A Simple Tool for Bounding the Deviation of Random Matrices on Geometric Sets

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin

Abstract Let A be an isotropic, sub-gaussian $m \times n$ matrix. We prove that the process $Z_x := \|Ax\|_2 - \sqrt{m}\,\|x\|_2$ has sub-gaussian increments, that is, $\|Z_x - Z_y\|_{\psi_2} \leq C\|x-y\|_2$ for any $x, y \in \mathbb{R}^n$. Using this, we show that for any bounded set $T \subseteq \mathbb{R}^n$, the deviation of $\|Ax\|_2$ around its mean is uniformly bounded by the Gaussian complexity of T. We also prove a local version of this theorem, which allows for unbounded sets. These theorems have various applications, some of which are reviewed in this paper. In particular, we give a new result regarding model selection in the constrained linear model.

1 Introduction

Recall that a random variable Z is *sub-gaussian* if its distribution is dominated by a normal distribution. One of several equivalent ways to define this rigorously is to require the Orlicz norm

$$\|Z\|_{\psi_2} := \inf\left\{K > 0 : \mathbb{E}\psi_2(|Z|/K) \leq 1\right\}$$

C. Liaw
Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, BC, Canada V6T 1Z4
e-mail: cvliaw@cs.ubc.ca

A. Mehrabian
Department of Computer Science, University of British Columbia, 2366 Main Mall, Vancouver, BC, Canada V6T 1Z4

School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6
e-mail: abbasmehrabian@gmail.com

Y. Plan
Department of Mathematics, University of British Columbia, 1984 Mathematics Rd, Vancouver, BC, Canada V6T 1Z4
e-mail: yaniv@math.ubc.ca

R. Vershynin (✉)
Department of Mathematics, University of Michigan, 530 Church St., Ann Arbor, MI 48109, USA
e-mail: romanv@umich.edu

B. Klartag, E. Milman (eds.), *Geometric Aspects of Functional Analysis*,
Lecture Notes in Mathematics 2169, DOI 10.1007/978-3-319-45282-1_18

to be finite, for the Orlicz function $\psi_2(x) = \exp(x^2) - 1$. Also recall that a random vector X in $\mathbb{R}^n$ is sub-gaussian if all of its one-dimensional marginals are sub-gaussian random variables; this is quantified by the norm

$$\|X\|_{\psi_2} := \sup_{\theta \in S^{n-1}} \big\| \langle X, \theta \rangle \big\|_{\psi_2}.$$

For basic properties and examples of sub-gaussian random variables and vectors, see e.g. [27].

In this paper we study *isotropic, sub-gaussian random matrices* A. This means that we require the rows A_i of A to be independent, isotropic, and sub-gaussian random vectors:

$$\mathbb{E} A_i A_i^{\mathsf{T}} = I, \quad \|A_i\|_{\psi_2} \le K. \tag{1}$$

In Remark 1 below we show how to remove the isotropic assumption.

Suppose A is an $m \times n$ isotropic, sub-gaussian random matrix, and $T \subset \mathbb{R}^n$ is a given set. We are wondering when A acts as an approximate isometry on T, that is, when $\|Ax\|_2$ concentrates near the value $(\mathbb{E}\|Ax\|_2^2)^{1/2} = \sqrt{m}\|x\|_2$ uniformly over vectors $x \in T$.

Such a uniform deviation result must somehow depend on the "size" of the set T. A simple way to quantify the size of T is through the *Gaussian complexity*

$$\gamma(T) := \mathbb{E} \sup_{x \in T} |\langle g, x \rangle| \quad \text{where } g \sim N(0, I_n). \tag{2}$$

One can often find in the literature the following translation-invariant cousin of Gaussian complexity, called the *Gaussian width* of T:

$$w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle = \frac{1}{2} \mathbb{E} \sup_{x \in T - T} \langle g, x \rangle .$$

These two quantities are closely related. Indeed, a standard calculation shows that

$$\frac{1}{3}\big[w(T) + \|y\|_2\big] \le \gamma(T) \le 2\big[w(T) + \|y\|_2\big] \quad \text{for every } y \in T. \tag{3}$$

The reader is referred to [19, Sect. 2], [28, Sect. 3.5] for other basic properties of Gaussian width. Our main result is that the deviation of $\|Ax\|_2$ over T is uniformly bounded by the Gaussian complexity of T.

Theorem 1 (Deviation of Random Matrices on Sets) *Let A be an isotropic, sub-gaussian random matrix as in (1), and T be a bounded subset of $\mathbb{R}^n$. Then*

$$\mathbb{E} \sup_{x \in T} \big| \|Ax\|_2 - \sqrt{m}\|x\|_2 \big| \le CK^2 \cdot \gamma(T).$$

(Throughout, c and C denote absolute constants that may change from line to line). For Gaussian random matrices A, this theorem follows from a result of Schechtman [23]. For sub-gaussian random matrices A, one can find related results in [4, 10, 13]. Comparisons with these results can be found in Sect. 3.

The dependence of the right-hand-side of this theorem on T is essentially optimal. This is not hard to see for $m = 1$ by a direct calculation. For general m, optimality follows from several consequences of Theorem 1 that are known to be sharp; see Sect. 2.5.

We do not know if the dependence on K in the theorem is optimal or if the dependence can be improved to linear. However, none of the previous results have shown a linear dependence on K even in partial cases.

Remark 1 (Removing Isotropic Condition) Theorem 1 and the results below may also be restated without the assumption that A is isotropic using a simple linear transformation. Indeed, suppose that instead of being isotropic, each row of A satisfies $\mathbb{E}A_i A_i^T = \Sigma$ for some invertible covariance matrix Σ. Consider the whitened version $B_i := \sqrt{\Sigma^{-1}} A_i$. Note that $\|B_i\|_{\psi_2} \le \|\sqrt{\Sigma^{-1}}\| \cdot \|A_i\|_{\psi_2} \le \|\sqrt{\Sigma^{-1}}\| \cdot K$. Let B be the random matrix whose ith row is B_i. Then

$$\begin{aligned}\mathbb{E} \sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m} \|\sqrt{\Sigma} x\|_2 \right| &= \mathbb{E} \sup_{x \in T} \left| \|B\sqrt{\Sigma} x\|_2 - \sqrt{m} \|\sqrt{\Sigma} x\|_2 \right| \\ &= \mathbb{E} \sup_{x \in \sqrt{\Sigma} T} \left| \|Bx\|_2 - \sqrt{m} \|x\|_2 \right| \\ &\le C \|\Sigma^{-1}\| K^2 \gamma(\sqrt{\Sigma} T).\end{aligned}$$

The last line follows from Theorem 1. Note also that $\gamma(\sqrt{\Sigma} T) \le \|\sqrt{\Sigma}\| \gamma(T) = \sqrt{\|\Sigma\|} \gamma(T)$, which follows from Sudakov-Fernique's inequality. Summarizing, our bounds can be extended to anisotropic distributions by including in them the smallest and largest eigenvalues of the covariance matrix Σ.

Our proof of Theorem 1 given in Sect. 4.1 is particularly simple, and is inspired by the approach of Schechtman [23]. He showed that for Gaussian matrices A, the random process $Z_x := \|Ax\|_2 - (\mathbb{E}\|Ax\|_2^2)^{1/2}$ indexed by points $x \in \mathbb{R}^n$, has sub-gaussian increments, that is

$$\|Z_x - Z_y\|_{\psi_2} \le C \|x - y\|_2 \quad \text{for every } x, y \in \mathbb{R}^n. \tag{4}$$

Then Talagrand's Majorizing Measure Theorem implies the desired conclusion that[1] $\mathbb{E} \sup_{x \in T} |Z_x| \lesssim \gamma(T)$.

However, it should be noted that G. Schechtman's proof of (4) makes heavy use of the rotation invariance property of the Gaussian distribution of A. When A is only sub-gaussian, there is no rotation invariance to rely on, and it was unknown if one

[1] In this paper, we sometimes hide absolute constants in the inequalities marked $\lesssim$.

can transfer G. Schechtman's argument to this setting. This is precisely what we do here: we show that, perhaps surprisingly, the sub-gaussian increment property (4) holds for general sub-gaussian matrices A.

Theorem 2 (Sub-gaussian Process) *Let A be an isotropic, sub-gaussian random matrix as in (1). Then the random process*

$$Z_x := \|Ax\|_2 - (\mathbb{E}\|Ax\|_2^2)^{1/2} = \|Ax\|_2 - \sqrt{m}\|x\|_2$$

has sub-gaussian increments:

$$\|Z_x - Z_y\|_{\psi_2} \leq CK^2\|x - y\|_2 \quad \text{for every } x, y \in \mathbb{R}^n. \tag{5}$$

The proof of this theorem, given in Sect. 5, essentially consists of a couple of non-trivial applications of Bernstein's inequality; parts of the proof are inspired by G. Schechtman's argument. Applying Talagrand's Majorizing Measure Theorem (see Theorem 8 below), we immediately obtain Theorem 1.

We also prove a high-probability version of Theorem 1.

Theorem 3 (Deviation of Random Matrices on Sets: Tail Bounds) *Under the assumptions of Theorem 1, for any $u \geq 0$ the event*

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leq CK^2\big[w(T) + u \cdot \mathrm{rad}(T)\big]$$

holds with probability at least $1 - \exp(-u^2)$. Here $\mathrm{rad}(T) := \sup_{x \in T} \|x\|_2$ denotes the radius of T.

This result will be deduced in Sect. 4.1 from a high-probability version of Talagrand's theorem.

In the light of the equivalence (3), notice that Theorem 3 implies the following simpler but weaker bound

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leq CK^2 u \cdot \gamma(T) \tag{6}$$

if $u \geq 1$. Note that even in this simple bound, $\gamma(T)$ cannot be replaced with the Gaussian width $w(T)$, e.g. the result would fail for a singleton T. This explains why the radius of T appears in Theorem 3.

Restricting the set T to the unit sphere, we obtain the following corollary.

Corollary 1 *Under the assumptions of Theorem 1, for any $u \geq 0$ the event*

$$\sup_{x \in T \cap S^{n-1}} \left| \|Ax\|_2 - \sqrt{m} \right| \leq CK^2 \left[w(T \cap S^{n-1}) + u\right]$$

holds with probability at least $1 - \exp(-u^2)$.

In Theorems 1 and 3, we assumed that the set T is bounded. For unbounded sets, we can still prove a 'local version' of Theorem 3. Let us state a simpler form of this result here. In Sect. 6, we will prove a version of the following theorem with a better probability bound.

Theorem 4 (Local Version) *Let $(Z_x)_{x\in\mathbb{R}^n}$ be a random process with sub-gaussian increments as in (5). Assume that the process is homogeneous, that is, $Z_{\alpha x} = \alpha Z_x$ for any $\alpha \geq 0$. Let T be a star-shaped[2] subset of $\mathbb{R}^n$, and let $t \geq 1$. With probability at least $1 - \exp(-t^2)$, we have*

$$|Z_x| \leq t \cdot CK^2\gamma\left(T \cap \|x\|_2 B_2^n\right) \quad \text{for all } x \in T. \tag{7}$$

Combining with Theorem 2, we immediately obtain the following result.

Theorem 5 (Local Version of Theorem 3) *Let A be an isotropic, sub-gaussian random matrix as in (1), and let T be a star-shaped subset of $\mathbb{R}^n$, and let $t \geq 1$. With probability at least $1 - \exp(-t^2)$, we have*

$$\left|\|Ax\|_2 - \sqrt{m}\|x\|_2\right| \leq t \cdot CK^2\gamma\left(T \cap \|x\|_2 B_2^n\right) \quad \text{for all } x \in T. \tag{8}$$

Remark 2 We note that Theorems 4 and 5 can also apply when T is not a star-shaped set, simply by considering the smallest star-shaped set that contains T:

$$\operatorname{star}(T) := \bigcup_{\lambda\in[0,1]} \lambda T.$$

Then one only needs to replace T by $\operatorname{star}(T)$ in the right-hand side of Eqs. (7) and (8).

Results of the type of Theorems 1, 3 and 5 have been useful in a variety of applications. For completeness, we will review some of these applications in the next section.

2 Applications

Random matrices have proven to be useful both for modeling data and transforming data in a variety of fields. Thus, the theory of this paper has implications for several applications. A number of classical theoretical discoveries as well as some new results follow directly from our main theorems. In particular, the local version of our theorem (Theorem 5), allows a new result in model selection under the constrained linear model, with applications in *compressed sensing*. We give details below.

[2]Recall that a set T is called star-shaped if $t \in T$ implies $\lambda t \in T$ for all $\lambda \in [0, 1]$.

2.1 Singular Values of Random Matrices

The singular values of a random matrix are an important topic of study in random matrix theory. A small sample includes covariance estimation [26], stability in numerical analysis [29], and quantum state tomography [8].

Corollary 1 may be specialized to bound the singular values of a sub-gaussian matrix. Indeed, take $T = S^{n-1}$ and note that $w(T) \leq \sqrt{n}$. Then the corollary states that, with high probability,

$$\left| \|Ax\|_2 - \sqrt{m} \right| \leq CK^2\sqrt{n} \qquad \text{for all } x \in S^{n-1}.$$

This recovers the well-known result that, with high probability, all of the singular values of A reside in the interval $[\sqrt{m} - CK^2\sqrt{n}, \sqrt{m} + CK^2\sqrt{n}]$ (see [27]). When $nK^4 \ll m$, all of the singular values concentrate around $\sqrt{m}$. In other words, a tall random matrix is well conditioned with high probability.

2.2 Johnson-Lindenstrauss Lemma

The Johnson-Lindenstrauss lemma [9] describes a simple and effective method of dimension reduction. It shows that a (finite) set of data vectors $\mathcal{X}$ belonging to a very high-dimensional space, $\mathbb{R}^n$, can be mapped to a much lower dimensional space while roughly preserving pairwise distances. This is useful from a computational perspective since the storage space and the speed of computational tasks both improve in the lower dimensional space. Further, the mapping can be done simply by multiplying each vector by the random matrix $A/\sqrt{m}$.

The classic Johnson-Lindenstrauss lemma follows immediately from our results. Indeed, take $T' = \mathcal{X} - \mathcal{X}$. To construct T, remove the 0 vector from T' and project all of the remaining vectors onto S^{n-1} (by normalizing). Since T belongs to the sphere and has fewer than $|\mathcal{X}|^2$ elements, it is not hard to show that $\gamma(T) \leq C\sqrt{\log|\mathcal{X}|}$. Then by Corollary 1, with high probability,

$$\sup_{x \in T} \left| \frac{1}{\sqrt{m}} \|Ax\|_2 - 1 \right| \leq \frac{CK^2\sqrt{\log|\mathcal{X}|}}{\sqrt{m}}.$$

Equivalently, for all $x, y \in \mathcal{X}$

$$(1-\delta)\|x-y\|_2 \leq \frac{1}{\sqrt{m}}\|A(x-y)\|_2 \leq (1+\delta)\|x-y\|_2, \qquad \delta = \frac{CK^2\sqrt{\log|\mathcal{X}|}}{\sqrt{m}}.$$

This is the classic Johnson-Lindenstrauss lemma. It shows that as long as $m \gg K^4 \log|\mathcal{X}|$, the mapping $x \to Ax/\sqrt{m}$ nearly preserves pair-wise distances. In other

words, $\mathcal{X}$ may be embedded into a space of dimension slightly larger than $\log|\mathcal{X}|$ while preserving distances.

In contrast to the classic Johnson-Lindenstrauss lemma that applies only to finite sets $\mathcal{X}$, the argument above based on Corollary 1 allows $\mathcal{X}$ to be infinite. In this case, the size of $\mathcal{X}$ is quantified using the notion of Gaussian width instead of cardinality.

To get even more precise control of the geometry of $\mathcal{X}$ in Johnson-Lindenstrauss lemma, we may use the local version of our results. To this end, apply Theorem 5 combined with Remark 2 to the set $T = \mathcal{X} - \mathcal{X}$. This shows that with high probability, for all $x, y \in \mathcal{X}$,

$$\left| \frac{1}{\sqrt{m}} \|A(x-y)\|_2 - \|x-y\|_2 \right| \le \frac{CK^2\gamma\left(\operatorname{star}(\mathcal{X}-\mathcal{X}) \cap \|x-y\|_2 B_2^n\right)}{\sqrt{m}}. \tag{9}$$

One may recover the classic Johnson-Lindenstrauss lemma from the above bound using the containment $\operatorname{star}(\mathcal{X}-\mathcal{X}) \subset \operatorname{cone}(\mathcal{X}-\mathcal{X})$. However, the above result also applies to infinite sets, and further can benefit when $\mathcal{X}-\mathcal{X}$ has different structure at different scales, e.g., when $\mathcal{X}$ has clusters.

2.3 Gordon's Escape Theorem

In [7], Gordon answered the following question: *Let T be an arbitrary subset of S^{n-1}. What is the probability that a random subspace has nonempty intersection with T?* Gordon showed that this probability is small provided that the codimension of the subspace exceeds $w(T)$. This result also follows from Corollary 1 for a general model of random subspaces.

Indeed, let A be an isotropic, sub-gaussian $m \times n$ random matrix as in (1). Then its kernel $\ker A$ is a *random subspace* in $\mathbb{R}^n$ of dimension at least $n-m$. Corollary 1 implies that, with high probability,

$$\ker A \cap T = \emptyset \tag{10}$$

provided that $m \ge CK^4 w(T)^2$. To see this, note that in this case Corollary 1 yields that $\left| \|Ax\|_2 - \sqrt{m} \right| < \sqrt{m}$ for all $x \in T$, so $\|Ax\|_2 > 0$ for all $x \in T$, which in turn is equivalent to (10).

We also note that there is an equivalent version of the above result when T is a cone. Then, with high probability,

$$\ker A \cap T = \{0\} \quad \text{provided that} \quad m \ge CK^4 \gamma(T \cap S^{n-1})^2. \tag{11}$$

The conical version follows from the spherical version by expanding the sphere into a cone.

2.4 Sections of Sets by Random Subspaces: The M* Theorem

The M^* theorem [14, 15, 18] answers the following question: *Let T be an arbitrary subset of $\mathbb{R}^n$. What is the diameter of the intersection of a random subspace with T?* We may bound the radius of this intersection (which of course bounds the diameter) using our main results, and again for a general model of random subspaces.

Indeed, let us consider the kernel of an $m \times n$ random matrix A as in the previous section. By Theorem 3 (see (6)), we have

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leq CK^2\gamma(T) \tag{12}$$

with high probability. On the event that the above inequality holds, we may further restrict the supremum to $\ker A \cap T$, giving

$$\sup_{x \in \ker A \cap T} \sqrt{m}\|x\|_2 \leq CK^2\gamma(T).$$

The left-hand side is $\sqrt{m}$ times the radius of $T \cap \ker A$. Thus, with high probability,

$$\operatorname{rad}(\ker A \cap T) \leq \frac{CK^2\gamma(T)}{\sqrt{m}}. \tag{13}$$

This is a classical form of the so-called M^* estimate. It is typically used for sets T that contain the origin. In these cases, the Gaussian complexity $\gamma(T)$ can be replaced by Gaussian width $w(T)$. Indeed, (3) with $y = 0$ implies that these two quantities are equivalent.

2.5 The Size of Random Linear Images of Sets

Another question that can be addressed using our main results is how the size of a set T in $\mathbb{R}^n$ changes under the action of a random linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$. Applying (6) and the triangle inequality, we obtain

$$\operatorname{rad}(AT) \leq \sqrt{m} \cdot \operatorname{rad}(T) + CK^2\gamma(T) \tag{14}$$

with high probability. This result has been known for random projections, where $A = \sqrt{n}P$ and P is the orthogonal projection onto a random m-dimensional subspace in $\mathbb{R}^n$ drawn according to the Haar measure on the Grassmanian, see [2, Proposition 5.7.1].

It is also known that the bound (14) is sharp (up to absolute constant factor) even for random projections, see [2, Sect. 5.7.1]. This in particular implies optimality of the bound in our main result, Theorem 1.

2.6 Signal Recovery from the Constrained Linear Model

The constrained linear model is the backbone of many statistical and signal processing problems. It takes the form

$$y = Ax + z, \qquad x \in T, \tag{15}$$

where $x \in T \subset \mathbb{R}^n$ is unknown, $y \in \mathbb{R}^m$ is a vector of known observations, the measurement matrix $A \in \mathbb{R}^{m \times n}$ is known, and $z \in \mathbb{R}^m$ is unknown noise which can be either fixed or random and independent of A.

For example, in the statistical linear model, A is a matrix of explanatory variables, and x is a coefficient vector. It is common to assume, or enforce, that only a small percentage of the explanatory variables are significant. This is encoded by taking T to be the set of vectors with less than s non-zero entries, for some $s \leq n$. In other words, T encodes *sparsity*. In another example, y is a vector of MRI measurements [12], in which case x is the image to be constructed. Natural images have quite a bit of structure, which may be enforced by bounding the total variation, or requiring sparsity in a certain dictionary, each of which gives a different constraint set T. There are a plethora of other applications, with various constraint sets T, including low-rank matrices, low-rank tensors, non-negative matrices, and structured sparsity. In general, a goal of interest is to estimate x.

When T is a linear subspace, it is standard to estimate x via least squares regression, and the performance of such an estimator is well known. However, when T is non-linear, the problem can become quite complicated, both in designing a tractable method to estimate x and also analyzing the performance. The field of compressed sensing [5, 6] gives a comprehensive treatment of the case when T encodes sparsity, showing that convex programming can be used to estimate x, and that enforcing the sparse structure this way gives a substantial improvement over least squares regression. A main idea espoused in compressed sensing is that random matrices A give near optimal recovery guarantees.

Predating, but especially following, the works in compressed sensing, there have also been several works which tackle the general case, giving results for arbitrary T [1, 3, 11, 16, 17, 20, 21, 25]. The deviation inequalities of this paper allow for a general treatment as well. We will first show how to recover several known signal recovery results, and then give a new result in Sect. 2.7.

Consider the constrained linear model (15). A simple and natural way to estimate the unknown signal x is to solve the optimization problem

$$\hat{x} := \arg\min_{x' \in T} \|Ax' - y\|_2^2 \tag{16}$$

We note that depending on T, the constrained least squares problem (16) may be computationally tractable or intractable. We do not focus on algorithmic issues here, but just note that T may be replaced by a larger tractable set (e.g., convexified) to aid computation.

Our goal is to bound the Euclidean norm of the error

$$h := \hat{x} - x.$$

Since $\hat{x}$ minimizes the squared error, we have $\|A\hat{x} - y\|_2^2 \leq \|Ax - y\|_2^2$. Simplifying this, we obtain

$$\|Ah\|_2^2 \leq 2\langle h, A^T z\rangle. \tag{17}$$

We now proceed to control $\|h\|_2$ depending on the structure of T.

2.6.1 Exact Recovery

In the noiseless case where $z = 0$, inequality (17) simplifies and we have

$$h \in \ker A \cap (T - x). \tag{18}$$

(The second constraint here follows since $h = \hat{x} - x$ and $\hat{x} \in T$.)

In many cases of interest, $T - x$ is a cone, or is contained in a cone, which is called the *tangent cone* or *descent cone*. Gordon-type inequality (11) then implies that $h = 0$, and thus we have *exact recovery* $\hat{x} = x$, provided that the number of observations m significantly exceeds the Gaussian complexity of this cone: $m \geq CK^4\gamma((T - x) \cap S^{n-1})^2$.

For example, if x is a sparse vector with s non-zero entries, and T is an appropriately scaled ℓ_1 ball, then $T - x$ is contained in a tangent cone, D, satisfying $\gamma(D)^2 \leq Cs\log(n/s)$. This implies that $\hat{x} = x$ with high probability, provided $m \geq CK^4 s\log(n/s)$.

2.6.2 Approximate Recovery

In the cases where $T - x$ is not a cone or cannot be extended to a narrow cone (for example, when x lies in the interior of T), we can use the M^* Theorem for the analysis of the error. Indeed, combining (18) with (13), we obtain

$$\|h\|_2 \leq \frac{CK^2 w(T)}{\sqrt{m}}.$$

Here we also used that since $T - T$ contains the origin, we have $\gamma(T - T) \sim w(T)$ according to (3). In particular, this means that x can be estimated up to an additive error of ε in the Euclidean norm provided that the number of observations satisfies $m \geq CK^4 w(T)^2/\varepsilon^2$.

For a more detailed description of the M^* Theorem, Gordon's Escape Theorem, and their implications for the constrained linear model, see [28].

2.7 Model Selection for Constrained Linear Models

It is often unknown precisely what constraint set to use for the constrained linear model, and practitioners often experiment with different constraint sets to see which gives the best performance. This is a form of model selection. We focus on the case when the form of the set is known, but the scaling is unknown. For example, in compressed sensing, it is common to assume that x is *compressible*, i.e., that it can be well approximated by setting most of its entries to 0. This can be enforced by assuming that x belongs to a scaled ℓ_p ball for some $p \in (0, 1]$. However, generally it is not known what scaling to use for this ℓ_p ball.

Despite this need, previous theory concentrates on controlling the error for one fixed choice of the scaling. Thus, a practitioner who tries many different scalings cannot be sure that the error bounds will hold uniformly over all such scalings. In this subsection, we remove this uncertainty by showing that the error in constrained least squares can be controlled simultaneously for an infinite number of scalings of the constraint set.

Assume $x \in T$, but the precise scaling of T is unknown. Thus, x is estimated using a scaled version of T:

$$\hat{x}_\lambda := \arg\min_{x' \in \lambda T} \|Ax' - y\|_2^2, \qquad \lambda \geq 1. \tag{19}$$

The following corollary controls the estimation error.

Corollary 2 *Let T be a convex, symmetric set. Given $\lambda \geq 1$, let $\hat{x}_\lambda$ be the solution to (19). Let $h_\lambda := \hat{x}_\lambda - x$, let $v_\lambda = h_\lambda/(1+\lambda)$, and let $\delta = \|v_\lambda\|_2$. Then with probability at least 0.99, the following occurs. For every $\lambda \geq 1$,*

$$\delta \leq \frac{CK^2\gamma(T \cap \delta B_2^n)}{\sqrt{m}} + CK\sqrt{\frac{\gamma(T \cap \delta B_2^n) \cdot \|z\|_2}{m(1+\lambda)}}. \tag{20}$$

The corollary is proven using Theorem 5. To our knowledge, this corollary is new. It recovers previous results that only apply to a single, fixed λ, as in [11, 20]. It is known to be nearly minimax optimal for many constraint sets of interest and for stochastic noise term z, in which case $\|z\|_2$ would be replaced by its expected value [21].

The rather complex bound of Eq. (20) seems necessary in order to allow generality. To aid understanding, we specialize the result to a very simple set—a linear subspace—for which the behaviour of constrained least squares is well known, the scaling becomes irrelevant, and the result simplifies significantly. When T is a d-dimensional subspace, we may bound the Gaussian complexity as $\gamma(T \cap \delta B_2) \leq \delta\sqrt{d}$. Plugging in the bound on $\gamma(T \cap \delta B_2^n)$ into (20), substituting h_λ back in, and massaging the equation gives

$$\|h_\lambda\|_2^2 \leq CK^4 \cdot \frac{d\|z\|_2^2}{m^2} \qquad \text{as long as} \qquad m \geq CK^4 d.$$

If z is Gaussian noise with standard deviation σ, then it's norm concentrates around $\sqrt{m}\sigma$, giving (with high probability)

$$\|h_\lambda\|_2^2 \le CK^4 \cdot \frac{d\sigma^2}{m} \qquad \text{as long as} \qquad m \ge CK^4 d.$$

In other words, the performance of least squares is proportional to the noise level multiplied by the dimension of the subspace, and divided by the number of observations, m. This is well known.

In this corollary, for simplicity we assumed that T is convex and symmetric. Note that this already allows constraint sets of interest, such as the ℓ_1 ball. However, this assumption can be weakened. All that is needed is for $T - \lambda T$ to be contained in a scaled version of T, and to be star shaped. This also holds, albeit for more complex scalings, for arbitrary ℓ_p balls with $p > 0$.

Proof (of Corollary 2) For simplicity of notation, we assume $K \le 10$ (say), and absorb K into other constants. The general case follows the same proof. First note that $h_\lambda \in \lambda T - T$. Since T is convex and symmetric, we have $\lambda T - T \subset (1+\lambda)T$ and as $v_\lambda = h_\lambda/(1+\lambda)$, we get

$$v_\lambda \in T. \tag{21}$$

Moreover, (17) gives

$$\|Av_\lambda\|_2^2 \le \frac{\langle v_\lambda, A^T z\rangle}{1+\lambda}, \qquad v_\lambda \in T. \tag{22}$$

We will show that, with high probability, any vector v_λ satisfying (21) and (22) has a small norm, thus completing the proof. We will do this by upper bounding $\langle v_\lambda, A^T z\rangle$ and lower bounding $\|Av_\lambda\|_2$ by $\|v_\lambda\|_2$ minus a deviation term.

For the former goal, let $w := A^T z/\|z\|_2$. Recall that the noise vector z is fixed (and in case z random and independent of A, condition on z to make it fixed). Then w is a sub-gaussian vector with independent entries whose sub-gaussian norm is upper-bounded by a constant; see [27]. Thus, the random process $Z_x := \langle x, w\rangle$ has the sub-gaussian increments required in Theorem 4 (again, see [27]). By this theorem, with probability ≥ 0.995,

$$|Z_x| \le C\gamma(T \cap \|x\|_2 B_2^n) \qquad \text{for all } x \in T.$$

Let F be the 'good' event that the above equation holds.

To control $\|Av_\lambda\|_2$, consider the 'good' event G that

$$\|Ax\|_2 \ge \sqrt{m}\|x\|_2 - C\gamma(T \cap \|x\|_2 B_2^n) \qquad \text{for all } x \in T.$$

By Theorem 5, G holds with probability at least 0.995.

Now, suppose that both G and F hold (which occurs with probability at least 0.99 by the union bound). We will show that for every $\lambda > 1$, v_λ is controlled. The event G gives

$$\langle v_\lambda, A^T z \rangle \le C\gamma(T \cap \|v_\lambda\|_2 B_2^n) \cdot \|z\|_2.$$

The event F gives

$$\|Av_\lambda\|_2 \ge \sqrt{m}\|v_\lambda\|_2 - C\gamma(T \cap \|v_\lambda\|_2 B_2^n).$$

Taking square roots of both sides of (22) and plugging in these two inequalities gives (20). □

3 Comparison with Known Results

Several partial cases of our main results have been known. As we already mentioned, the special case of Theorem 1 where the entries of A have standard normal distribution follows from the main result of the paper by Schechtman [23].

Generalizing the result of [23], Klartag and Mendelson proved the following theorem.

Theorem 6 (Theorem 4.1 in [10]) *Let A be an isotropic, sub-gaussian random matrix as in (1), and let $T \subseteq S^{n-1}$. Assume that $w(T) \ge C'(K)$.[3] Then with probability larger than $1/2$,*

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m} \right| \le C(K) w(T). \tag{23}$$

Here $C'(K)$ and $C(K)$ may depend on K only.

A similar but slightly more informative statement follows from our main results. Indeed, Corollary 1 gives the same conclusion, but with explicit dependence on K (the sub-gaussian norms of the rows of A) as well as probability of success. Moreover, our general results, Theorems 1 and 3, do not require the set T to lie on the unit sphere.

Another related result was proved by S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann.

[3] This restriction is not explicitly mentioned in the statement of Theorem 4.1 in [10], but it is used in the proof. Indeed, this result is derived from their Theorem 1.3, which explicitly requires that $\gamma(T)$ be large enough. Without such requirement, Theorem 4.1 in [10] fails e.g. when T is a singleton, since in that case we have $w(T) = 0$.

Theorem 7 (Theorem 2.3 in [13]) *Let A be an isotropic, sub-gaussian random matrix as in (1), and T be a star-shaped subset of $\mathbb{R}^n$. Let $0 < \theta < 1$. Then with probability at least $1 - \exp(-c\theta^2 m/K^4)$ we have that all vectors $x \in T$ with*

$$\|x\|_2 \geq r^* := \inf\left\{\rho > 0 : \rho \geq CK^2\gamma\left(T \cap \rho \cdot S^{n-1}\right)/(\theta\sqrt{m})\right\}$$

satisfy

$$(1-\theta)\,\|x\|_2^2 \leq \frac{\|Ax\|_2^2}{m} \leq (1+\theta)\,\|x\|_2^2\,.$$

Applying our Theorem 3 to the bounded set $T \cap r^* \cdot S^{n-1}$ precisely implies Theorem 7 with the same failure probability (up to the values of the absolute constants c, C). Moreover, our Theorem 3 treats all $x \in T$ uniformly, whereas Theorem 7 works only for x with large norm.

Yet another relevant result was proved by Dirksen [4, Theorem 5.5]. He showed that the inequality

$$\begin{aligned}\left|\|Ax\|_2^2 - m\|x\|_2^2\right| \lesssim\ & K^2 w(T)^2 + \sqrt{m}K^2\,\mathrm{rad}(T)w(T) \\ & + u\sqrt{m}K^2\,\mathrm{rad}(T)^2 + u^2K^2\,\mathrm{rad}(T)^2 \end{aligned} \tag{24}$$

holds uniformly over $x \in T$ with probability at least $1 - \exp(-u^2)$. To compare with our results, one can see that Theorem 3 implies that, with the same probability,

$$\begin{aligned}\left|\|Ax\|_2^2 - m\|x\|_2^2\right| \lesssim\ & K^4 w(T)^2 + \sqrt{m}K^2\|x\|_2 w(T) \\ & + u\sqrt{m}K^2\,\mathrm{rad}(T)\|x\|_2 + uK^4\,\mathrm{rad}(T)w(T) + u^2K^4\,\mathrm{rad}(T)^2,\end{aligned}$$

which is stronger than (24) when $K = O(1)$ and $m \gtrsim n$, since then $\|x\|_2 \leq \mathrm{rad}(T)$ and $w(T) \lesssim \sqrt{m}\,\mathrm{rad}(T)$.

4 Preliminaries

4.1 Majorizing Measure Theorem, and Deduction of Theorems 1 and 3

As we mentioned in the Introduction, Theorems 1 and 3 follow from Theorem 2 via Talagrand's Majorizing Measure Theorem (and its high-probability counterpart). Let us state this theorem specializing to processes that are indexed by points in $\mathbb{R}^n$. For $T \subset \mathbb{R}^n$, let $\mathrm{diam}(T) := \sup_{x,y\in T} \|x - y\|_2$.

Theorem 8 (Majorizing Measure Theorem) *Consider a random process $(Z_x)_{x\in T}$ indexed by points x in a bounded set $T \subset \mathbb{R}^n$. Assume that the process has sub-gaussian increments, that is there exists $M \geq 0$ such that*

$$\|Z_x - Z_y\|_{\psi_2} \leq M\|x-y\|_2 \quad \text{for every } x, y \in T. \tag{25}$$

Then

$$\mathbb{E}\sup_{x,y\in T}|Z_x - Z_y| \leq CM\mathbb{E}\sup_{x\in T}\langle g, x\rangle,$$

where $g \sim N(0, I_n)$. Moreover, for any $u \geq 0$, the event

$$\sup_{x,y\in T}|Z_x - Z_y| \leq CM\big[\mathbb{E}\sup_{x\in T}\langle g, x\rangle + u\,\mathrm{diam}(T)\big]$$

holds with probability at least $1 - \exp(-u^2)$.

The first part of this theorem can be found e.g. in [24, Theorems 2.1.1, 2.1.5]. The second part, a high-probability bound, is borrowed from [4, Theorem 3.2].

Let us show how to deduce Theorems 1 and 3. According to Theorem 2, the random process $Z_x := \|Ax\|_2 - \sqrt{m}\|x\|_2$ satisfies the hypothesis (25) of the Majorizing Measure Theorem 8 with $M = CK^2$. Fix an arbitrary $y \in T$ and use the triangle inequality to obtain

$$\mathbb{E}\sup_{x\in T}|Z_x| \leq \mathbb{E}\sup_{x\in T}|Z_x - Z_y| + \mathbb{E}|Z_y|. \tag{26}$$

Majorizing Measure Theorem bounds the first term: $\mathbb{E}\sup_{x\in T}|Z_x - Z_y| \lesssim K^2 w(T)$. (We suppress absolute constant factors in this inequality and below.) The second term can be bounded more easily as follows: $\mathbb{E}|Z_y| \lesssim \|Z_y\|_{\psi_2} \lesssim K^2\|y\|_2$, where we again used Theorem 2 with $x = 0$. Using (3), we conclude that

$$\mathbb{E}\sup_{x\in T}|Z_x| \lesssim K^2(w(T) + \|y\|_2) \lesssim K^2\gamma(T),$$

as claimed in Theorem 1.

We now prove Theorem 3. Since adding 0 to a set does not change its radius, we may assume that $0 \in T$. Let $Z_x := \|Ax\|_2 - \sqrt{m}\,\|x\|_2$. Since $Z_0 = 0$, and since Z_x has sub-gaussian increments by Theorems 2, 8 gives that with probability at least $1 - \exp(-u^2)$,

$$\begin{aligned}\sup_{x\in T}|Z_x| = \sup_{x\in T}|Z_x - Z_0| &\lesssim K^2\big[\mathbb{E}\sup_{x\in T}\langle g, x\rangle + u\cdot\mathrm{diam}(T)\big] \\ &\lesssim K^2\big[\mathbb{E}\sup_{x\in T}\langle g, x\rangle + u\cdot\mathrm{rad}(T)\big].\end{aligned}$$

□

4.2 Sub-exponential Random Variables, and Bernstein's Inequality

Our argument will make an essential use of Bernstein's inequality for sub-exponential random variables. Let us briefly recall the relevant notions, which can be found, e.g., in [27]. A random variable Z is *sub-exponential* if its distribution is dominated by an exponential distribution. More formally, Z is sub-exponential if the Orlicz norm

$$\|Z\|_{\psi_1} := \inf\left\{K > 0 : \ \mathbb{E}\psi_1(|Z|/K) \leq 1\right\}$$

is finite, for the Orlicz function $\psi_1(x) = \exp(x) - 1$. Every sub-gaussian random variable is sub-exponential. Moreover, an application of Young's inequality implies the following relation for any two sub-gaussian random variables X and Y:

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}. \tag{27}$$

The classical Bernstein's inequality states that a sum of independent sub-exponential random variables is dominated by a mixture of sub-gaussian and sub-exponential distributions.

Theorem 9 (Bernstein-Type Deviation Inequality, See e.g. [27]) *Let $X_1, \ldots, X_m$ be independent random variables, which satisfy $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_1} \leq L$. Then*

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} X_i\right| > t\right\} \leq 2\exp\Big[-cm\min\Big(\frac{t^2}{L^2}, \frac{t}{L}\Big)\Big], \quad t \geq 0.$$

5 Proof of Theorem 2

Proposition 1 (Concentration of the Norm) *Let $X \in \mathbb{R}^m$ be a random vector with independent coordinates X_i that satisfy $\mathbb{E}X_i^2 = 1$ and $\|X_i\|_{\psi_2} \leq K$. Then*

$$\Big\|\|X\|_2 - \sqrt{m}\Big\|_{\psi_2} \leq CK^2.$$

Remark 3 If $\mathbb{E}X_i = 0$, this proposition follows from [22, Theorem 2.1], whose proof uses the Hanson-Wright inequality.

Proof Let us apply Bernstein's deviation inequality (Theorem 9) for the sum of independent random variables $\|X\|_2^2 - m = \sum_{i=1}^m (X_i^2 - 1)$. These random variables have zero means and sub-exponential norms

$$\|X_i^2 - 1\|_{\psi_1} \leq 2\|X_i^2\|_{\psi_1} \leq 2\|X_i\|_{\psi_2}^2 \leq 2K^2.$$

(Here we used a simple centering inequality which can be found e.g. in [27, Remark 5.18] and the inequality (27).) Bernstein's inequality implies that

$$\mathbb{P}\left\{\left|\|X\|_2^2 - m\right| > tm\right\} \le 2\exp\Big[-cm\min\Big(\frac{t^2}{K^4}, \frac{t}{K^2}\Big)\Big], \quad t \ge 0. \tag{28}$$

To deduce a concentration inequality for $\|X\|_2 - \sqrt{m}$ from this, let us employ the numeric bound $|x^2 - m| \ge \sqrt{m}\,|x - \sqrt{m}|$ valid for all $x \ge 0$. Using this together with (28) for $t = s/\sqrt{m}$, we obtain

$$\begin{aligned}\mathbb{P}\left\{\left|\|X\|_2 - \sqrt{m}\right| > s\right\} &\le \mathbb{P}\left\{\left|\|X\|_2^2 - m\right| > s\sqrt{m}\right\} \\ &\le 2\exp(-cs^2/K^4) \quad \text{for } s \le K^2\sqrt{m}.\end{aligned}$$

To handle large s, we proceed similarly but with a different numeric bound, namely $|x^2 - m| \ge (x - \sqrt{m})^2$ which is valid for all $x \ge 0$. Using this together with (28) for $t = s^2/m$, we obtain

$$\begin{aligned}\mathbb{P}\left\{\left|\|X\|_2 - \sqrt{m}\right| > s\right\} &\le \mathbb{P}\left\{\left|\|X\|_2^2 - m\right| > s^2\right\} \\ &\le 2\exp(-cs^2/K^2) \quad \text{for } s \ge K\sqrt{m}.\end{aligned}$$

Since $K \ge 1$, in both cases we bounded the probability in question by $2\exp(-cs^2/K^4)$. This completes the proof. □

Lemma 1 (Concentration of a Random Matrix on a Single Vector) *Let A be an isotropic, sub-gaussian random matrix as in* (1)*. Then*

$$\Big\|\|Ax\|_2 - \sqrt{m}\Big\|_{\psi_2} \le CK^2 \quad \text{for every } x \in S^{n-1}.$$

Proof The coordinates of the vector $Ax \in \mathbb{R}^m$ are independent random variables $X_i := \langle A_i, x\rangle$. The assumption that $\mathbb{E}A_iA_i^{\mathsf{T}} = I$ implies that $\mathbb{E}X_i^2 = 1$, and the assumption that $\|A_i\|_{\psi_2} \le K$ implies that $\|X_i\|_{\psi_2} \le K$. The conclusion of the lemma then follows from Proposition 1. □

Lemma 1 can be viewed as a partial case of the increment inequality of Theorem 2 for $x \in S^{n-1}$ and $y = 0$, namely

$$\|Z_x\|_{\psi_2} \le CK^2 \quad \text{for every } x \in S^{n-1}. \tag{29}$$

Our next intermediate step is to extend this by allowing y to be an arbitrary unit vector.

Lemma 2 (Sub-Gaussian Increments for Unit Vectors) *Let A be an isotropic, sub-gaussian random matrix as in* (1)*. Then*

$$\Big\|\|Ax\|_2 - \|Ay\|_2\Big\|_{\psi_2} \le CK^2\|x - y\|_2 \quad \text{for every } x, y \in S^{n-1}.$$

Proof Given $s \geq 0$, we will bound the tail probability

$$p := \mathbb{P}\left\{ \frac{\big|\|Ax\|_2 - \|Ay\|_2\big|}{\|x-y\|_2} > s \right\}. \tag{30}$$

Case 1: $s \geq 2\sqrt{m}$. Using the triangle inequality we have $|\|Ax\|_2 - \|Ay\|_2| \leq \|A(x-y)\|_2$. Denoting $u := (x-y)/\|x-y\|_2$, we find that

$$p \leq \mathbb{P}\left\{\|Au\|_2 > s\right\} \leq \mathbb{P}\left\{\|Au\|_2 - \sqrt{m} > s/2\right\} \leq \exp(-Cs^2/K^4).$$

Here the second bound holds since $s \geq 2\sqrt{m}$, and the last bound follows by Lemma 1.

Case 2: $s \leq 2\sqrt{m}$. Multiplying both sides of the inequality defining p in (30) by $\|Ax\|_2 + \|Ay\|_2$, we can write p as

$$p = \mathbb{P}\left\{|Z| > s\big(\|Ax\|_2 + \|Ay\|_2\big)\right\} \quad \text{where} \quad Z := \frac{\|Ax\|_2^2 - \|Ay\|_2^2}{\|x-y\|_2}.$$

In particular,

$$p \leq \mathbb{P}\left\{|Z| > s\|Ax\|_2\right\} \leq \mathbb{P}\left\{|Z| > \frac{s\sqrt{m}}{2}\right\} + \mathbb{P}\left\{\|Ax\|_2 \leq \frac{\sqrt{m}}{2}\right\} =: p_1 + p_2.$$

We may bound p_2 using Lemma 1:

$$p_2 \leq 2\exp\Big(-\frac{(\sqrt{m}/2)^2}{C^2K^4}\Big) = 2\exp\Big(-\frac{m}{4C^2K^4}\Big) \leq 2\exp\Big(-\frac{s^2}{16C^2K^4}\Big). \tag{31}$$

Next, to bound p_1, it will be useful to write Z as

$$Z = \frac{\langle A(x-y), A(x+y)\rangle}{\|x-y\|_2} = \langle Au, Av\rangle, \quad \text{where} \quad u := \frac{x-y}{\|x-y\|_2}, \quad v := x+y.$$

Since the coordinates of Au and Av are $\langle A_i, u\rangle$ and $\langle A_i, v\rangle$ respectively, Z can be represented as a sum of independent random variables:

$$Z = \sum_{i=1}^{m} \langle A_i, u\rangle \langle A_i, v\rangle. \tag{32}$$

Note that each of these random variables $\langle A_i, u\rangle \langle A_i, v\rangle$ has zero mean, since

$$\mathbb{E}\,\langle A_i, x-y\rangle \langle A_i, x+y\rangle = \mathbb{E}\big[\,\langle A_i, x\rangle^2 - \langle A_i, y\rangle^2\,\big] = 1 - 1 = 0.$$

(Here we used the assumptions that $\mathbb{E}A_i A_i^{\mathsf{T}} = I$ and $\|x\|_2 = \|y\|_2 = 1$.) Moreover, the assumption that $\|A_i\|_{\psi_2} \le K$ implies that $\| \langle A_i, u \rangle \|_{\psi_2} \le K\|u\|_2 = K$ and $\| \langle A_i, v \rangle \|_{\psi_2} \le K\|v\|_2 \le 2K$. Recalling inequality (27), we see that $\langle A_i, u \rangle \langle A_i, v \rangle$ are sub-exponential random variables with $\| \langle A_i, u \rangle \langle A_i, v \rangle \|_{\psi_1} \le CK^2$. Thus we can apply Bernstein's inequality (Theorem 9) to the sum of mean zero, sub-exponential random variables in (32), and obtain

$$p_1 = \mathbb{P}\left\{ |Z| > \frac{s\sqrt{m}}{2} \right\} \le 2\exp(-cs^2/K^4), \quad \text{since } s \le 2K^2\sqrt{m}.$$

Combining this with the bound on p_2 obtained in (31), we conclude that

$$p = p_1 + p_2 \le 2\exp(-cs^2/K^4).$$

This completes the proof. □

Finally, we are ready to prove the increment inequality in full generality, for all $x, y \in \mathbb{R}^n$.

Proof (of Theorem 2) Without loss of generality we may assume that $\|x\|_2 = 1$ and $\|y\|_2 \ge 1$. Consider the unit vector $\bar{y} := y/\|y\|_2$ and apply the triangle inequality to get

$$\|Z_x - Z_y\|_{\psi_2} \le \|Z_x - Z_{\bar{y}}\|_{\psi_2} + \|Z_{\bar{y}} - Z_y\|_{\psi_2} =: R_1 + R_2.$$

By Lemma 2, $R_1 \le CK^2\|x - \bar{y}\|_2$. Next, since $\bar{y}$ and y are collinear, we have $R_2 = \|\bar{y} - y\|_2 \cdot \|Z_{\bar{y}}\|_{\psi_2}$. Since $\bar{y} \in S^{n-1}$, inequality (29) states that $\|Z_{\bar{y}}\|_{\psi_2} \le CK^2$, and we conclude that $R_2 \le CK^2\|\bar{y} - y\|_2$. Combining the bounds on R_1 and R_2, we obtain

$$\|Z_x - Z_y\|_{\psi_2} \le CK^2\big(\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2\big).$$

It is not difficult to check that since $\|y\|_2 \ge 1$, we have $\|x - \bar{y}\|_2 \le \|x - y\|_2$ and $\|\bar{y} - y\|_2 \le \|x - y\|_2$. This completes the proof. □

6 Proof of Theorem 4

We will prove a slightly stronger statement. For $r > 0$, define

$$E_r := \sup_{x \in \frac{1}{r}T \cap B_2^n} |Z_x|.$$

Set $W := \lim_{r \to \mathrm{rad}(T)-} \gamma\left(\frac{1}{r}T \cap B_2^n\right)$. Since $\frac{1}{r}T \cap B_2^n$ contains at least one point on the boundary for every $r < \mathrm{rad}(T)$, it follows that $W \ge \sqrt{2/\pi}$. We will show that,

with probability at least $1 - \exp\left(-c't^2W^2\right)$, one has

$$E_r \leq t \cdot CK^2\gamma\left(\frac{1}{r}T \cap B_2^n\right) \text{ for all } r \in (0,\infty),$$

which, when combined with the assumption of homogeneity, will clearly imply the theorem with a stronger probability.

Fix $\varepsilon > 0$. Let $\varepsilon = r_0 < r_1 < \ldots < r_N$ be a sequence of real numbers satisfying the following conditions:

- $\gamma\left(\frac{1}{r_i}T \cap B_2^n\right) = 2 \cdot \gamma\left(\frac{1}{r_{i+1}}T \cap B_2^n\right)$ for $i = 0, 1, \ldots, N-1$, and
- $\gamma\left(\frac{1}{r_N}T \cap B_2^n\right) \leq 2 \cdot W$.

The quantities $r_1, \ldots, r_N$ exist since the map $r \mapsto \gamma\left(\frac{1}{r}T \cap B_2^n\right)$ is decreasing and continuous when T is star-shaped.

Applying the Majorizing Measure Theorem 8 to the set $\frac{1}{r}T \cap B_2^n$ and noting that $Z_0 = 0$, we obtain that

$$E_r \lesssim K^2\left[\gamma\left(\frac{1}{r}T \cap B_2^n\right) + u\right]$$

with probability at least $1 - \exp(-u^2)$. Set $c := 10 \cdot \sqrt{\frac{\pi}{2}} \geq 10/W$ and use the above inequality for $u = ct\gamma\left(\frac{1}{r}T \cap B_2^n\right)$. We get

$$E_r \lesssim t \cdot K^2\gamma\left(\frac{1}{r}T \cap B_2^n\right) \tag{33}$$

holds with probability at least $1 - \exp\left(-c^2t^2\gamma\left(\frac{1}{r}T \cap B_2^n\right)^2\right)$. Thus for each $i \in \{0, 1, \ldots, N\}$, we have

$$E_{r_i} \lesssim t \cdot K^2\gamma\left(\frac{1}{r_i}T \cap B_2^n\right) \tag{34}$$

with probability at least

$$1 - \exp\left(-c^2t^2 4^{N-i}\gamma\left(\frac{1}{r_N}T \cap B_2^n\right)^2\right) \geq 1 - \exp\left(-c^2t^2 4^{N-i}W^2\right).$$

By our choice of c and the union bound, (34) holds for all i simultaneously with probability at least

$$1 - \sum_{i=0}^{N} \exp\left(-c^2t^2 4^{N-i}W^2\right) \geq 1 - 2 \cdot \exp(-100t^2W^2) =: 1 - \exp(-c't^2W^2).$$

We now show that if (34) holds for all i, then (33) holds for all $r \in (\varepsilon, \infty)$. This is done via an approximation argument. To this end, assume that (34) holds and let $r \in (r_{i-1}, r_i)$ for some $i \in [N]$. Since T is star-shaped, we have $\frac{1}{r} T \cap B_2^n \subseteq \frac{1}{r_{i-1}} T \cap B_2^n$, so

$$E_r \leq E_{r_{i-1}} \lesssim t \cdot K^2 \gamma \left(\frac{1}{r_{i-1}} T \cap B_2^n \right) = 2t \cdot K^2 \gamma \left(\frac{1}{r_i} T \cap B_2^n \right) \leq 2t \cdot K^2 \gamma \left(\frac{1}{r} T \cap B_2^n \right).$$

Also, for $\mathrm{rad}(T) \geq r > r_N$ we have

$$E_r \lesssim t \cdot K^2 \gamma \left(\frac{1}{r_N} T \cap B_2^n \right) \leq 2t \cdot K^2 W \leq 2t \cdot K^2 \gamma \left(\frac{1}{r} T \cap B_2^n \right).$$

Let F_k be the event that (33) holds for all $r \in (1/k, \infty)$. We have just shown that $\mathbb{P}\{F_k\} \geq 1 - \exp\left(-c' t^2 W^2\right)$ for all $k \in \mathbb{N}$. As $F_1 \supseteq F_2 \supseteq \ldots$ and $\cap_k F_k =: F_\infty$ is the event that (33) holds for all $r \in (0, \infty)$, it follows by continuity of measure that $\mathbb{P}\{F_\infty\} \geq 1 - \exp\left(-c' t^2 W^2\right)$, thus completing the proof.

7 Further Thoughts

In the definition of Gaussian complexity $\gamma(T) = \mathbb{E} \sup_{x \in T} |\langle g, x \rangle|$, the absolute value is essential to make Theorem 1 hold. In other words, the bound would fail if we replace $\gamma(T)$ by the Gaussian width $w(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$. This can be seen by considering a set T that consists of a single point.

However, *one-sided* deviation inequalities do hold for Gaussian width. Thus a one-sided version of Theorem 1 states that

$$\mathbb{E} \sup_{x \in T} \left(\|Ax\|_2 - \sqrt{m} \|x\|_2 \right) \leq CK^2 \cdot w(T), \tag{35}$$

and the same bound holds for $\mathbb{E} \sup_{x \in T} \left(- \|Ax\|_2 + \sqrt{m} \|x\|_2 \right)$. To prove (35), one modifies the argument in Sect. 4.1 as follows. Fix a $y \in T$. Since $\mathbb{E}\|Ay\|_2 \leq \left(\mathbb{E}\|Ay\|_2^2\right)^{1/2} = \sqrt{m}\|y\|_2$, we have $\mathbb{E} Z_y \leq 0$, thus

$$\mathbb{E} \sup_{x \in T} Z_x \leq \mathbb{E} \sup_{x \in T} (Z_x - Z_y) \leq \mathbb{E} \sup_{x \in T} |Z_x - Z_y| \lesssim K^2 w(T)$$

where the last bound follows by Majorizing Measure Theorem 8. Thus in this argument there is no need to separate the term $\mathbb{E}|Z_y|$ as was done before in Eq. (26).

Acknowledgements Christopher Liaw is partially supported by an NSERC graduate scholarship. Abbas Mehrabian is supported by an NSERC Postdoctoral Fellowship. Yaniv Plan is partially supported by NSERC grant 22R23068. Roman Vershynin is partially supported by NSF grant DMS 1265782 and USAF Grant FA9550-14-1-0009.

References

1. D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Living on the edge: phase transitions in convex programs with random data. Inf. Inference **3**(3), 224–294 (2014). doi:10.1093/imaiai/iau005. http://dx.doi.org/10.1093/imaiai/iau005
2. S. Artstein-Avidan, A. Giannopoulos, V.D. Milman, Asymptotic geometric analysis. Part I, in *Mathematical Surveys and Monographs*, vol. 202 (American Mathematical Society, Providence, RI, 2015)
3. V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
4. S. Dirksen, Tail bounds via generic chaining. Electron. J. Probab. **20**(53), 1–29 (2015). doi:10.1214/EJP.v20-3760. http://dx.doi.org/10.1214/EJP.v20-3760
5. Y.C. Eldar, G. Kutyniok, *Compressed Sensing: Theory and Applications* (Cambridge University Press, Cambridge, 2012)
6. S. Foucart, H. Rauhut, A mathematical introduction to compressive sensing, in *Applied and Numerical Harmonic Analysis* (Birkhäuser/Springer, New York, 2013) doi:10.1007/978-0-8176-4948-7. http://dx.doi.org/10.1007/978-0-8176-4948-7
7. Y. Gordon, On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$, in *Geometric Aspects of Functional Analysis (1986/87)*. Lecture Notes in Mathematics, vol. 1317 (Springer, Berlin, 1988), pp. 84–106. doi:10.1007/BFb0081737. http://dx.doi.org/10.1007/BFb0081737
8. D. Gross, Y.K. Liu, S.T. Flammia, S. Becker, J. Eisert, Quantum state tomography via compressed sensing. Phys. Rev. Lett. **105**(15), 150,401 (2010)
9. W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. **26**(1), 189–206 (1984)
10. B. Klartag, S. Mendelson, Empirical processes and random projections. J. Funct. Anal. **225**(1), 229–245 (2005). doi:10.1016/j.jfa.2004.10.009. http://dx.doi.org/10.1016/j.jfa.2004.10.009
11. G. Lecué, S. Mendelson, Learning subgaussian classes: upper and minimax bounds (2013). Available at http://arxiv.org/abs/1305.4825
12. M. Lustig, D. Donoho, J.M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging. Magn. Reson. Med. **58**(6), 1182–1195 (2007)
13. S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Reconstruction and subgaussian operators in asymptotic geometric analysis. Geom. Funct. Anal. **17**(4), 1248–1282 (2007). doi:10.1007/s00039-007-0618-7. http://dx.doi.org/10.1007/s00039-007-0618-7
14. V.D. Milman, Geometrical inequalities and mixed volumes in the local theory of Banach spaces. Astérisque **131**, 373–400 (1985)
15. V.D. Milman, Random subspaces of proportional dimension of finite dimensional normed spaces: approach through the isoperimetric inequality, in *Banach Spaces* (Springer, Berlin, 1985), pp. 106–115
16. S. Oymak, C. Thrampoulidis, B. Hassibi, Simple bounds for noisy linear inverse problems with exact side information (2013). Available at http://arxiv.org/abs/1312.0641
17. S. Oymak, C. Thrampoulidis, B. Hassibi, The squared-error of generalized lasso: a precise analysis, in *51st Annual Allerton Conference on Communication, Control, and Computing*, IEEE (2013), pp. 1002–1009
18. A. Pajor, N. Tomczak-Jaegermann, Subspaces of small codimension of finite-dimensional banach spaces. Proc. Am. Math. Soc. **97**(4), 637–642 (1986)

19. Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inform. Theory **59**(1), 482–494 (2013). doi:10.1109/TIT.2012.2207945. http://dx.doi.org/10.1109/TIT.2012.2207945
20. Y. Plan, R. Vershynin, The generalized lasso with non-linear observations. IEEE Trans. Inform. Theory **62**(3), 1528–1537 (2016). doi:10.1109/TIT.2016.2517008
21. Y. Plan, R. Vershynin, E. Yudovina, High-dimensional estimation with geometric constraints (2014). Available at http://arxiv.org/abs/1404.3749
22. M. Rudelson, R. Vershynin, Hanson-Wright inequality and sub-Gaussian concentration. Electron. Commun. Probab. **18**(82), 1–9 (2013). doi:10.1214/ECP.v18-2865. http://dx.doi.org/10.1214/ECP.v18-2865
23. G. Schechtman, Two observations regarding embedding subsets of Euclidean spaces in normed spaces. Adv. Math. **200**(1), 125–135 (2006). doi:10.1016/j.aim.2004.11.003. http://dx.doi.org/10.1016/j.aim.2004.11.003
24. M. Talagrand, The generic chaining: upper and lower bounds of stochastic processes, in *Springer Monographs in Mathematics* (Springer, Berlin, 2005)
25. C. Thrampoulidis, S. Oymak, B. Hassibi, Simple error bounds for regularized noisy linear inverse problems, in *IEEE International Symposium on Information Theory (ISIT)*, IEEE (2014), pp. 3007–3011
26. R. Vershynin, How close is the sample covariance matrix to the actual covariance matrix? J. Theor. Probab. **25**(3), 655–686 (2012)
27. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing* (Cambridge University Press, Cambridge, 2012), pp. 210–268
28. R. Vershynin, Estimation in high dimensions: a geometric perspective, in *Sampling Theory, A Renaissance* (Birkhauser, Basel, 2015), pp. 3–66
29. J. Von Neumann, *Collected Works*, ed. by A.H. Taub (Pergamon, Oxford, 1961)