in Z\}$,

$$\mathop{\mathbf{Pr}}_{(\boldsymbol{a},\boldsymbol{b})\sim Z}\left[|H_h(\boldsymbol{a},\boldsymbol{b})| \ge \frac{4e\sqrt{q}\cdot\varepsilon}{d^{1/2-1/p}}\right] \le r2^{-q}.$$

We may also apply Cauchy-Schwartz inequality to (11) to say that for every $a, b \in \{-1,1\}^d$,

$$\begin{aligned} H_h(a,b) &\le \left(\frac{1}{d!}\sum_{\pi\in\mathcal{S}_d}\sum_{S\subset[d]}(\varepsilon d^{1/p})^{2|S|}\right)^{1/2}\left(\frac{1}{d!}\sum_{\pi\in\mathcal{S}_d}\sum_{S\subset[d]}\widehat{h_\pi}(S)^2\right)^{1/2} \\ &\le (1+\varepsilon^2 d^{2/p})^{d/2}\cdot\|h\|_D = (1+\varepsilon^2 d^{2/p})^{d/2}. \end{aligned}$$

This, in turn, implies

$$\mathop{\mathbf{E}}_{(\boldsymbol{a},\boldsymbol{b})\sim Z}[|H_h(\boldsymbol{a},\boldsymbol{b})|] \lesssim \frac{\sqrt{\log r}\cdot\varepsilon}{d^{1/2-1/p}} + \left(1+\varepsilon^2 d^{2/p}\right)^{d/2}\cdot r2^{-d/(2e)} \lesssim \frac{\sqrt{\log r}\cdot\varepsilon}{d^{1/2-1/p}} + r\cdot 2^{-d/6}$$

when ε is a small constant times $d^{-1/p}$. Therefore, we have $\mathbf{E}_Z[|H_h(\boldsymbol{a},\boldsymbol{b})|] \le \kappa$ for all subsets containing at least $2^{2d}/r$ distributions, where $r = \exp(\Omega(\min\{\frac{\kappa^2 d^{1-2/p}}{\varepsilon^2}, d + \log\kappa\}))$. ■

We now prove the concentration inequality for $H_h(\boldsymbol{a},\boldsymbol{b})$ used in the proof of Lemma 22.

Lemma 23 *Let $h\colon B_{S_p} \to \mathbb{R}$ satisfy $\|h\|_D = 1$, and let $H_h\colon \{-1,1\}^d \times \{-1,1\}^d \to \mathbb{R}$ be the function in (12). Then, for any $2 \le q \le d/(2e)$, $\mathbf{Pr}_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}[|H_h(\boldsymbol{a},\boldsymbol{b})| > \frac{4e\sqrt{q}\varepsilon}{d^{1/2-1/p}}] \le 2^{-q}$.*

To prove this lemma, we set up additional technical machinery. Recall that for any $\rho \in [-1,\infty)$ the *noise operator* $\mathscr{T}_\rho$ is the linear operator on Boolean functions, defined so that for any Boolean function $f\colon \{-1,1\}^m \to \mathbb{R}$ with Fourier expansion $f(x) = \sum_{S\subseteq[m]}\widehat{f}(S)\chi_S(x)$ where $\chi_S(x) = \prod_{i\in S}x_i$, we have $\mathscr{T}_\rho f(x) = \sum_{S\subseteq[m]}\rho^{|S|}\widehat{f}(S)\chi_S(x)$.[2] We will use the following version of the hypercontractivity theorem, which will allow us to bound moments of random Boolean functions.

Theorem 24 ($(2,q)$-Hypercontractivity, Chapter 9 in O'Donnell (2014)) *Let $f\colon \{-1,1\}^m \to \mathbb{R}$, and let $2 \le q \le \infty$. Then for $\rho = 1/\sqrt{q-1}$, $\mathbf{E}_{\boldsymbol{x}\sim\{-1,1\}^m}[|\mathscr{T}_\rho f(\boldsymbol{x})|^q] \le \mathbf{E}_{\boldsymbol{x}\sim\{-1,1\}^m}\left[f(\boldsymbol{x})^2\right]^{q/2}$.*

Proof [Proof of Lemma 23] Define the auxiliary Boolean function $g\colon \{-1,1\}^d \times \{-1,1\}^d \to \mathbb{R}$ by

$$g(a,b) = \frac{1}{d!}\sum_{t=1}^{d}\sum_{\substack{S,T\subseteq[d]\\|S|=|T|=t}}\Gamma_{S,T}\cdot\chi_S(a)\chi_T(b),$$

2. The operator $\mathscr{T}_\rho$ is typically only defined for $\rho \in [-1,1]$, but one may naturally extend this definition to $\rho > 1$, see e.g. O'Donnell (2014).

for $\Gamma_{S,T}$ as in (12). Note that for $\sigma = \sqrt{\varepsilon d^{1/p}(q-1)}$ and $\rho = 1/\sqrt{q-1}$, $H_h(a,b) = \mathscr{T}_\rho \mathscr{T}_\sigma g(a,b)$. For all $2 \le q \le \infty$, we have

$$\Pr_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}\left[|H_h(\boldsymbol{a},\boldsymbol{b})| > \alpha\right] \le \frac{\mathbf{E}_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}\left[|H_h(\boldsymbol{a},\boldsymbol{b})|^q\right]}{\alpha^q}$$
$$\le \frac{\mathbf{E}_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}\left[\mathscr{T}_\sigma g(\boldsymbol{a},\boldsymbol{b})^2\right]^{q/2}}{\alpha^q}, \tag{13}$$

where the first inequality follows from Markov's inequality, and the second from $(2,q)$-hypercontractivity (Theorem 24). By Parseval's identity, observe that

$$\mathop{\mathbf{E}}_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}\left[\mathscr{T}_\sigma g(a,b)^2\right] \le \varepsilon^2 d^{2/p}\left(\frac{1}{d!}\right)^2 \sum_{t=1}^{d} q^t \sum_{\substack{S,T\subseteq[d]\\|S|=|T|=t}} \Gamma_{S,T}^2.$$

For any fixed $1 \le t \le d$, recall from (12) that

$$\sum_{\substack{S,T\subseteq[d]\\|S|=|T|=t}} \Gamma_{S,T}^2 = \sum_{\substack{S,T\subseteq[d]\\|S|=|T|=t}} \left(\sum_{\substack{\pi\in\mathcal{S}_d\\\pi(S)=T}} \widehat{h_\pi}(S)\right)^2 \overset{(a)}{\le} (d-t)! \sum_{\substack{S,T\subseteq[d]\\|S|=|T|=t}} \sum_{\substack{\pi\in\mathcal{S}_d\\\pi(S)=T}} \widehat{h_\pi}(S)^2$$
$$= (d-t)! \sum_{\pi\in\mathcal{S}_d} \sum_{\substack{S\subseteq[d]\\|S|=t}} \widehat{h_\pi}(S)^2 \overset{(b)}{\le} (d-t)!d!,$$

where (a) follows by Cauchy-Schwarz, and (b) follows since $\frac{1}{d!}\sum_{\pi\in\mathcal{S}_d}\sum_{S\subset[d]}\widehat{h_\pi}(S)^2 = 1$, as $\|h\|_D = 1$. Summing over all $t \in [d]$, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}\left[\mathscr{T}_\sigma g(\boldsymbol{a},\boldsymbol{b})^2\right] \le \varepsilon^2 d^{2/p}\left(\frac{1}{d!}\right)^2 \sum_{t=1}^{d} q^t (d-t)!d! = \varepsilon^2 d^{2/p} \sum_{t=1}^{d} \frac{q^t(d-t)!}{d!}$$
$$= q\varepsilon^2 d^{2/p-1} \sum_{t=0}^{d-1} \frac{q^t(d-t-1)!}{(d-1)!}, \tag{14}$$

and using Stirling's approximation,

$$\sum_{t=0}^{d-1} \frac{q^t(d-t-1)!}{(d-1)!} \le \sum_{t=0}^{d-1} eq^t \sqrt{\frac{d-t-1}{d-1}}\left(\frac{(d-t-1)}{e}\right)^{d-t-1}\left(\frac{e}{d-1}\right)^{d-1}$$
$$\le e\sum_{t=0}^{d-1}\left(\frac{eq}{d}\right)^t \le 2e,$$

for all $q \le d/(2e)$. Therefore (14) simplifies to give $\mathbf{E}_{\boldsymbol{a},\boldsymbol{b}\sim\{-1,1\}^d}\left[\mathscr{T}_\sigma g(\boldsymbol{a},\boldsymbol{b})^2\right] \le 2eq\varepsilon^2 d^{2/p-1}$, for all $q \le d/(2e)$, and plugging this bound into (13) while letting $\alpha = 4e\sqrt{q}\cdot\varepsilon/d^{1/2-1/p}$, we obtain the desired concentration bound. ■

References

Alexandr Andoni, Huy L Nguyen, Aleksandar Nikolov, Ilya Razenshteyn, and Erik Waingarten. Approximate near neighbors for general symmetric norms. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 902–913. ACM, 2017.

Alexandr Andoni, Chengyu Lin, Ying Sheng, Peilin Zhong, and Ruiqi Zhong. Subspace embedding and linear regression with orlicz norm. *arXiv preprint arXiv:1806.06430*, 2018.

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Invent. Math.*, 115(3):463–482, 1994. ISSN 0020-9910. doi: 10.1007/BF01231769. URL https://doi.org/10.1007/BF01231769.

Jarosław Błasiok, Vladimir Braverman, Stephen R Chestnut, Robert Krauthgamer, and Lin F Yang. Streaming symmetric norms via measure concentration. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 716–729. ACM, 2017.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 73–84. IEEE, 2017.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

Vitaly Feldman. A general characterization of the statistical query complexity. *arXiv preprint arXiv:1608.02198*, 2016a.

Vitaly Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016b. doi: 10.1007/978-1-4939-2864-4_401. URL https://doi.org/10.1007/978-1-4939-2864-4_401.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the 45th ACM Symposium on the Theory of Computing (STOC '2013)*, 2013a.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 655–664. ACM, 2013b.

Vitaly Feldman, Cristóbal Guzmán, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the 28th ACM-SIAM Symposium on Discrete Algorithms (SODA '2017)*, 2017.

Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. *SIAM Journal on Computing*, 47(4):1294–1338, 2018.

Fritz John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*, pages 187–204. Interscience Publishers, Inc., New York, N. Y., 1948.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Adam R Klivans and Alexander A Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2-3):97–114, 2007.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. ISBN 978-3-642-20211-7. Isoperimetry and processes, Reprint of the 1991 edition.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Le Song, Santosh Vempala, John Wilmes, and Bo Xie. On the complexity of learning neural networks. In *Advances in Neural Information Processing Systems*, pages 5514–5522, 2017.

Zhao Song, David P Woodruff, and Peilin Zhong. Towards a zero-one law for entrywise low rank approximation. *arXiv preprint arXiv:1811.01442*, 2018.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516, 2016.

Przemyslaw Wojtaszczyk. *Banach spaces for analysts*, volume 25. Cambridge University Press, 1996.

Appendix A. Omitted Proofs

Proof [Proof of Lemma 15] Since $t < T_2(X)$, there exists a sequence $x' = (x'_1, \ldots, x'_m)$ such that

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{-1,1\}^m}\left[\left\|\sum_{i=1}^m \boldsymbol{\varepsilon}_i x'_i\right\|_X^2\right] \geq t\sum_{i=1}^m \|x'_i\|_X^2.$$

Lemma 4.5 in Ledoux and Talagrand (2011) gives that we can replace the Rademacher random variables with Gaussians, i.e. for a sequence of independent standard Gaussian random variables we have $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m$,

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}}\left[\left\|\sum_{i=1}^m \boldsymbol{\varepsilon}_i x'_i\right\|_X^2\right] \leq \frac{\pi}{2}\mathop{\mathbf{E}}_{\boldsymbol{g}}\left[\left\|\sum_{i=1}^m \boldsymbol{g}_i x'_i\right\|_X^2\right].$$

Let us assume, without loss of generality, that $\|x'_i\|_X \geq 1$ for every $i \in [n]$. For any x'_i, define the sequence $x'_{i,1}, \ldots, x'_{i,m_i}$ to consist of $\lfloor \|x'_i\|_X^2 \rfloor - 1$ copies of $x'_i/\|x'_i\|_X$ and a single copy of

$$(1 + \|x'_i\|_X^2 - \lfloor \|x'_i\|_X^2 \rfloor)^{1/2} \cdot \frac{x'_i}{\|x'_i\|_X}.$$

This guarantees that $1 \leq \|x'_{i,j}\|_X \leq 2$ for every $i \in [n]$ and $j \in m_i$. Observe also that, if $\boldsymbol{g}_{i,1}, \ldots, \boldsymbol{g}_{i,m_i}$ are independent standard Gaussian random variables, then $\sum_{j=1}^{m_i} \boldsymbol{g}_{i,j} x'_{i,j}$ is distributed identically to $\boldsymbol{g}_i x'_i$, and, moreover, $\sum_{j=1}^{m_i} \|x'_{i,j}\|_X^2 = \|x'_i\|_X^2$. Therefore, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{g}}\left[\left\|\sum_{i=1}^{m}\sum_{j=1}^{m_i} \boldsymbol{g}_{i,j} x'_{i,j}\right\|_X^2\right] \geq \frac{2t}{\pi} \sum_{i=1}^{m}\sum_{j=1}^{m_i} \|x'_{i,j}\|_X^2.$$

By the Gaussian version of the Khintchine-Kahane inequalities (Corollary 3.2. in Ledoux and Talagrand (2011)) and the Paley-Zygmund inequality, we have that for some absolute constant C',

$$\mathbf{Pr}\left[\left\|\sum_{i=1}^{m}\sum_{j=1}^{m_i} \boldsymbol{g}_{i,j} x'_{i,j}\right\|_X^2 \geq \frac{t}{C'} \sum_{i=1}^{m}\sum_{j=1}^{m_i} \|x'_{i,j}\|_X^2\right] \geq \frac{1}{2}.$$

We define the sequence $x = (x_1, \ldots, x_n)$ to contain N copies of each vector $x_{i,j}$, for some large enough integer N. By the central limit theorem, as $N \to \infty$, $\frac{1}{\sqrt{N}} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i x_i$ converges in probability to $\sum_{i=1}^{m}\sum_{j=1}^{m_i} \boldsymbol{g}_{i,j} x'_{i,j}$. Since $\|\cdot\|_X$ is continuous, this means that, for a large enough N,

$$\mathbf{Pr}\left[\left\|\sum_{i=1}^{m}\sum_{j=1}^{m_i} \boldsymbol{g}_{i,j} x'_{i,j} - \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i x_i\right\|_X^2 > \frac{t}{4C'} \sum_{i=1}^{m}\sum_{j=1}^{m_i} \|x'_{i,j}\|_X\right] \leq \frac{1}{4}.$$

Then, by the triangle inequality,

$$\mathbf{Pr}\left[\left\|\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i x_i\right\|_X^2 \geq \frac{t}{4C'} \sum_{i=1}^{n} \|x_i\|_X^2\right] \geq \frac{1}{4},$$

and the lemma follows with $C = 4C'$. ■