

ESTIMATION IN HIGH DIMENSIONS: A GEOMETRIC PERSPECTIVE

ROMAN VERSHYNIN

ABSTRACT. This tutorial paper provides an exposition of a flexible geometric framework for high dimensional estimation problems with constraints. The paper develops geometric intuition about high dimensional sets, justifies it with some results of asymptotic convex geometry, and demonstrates connections between geometric results and estimation problems. The theory is illustrated with applications to sparse recovery, matrix completion, quantization, linear and logistic regression and generalized linear models.

CONTENTS

1. Introduction	2
2. High dimensional estimation problems	4
3. An excursion into high dimensional convex geometry	6
4. From geometry to estimation: linear observations	14
5. High dimensional sections: proof of a general M^* bound	17
6. Consequences: estimation from noisy linear observations	19
7. Applications to sparse recovery and regression	21
8. Extensions from Gaussian to sub-gaussian distributions	26
9. Exact recovery	30
10. Low-rank matrix recovery and matrix completion	33
11. Single-bit observations via hyperplane tessellations	40
12. Single-bit observations via optimization, and applications to logistic regression	42
13. General non-linear observations via metric projection	46
14. Some extensions	49
References	51

Date: May 20, 2014.

Partially supported by NSF grant DMS 1265782 and USAF Grant FA9550-14-1-0009.

1. INTRODUCTION

1.1. Estimation with constraints. This paper provides an exposition of an emerging mathematical framework for high-dimensional *estimation problems with constraints*. In these problems, the goal is to estimate a point \mathbf{x} which lies in a certain known feasible set $K \subseteq \mathbb{R}^n$, from a small sample y_1, \dots, y_m of independent observations of \mathbf{x} . The point \mathbf{x} may represent a signal in signal processing, a parameter of a distribution in statistics, or an unknown matrix in problems of matrix estimation or completion. The feasible set K is supposed to represent properties that we know or want to impose on \mathbf{x} .

The geometry of the high dimensional set K is a key to understanding estimation problems. A powerful intuition about *what high dimensional sets look like* has been developed in the area known as *asymptotic convex geometry* [4, 22]. The intuition is supported by many rigorous results, some of which can be applied to estimation problems. The main goals of this paper are:

- (a) develop geometric intuition about high dimensional sets;
- (b) explain results of asymptotic convex geometry which validate this intuition;
- (c) demonstrate connections between high dimensional geometry and high dimensional estimation problems.

This paper is not a comprehensive survey but is rather a tutorial. It does not attempt to chart vast territories of high dimensional inference that lie on the interface of statistics and signal processing. Instead, this paper proposes a useful geometric viewpoint, which could help us find a common mathematical ground for many (and often dissimilar) estimation problems.

1.2. Quick examples. Before we proceed with a general theory, let us mention some concrete examples of estimation problems that will be covered here. A particular class of estimation problems with constraints is considered in the young field of *compressed sensing* [16, 36, 12, 32]. There K is supposed to represent *sparsity*, thus K usually consists of vectors that have few non-zero coefficients. Sometimes more restrictive *structured sparsity* assumptions are placed, where only certain arrangements of non-zero coefficients are allowed [5, 58]. The observations y_i in compressed sensing are assumed to be *linear* in \mathbf{x} , which means that $y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle$. Here \mathbf{a}_i are typically i.i.d. vectors drawn from some known distribution in \mathbb{R}^n (for example, normal).

Another example of estimation problems with constraints is the *matrix completion problem* [8, 9, 34, 29, 62, 59] where K consists of matrices with *low rank*, and y_1, \dots, y_m is a sample of matrix entries. Such observations are still linear in \mathbf{x} .

In general, observations do not have to be linear; good examples are *binary observations* $y_i \in \{-1, 1\}$, which satisfy $y_i = \text{sign}(\langle \mathbf{a}_i, \mathbf{x} \rangle)$, see [7, 33, 54, 55], and more generally $\mathbb{E} y_i = \theta(\langle \mathbf{a}_i, \mathbf{x} \rangle)$, see [56, 2, 57].

In statistics, these classes of estimation problems can be interpreted as *linear regression* (for linear observations with noise), *logistic regression* (for binary observations) and *generalized linear models* (for more general non-linear observations).

All these examples, and more, will be explored in this paper. However, our main goal is to advance a general approach, which would not be tied to a particular nature of the feasible set K . Some general estimation problems of this nature were considered in [49, 3] for linear observations and in [56, 55, 2, 57] for non-linear observations.

1.3. Plan of the paper. In Section 2.1, we introduce a general class of estimation problems with constraints. We explain how the constraints (given by feasible set K) represent *low-complexity structures*, which could make it possible to estimate \mathbf{x} from few observations.

In Section 3, we make a short excursion into the field of *asymptotic convex geometry*. We explain intuitively the shape of high-dimensional sets K and state some known results supporting this intuition. In view of estimation problems, we especially emphasize one of these results – the so-called M^* bound on the size of high-dimensional sections of K by a random subspace E . It depends on the single geometric parameter of K that quantifies the complexity of K ; this quantity is called the *mean width*. We discuss mean width in some detail, pointing out its connections to convex geometry, stochastic processes, and statistical learning theory.

In Section 4 we apply the M^* bound to the general estimation problem with linear observations. We formulate an estimator first as a convex feasibility problem (following [49]) and then as a convex optimization problem.

In Section 5 we prove a general form of the M^* bound. Our proof borrowed from [55] is quite simple and instructive. Once the M^* bound is stated in the language of stochastic processes, it follows quickly by application of symmetrization, contraction and rotation invariance.

In Section 6, we apply the general M^* bound to estimation problems; observations here are still linear but can be noisy. Examples of such problems include *sparse recovery* problems and *linear regression* with constraints, which we explore in Section 7.

In Section 8, we extend the theory from Gaussian to sub-gaussian observations. A sub-gaussian M^* bound (similar to the one obtained in [49]) is deduced from the previous (Gaussian) argument followed by an application of a deep comparison theorem of X. Fernique and M. Talagrand (see [64]).

In Section 9 we pass to *exact recovery* results, where an unknown vector \mathbf{x} can be inferred from the observations y_i without any error. We present a simple geometric argument based on Y. Gordon’s “escape through a mesh”

theorem [28]. This argument was first used in this context in [60] and was pushed forward for general feasible sets in [13].

In Section 10, we explore matrix estimation problems. We first show how the general theory applies to a *low-rank matrix recovery* problem. Then we address a *matrix completion* problem with a short and self-contained argument from [57].

Finally, we pass to non-linear observations. In Section 11, we consider *single-bit observations* $y_i = \text{sign} \langle \mathbf{a}_i, \mathbf{x} \rangle$. Analogously to linear observations, there is a clear geometric interpretation for these as well. Namely, the estimation problem reduces in this case to a *pizza cutting problem* about random hyperplane tessellations of K . We discuss a result from [55] on this problem, and we apply it to estimation by formulating it as a feasibility problem.

Similarly to what we did for linear observations, we replace the feasibility problem by optimization problem in Section 12. Unlike before, such replacement is not trivial. We present a simple and self-contained argument from [56] about estimation from single-bit observations via convex optimization.

In Section 13 we discuss the estimation problem for general (not only single-bit) observations following [57]. The new crucial new step of estimation is the metric projection onto the feasible set; this projection was studied recently in [14] and [57].

In Section 14, we outline some natural extensions of the results for general distributions and to a localized version of mean width.

1.4. Acknowledgements. The author is grateful to Yaniv Plan for useful comments and to Vladimir Koltchinskii for helpful discussions.

2. HIGH DIMENSIONAL ESTIMATION PROBLEMS

2.1. Estimating vectors from random observations. Suppose we want to estimate an unknown vector $\mathbf{x} \in \mathbb{R}^n$. In signal processing, \mathbf{x} could be a signal to be reconstructed, while in statistics \mathbf{x} may represent a parameter of a distribution. We assume that information about \mathbf{x} comes from a sample of independent and identically distributed observations $y_1, \dots, y_m \in \mathbb{R}$, which are drawn from a certain distribution which depends on \mathbf{x} :

$$y_i \sim \text{distribution}(\mathbf{x}), \quad i = 1, \dots, m.$$

So, we want to estimate $\mathbf{x} \in \mathbb{R}^n$ from the observation vector

$$\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m.$$

One example of this situation is the classical linear regression problem in statistics,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\nu}, \tag{2.1}$$

in which one wants to estimate the coefficient vector $\boldsymbol{\beta}$ from the observation vector \mathbf{y} . We will see many more examples later; for now let us continue with setting up the general mathematical framework.

2.2. Low complexity structures. It often happens that we know in advance, believe in, or want to enforce, some properties of the vector \mathbf{x} . We can formalize such extra information as the assumption that

$$\mathbf{x} \in K$$

where K is some fixed and known subset of \mathbb{R}^n , a *feasible set*. This is a very general and flexible assumption, as we are not stipulating any properties of the feasible set K .

To give a quick example, in the regression problem (2.1), one often believes that β is a sparse vector, i.e. among its coefficients only few are non-zero. This is important because it means that a few explanatory variables can adequately explain the dependent variable. So one could choose K to be a set of all s -sparse vectors in \mathbb{R}^n – those with at most s non-zero coordinates, for a fixed sparsity level $s \leq n$. More examples of natural feasible sets K will be given later.

Figure 1 illustrates the estimation problem. Sampling can be thought as of a map taking $\mathbf{x} \in K$ to $\mathbf{y} \in \mathbb{R}^m$; estimation is a map from $\mathbf{y} \in \mathbb{R}^m$ to $\hat{\mathbf{x}} \in K$ and is ideally the inverse of sampling.

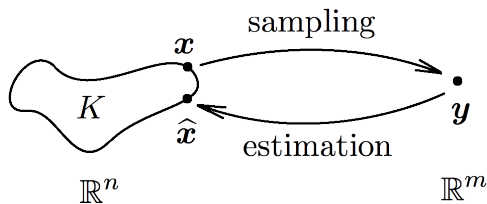


Figure 1. Estimation problem in high dimensions

How can a prior information encoded by K help in high-dimensional estimation? Let us start with a quick and non-rigorous argument based on the number of degrees of freedom. The unknown vector \mathbf{x} has n dimensions and the observation vector \mathbf{y} has m dimensions. So in principle, it should be possible to estimate \mathbf{x} from \mathbf{y} with

$$m = O(n)$$

observations. Moreover, this bound should be tight in general.

Now let us add the restriction that $\mathbf{x} \in K$. If K happens to be *low-dimensional*, with algebraic dimension $\dim(K) = d \ll n$, then \mathbf{x} has d degrees of freedom. Therefore, in this case the estimation should be possible with fewer observations,

$$m = O(d) = o(n).$$

It rarely happens that feasible sets of interest literally have small algebraic dimension. For example, the set of all s -sparse vectors in \mathbb{R}^n has full dimension n . Nevertheless, the intuition about low-dimensionality remains valid. Natural feasible sets, such as regression coefficient vectors, images, adjacency matrices of networks, do tend to have *low complexity*. Formally K

may live in an n -dimensional space where n can be very large, but the actual complexity of K , or “effective dimension” (which will formally quantify in Section 3.5.6) is often much smaller.

This intuition motivates the following three goals, which we will discuss in detail in this tutorial:

1. Quantify the complexity of general subsets K of \mathbb{R}^n .
2. Demonstrate that estimation can be done with few observations as long as the feasible set K has low complexity.
3. Design estimators that are algorithmically efficient.

We will start by developing intuition about the geometry of sets K in high dimensions. This will take us a short excursion into high dimensional convex geometry. Convexity assumption will not be imposed later, but it is going to be useful for developing a geometric intuition.

3. AN EXCURSION INTO HIGH DIMENSIONAL CONVEX GEOMETRY

High dimensional convex geometry studies convex subsets K of \mathbb{R}^n for large n . This area is sometimes also called asymptotic convex geometry (referring to n increasing to infinity) and geometric functional analysis. The tutorial [4] could be an excellent first contact with this area; the survey [26] and books [50, 53, 22] cover more material and in more depth.

3.1. What do high dimensional convex bodies look like? A central problem in high dimensional convex geometry is – *what do convex bodies look like in high dimensions?* A heuristic answer to this question is – a convex body K usually consists of a *bulk* and *outliers*. The bulk makes up most of the volume of K , but it is usually small in diameter. The outliers contribute little to the volume, but they are large in diameter.

If K is properly scaled, the bulk usually looks like a Euclidean ball. The outliers look like thin, long tentacles. This is best seen on Figure 2a, which depicts V. Milman’s vision of high dimensional convex sets [48]. This picture does not look convex, and there is a good reason for this. The volume in high dimensions scales differently than in low dimensions – dilating of a set by the factor 2 increases its volume by the factor 2^n . This is why it is not surprising that the tentacles contain exponentially less volume than the bulk. Such behavior is best seen if a picture looks “hyperbolic”. Although not convex, pictures like Figure 2 more accurately reflect the distribution of volume in higher dimensions.

Example 3.1 (The ℓ_1 ball). To illustrate this heuristic on a concrete example, consider the set

$$K = B_1^n = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\},$$

i.e. the unit ℓ_1 -ball in \mathbb{R}^n . The inscribed Euclidean ball in K , which we will denote by B , has diameter $2/\sqrt{n}$. One can then check that volumes of B

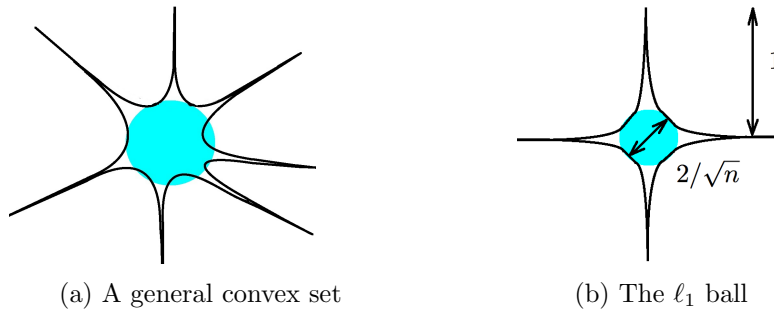


Figure 2. V. Milman’s “hyperbolic” drawings of high dimensional convex sets

and of K are comparable:

$$\text{vol}_n(B)^{1/n} \asymp \text{vol}_n(K)^{1/n} \asymp \frac{1}{n}.$$

Therefore, B (perhaps inflated by a constant factor) forms the bulk of K . It is round, makes up most of the volume of K , but has small diameter. The outliers of K are thin and long tentacles protruding quite far in the coordinate directions. This can be best seen in a hyperbolic drawing, see Figure 2b.

3.2. Concentration of volume. The heuristic representation of convex bodies just described can be supported by some rigorous results about *concentration of volume*.

These results assume that K is *isotropic*, which means that the random vector X distributed uniformly in K has zero mean and identity covariance:

$$\mathbb{E} X = 0, \quad \mathbb{E} X X^\top = I_n. \quad (3.1)$$

Isotropy is just an assumption of proper scaling – one can always make a convex body K with nonempty interior isotropic by applying a suitable invertible linear transformation.

With this scaling, most of the volume of K is located around the Euclidean sphere of radius \sqrt{n} . Indeed, taking traces of both sides of the second equation in (3.1), we obtain

$$\mathbb{E} \|X\|_2^2 = n.$$

Therefore, by Markov’s inequality, at least 90% of the volume of K is contained in a Euclidean ball of size $O(\sqrt{n})$. Much more powerful concentration results are known – the bulk of K lies very near the sphere of radius \sqrt{n} , and the outliers have exponentially small volume. This is the content of the two major results in high dimensional convex geometry, which we summarize in the following theorem.

Theorem 3.2 (Distribution of volume in high-dimensional convex sets). *Let K be an isotropic convex subset of \mathbb{R}^n , and let X be a random vector uniformly distributed in K . Then the following is true:*

1. (Concentration of volume) *For every $t \geq 1$, one has*

$$\mathbb{P} \{ \|X\|_2 > t\sqrt{n} \} \leq \exp(-ct\sqrt{n}).$$

2. (Thin shell) *For every $\varepsilon \in (0, 1)$, one has*

$$\mathbb{P} \left\{ \left| \|X\|_2 - \sqrt{n} \right| > \varepsilon\sqrt{n} \right\} \leq C \exp(-c\varepsilon^3 n^{1/2}).$$

Here and later in this paper, C, c denote positive absolute constants.

The concentration part of Theorem 3.2 is due to G. Paouris [51]; see [1] for an alternative and shorter proof. The thin shell part is an improved version of a result of B. Klartag [35], which is due to O. Guedon and E. Milman [30].

3.3. Low dimensional random sections. The intuition about bulk and outliers of high dimensional convex sets K can help us to understand what *random sections* of K should look like. Suppose E is a random subspace of \mathbb{R}^n with fixed dimension d , i.e. E is drawn at random from the Grassmanian manifold $G_{n,d}$ according to the Haar measure. What does the section $K \cap E$ look like on average?

If d is sufficiently small, then we should expect E to pass through the bulk of K and miss the outliers, as those have very small volume. Thus, if the bulk of K is a round ball,¹ we should expect the section $K \cap E$ to be a *round ball* as well; see Figure 3.

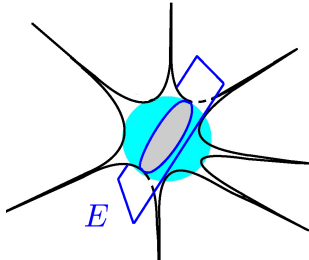


Figure 3. Random section of a high dimensional convex set

There is a rigorous result which confirms this intuition. It is known as Dvoretzky's theorem [20, 21], which we shall state in the form of V. Milman

¹This intuition is a good approximation to truth, but it should be corrected. While concentration of volume tells us that the bulk is *contained* in a certain Euclidean ball (and even in a thin spherical shell), it is not always true that the bulk *is* a Euclidean ball (or shell); a counterexample is the unit cube $[-1, 1]^n$. In fact, the cube is the worst convex set in the Dvoretzky theorem, which we are about to state.

[45]; expositions of this result can be found e.g. in [53, 22]. Dvoretzky-Milman's theorem has laid a foundation for the early development of asymptotic convex geometry. Informally, this result says that random sections of K of dimension $d \sim \log n$ are round with high probability.

Theorem 3.3 (Dvoretzky's theorem). *Let K be an origin-symmetric convex set in \mathbb{R}^n such that the ellipsoid of maximal volume contained in K is the unit Euclidean ball B_2^n . Fix $\varepsilon \in (0, 1)$. Let E be a random subspace of dimension $d = c\varepsilon^{-2} \log n$ drawn from the Grassmanian $G_{n,d}$ according to the Haar measure. Then there exists $R \geq 0$ such that with high probability (say, 0.99) we have*

$$(1 - \varepsilon) B(R) \subseteq K \cap E \subseteq (1 + \varepsilon) B(R).$$

Here $B(R)$ is the centered Euclidean ball of radius R in the subspace E .

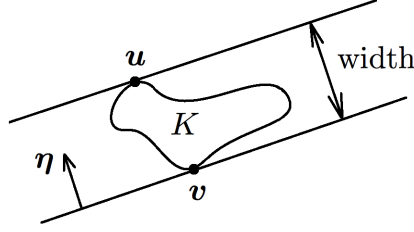
Several important aspects of this theorem are not mentioned here – in particular how, for a given convex set K , to compute the radius R and the largest dimension d of round sections of K . These aspects can be found in modern treatments of Dvoretzky theorem such as [53, 22].

3.4. High dimensional random sections? Dvoretzky's Theorem 3.3 describes the shape of *low* dimensional random sections $K \cap E$, those of dimensions $d \sim \log n$. Can anything be said about *high* dimensional sections, those with small codimension? In this more difficult regime, we can no longer expect such sections to be round. Instead, as the codimension decreases, the random subspace E becomes larger and it will probably pick more and more of the outliers (tentacles) of K . The shape of such sections $K \cap E$ is difficult to describe.

Nevertheless, it turns out that we can accurately predict the *diameter* of $K \cap E$. A bound on the diameter is known in asymptotic convex geometry as the low M^* estimate, or M^* bound. We will state this result in Section 3.6 and prove it in Section 5. For now, let us only mention that M^* bound is particularly attractive in applications as it depends only on two parameters – the codimension of E and a single geometric quantity, which informally speaking, measures the size of the bulk of K . This geometric quantity is called the *mean width* of K . We will pause briefly to discuss this important notion.

3.5. Mean width. The concept of mean width captures important geometric characteristics of sets in \mathbb{R}^n . One can mentally place it in the same category as other classical geometric quantities like volume and surface area.

Consider a bounded subset K in \mathbb{R}^n . (The convexity requirement will not be imposed from now on.) The *width* of K in the direction of a given unit vector $\eta \in S^{n-1}$ is defined as the width of the smallest slab between two parallel hyperplanes with normals η that contains K ; see Figure 4.

Figure 4. Width of K in the direction of η

Analytically, we can express the width in the direction of η as

$$\sup_{u \in K} \langle \eta, u - v \rangle = \sup_{z \in K - K} \langle \eta, z \rangle,$$

where $K - K = \{u - v : u, v \in K\}$ is the Minkowski sum of K and $-K$. Averaging over η uniformly distributed on the sphere S^{n-1} , we can define the *spherical mean width* of K :

$$\tilde{w}(K) := \mathbb{E} \sup_{z \in K - K} \langle \eta, z \rangle.$$

It will be convenient to replace the *spherical* random vector $\eta \sim \text{Unif}(S^{n-1})$ by the standard *Gaussian* random vector $g \sim N(0, I_n)$; the advantage is that g has independent coordinates while η does not.

Definition 3.4 (Gaussian mean width). *The Gaussian mean width of a bounded subset K of \mathbb{R}^n is defined as*

$$w(K) := \mathbb{E} \sup_{u \in K - K} \langle g, u \rangle, \quad (3.2)$$

where $g \sim N(0, I_n)$ is a standard Gaussian random vector in \mathbb{R}^n . We will often refer to Gaussian mean width as simply the mean width.

3.5.1. Simple properties of mean width. Observe first that the Gaussian mean width is about \sqrt{n} times larger than the spherical mean width. To see this, using rotation invariance we realize η as $\eta = g/\|g\|_2$. Next, we recall that the direction and magnitude of a standard Gaussian random vector are independent, so η is independent of $\|g\|_2$. It follows that

$$w(K) = \mathbb{E} \|g\|_2 \cdot \tilde{w}(K).$$

Further, the factor $\mathbb{E} \|g\|_2$ is of order \sqrt{n} ; this follows, for example, from known bounds on the χ^2 distribution:

$$c\sqrt{n} \leq \mathbb{E} \|g\|_2 \leq \sqrt{n} \quad (3.3)$$

where $c > 0$ is an absolute constant. Therefore, the Gaussian and spherical versions of mean width are equivalent (up to scaling factor \sqrt{n}), so it is mostly a matter of personal preference which version to work with. In this paper, we will mostly work with the Gaussian version.

Proposition 3.5. *The mean width is invariant under translations, orthogonal transformations, and taking convex hulls. \square*

Especially useful is the last property, which states that

$$w(\text{conv}(K)) = w(K). \quad (3.4)$$

This property will come handy later, when we consider convex relaxations of optimization problems.

3.5.2. Computing mean width on examples. Let us illustrate the notion of mean width on some simple examples.

Example 3.6. If K is the unit Euclidean ball B_2^n or sphere S^{n-1} , then

$$w(K) = \mathbb{E} \|\mathbf{g}\|_2 \leq \sqrt{n}$$

and also $w(K) \geq c\sqrt{n}$, by (3.3).

Example 3.7. Let K is a subset of B_2^n and it has *linear algebraic dimension* d . Then K lies in a d -dimensional unit Euclidean ball, so as before we have

$$w(K) \leq \sqrt{d}.$$

Example 3.8. Let K is a *finite subset* of B_2^n . Then

$$w(K) \leq C\sqrt{\log |K|}.$$

This follows from a known and simple computation of the expected maximum of $k = |K|$ Gaussian random variables.

Example 3.9 (Sparsity). Let K consist of all unit s -sparse vectors in \mathbb{R}^n – those with at most s non-zero coordinates:

$$K = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1, \|\mathbf{x}\|_0 \leq s\}.$$

Here $\|\mathbf{x}\|_0$ denotes the number of non-zero coordinates of \mathbf{x} . A simple computation (see e.g. [56, Lemma 2.3]) shows that

$$c\sqrt{s \log(n/s)} \leq w(K) \leq C\sqrt{s \log(n/s)}.$$

Example 3.10 (Low rank). Let K consist of $d_1 \times d_2$ matrices with unit Frobenius norm and rank at most r :

$$K = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F = 1, \text{rank}(X) \leq r\}.$$

We will see in Proposition 10.4,

$$w(K) \leq C\sqrt{r(d_1 + d_2)}.$$

3.5.3. Computing mean width algorithmically. Can we estimate the mean width of a given set K fast and accurately? Gaussian concentration of measure (see [53, 40, 39]) implies that, with high probability, the random variable

$$w(K, \mathbf{g}) = \sup_{\mathbf{u} \in K} \langle \mathbf{g}, \mathbf{u} \rangle$$

is close to its expectation $w(K)$. Therefore, to estimate $w(K)$, it is enough to generate a single realization of a random vector $\mathbf{g} \sim N(0, I_n)$ and compute $w(K, \mathbf{g})$; this should produce a good estimator of $w(K)$.

Since we can convexify K without changing the mean width by Proposition 3.5, computing this estimator is a *convex optimization problem* (and often even a linear problem if K is a polytope).

3.5.4. Computing mean width theoretically. Finding theoretical estimates on the mean width of a given set K is a non-trivial problem. It has been extensively studied in the areas of probability in Banach spaces and stochastic processes.

Two classical results in the theory of stochastic processes – Sudakov’s inequality (see [40, Theorem 3.18]) and Dudley’s inequality (see [40, Theorem 11.17]) – relate the mean width to the metric entropy of K . Let $N(K, t)$ denote the smallest number of Euclidean balls of radius t whose union covers K . Usually $N(K, t)$ is referred to as a *covering number* of K , and $\log N(K, t)$ is called the *metric entropy* of K .

Theorem 3.11 (Sudakov’s and Dudley’s inequalities). *For any bounded subset K of \mathbb{R}^n , we have*

$$c \sup_{t>0} t \sqrt{\log N(K, t)} \leq w(K) \leq C \int_0^\infty \sqrt{\log N(K, t)} dt.$$

The lower bound is Sudakov’s inequality and the upper bound is Dudley’s inequality.

Neither Sudakov’s nor Dudley’s inequality are tight for all sets K . A more advanced method of *generic chaining* produces a tight (but also more complicated) estimate of the mean width in terms of *majorizing measures*; see [64].

Let us only mention some other known ways to control mean width. In some cases, *comparison inequalities* for Gaussian processes can be useful, especially Slepian’s and Gordon’s; see [40, Section 3.3]. There is also a combinatorial approach to estimating the mean width and metric entropy, which is based on *VC-dimension* and its generalizations; see [41, 44].

3.5.5. Mean width and Gaussian processes. The theoretical tools of estimating mean width we just mentioned, including Sudakov’s, Dudley’s, Slepian’s and Gordon’s inequalities, have been developed in the context of stochastic processes. To see the connection, consider the Gaussian random variables $G_{\mathbf{u}} = \langle \mathbf{g}, \mathbf{u} \rangle$ indexed by points $\mathbf{u} \in \mathbb{R}^n$. The collection of these random variables $(G_{\mathbf{u}})_{\mathbf{u} \in K-K}$ forms a *Gaussian process*, and the mean width measures the size of this process:

$$w(K) = \mathbb{E} \sup_{\mathbf{u} \in K-K} G_{\mathbf{u}}.$$

In some sense, any Gaussian process can be approximated by a process of this form. We will return to the connection between mean width and Gaussian processes in Section 5 where we prove the M^* bound.

3.5.6. Mean width, complexity and effective dimension. In the context of stochastic processes, Gaussian mean width (and its non-gaussian variants) play an important role in statistical learning theory. There it is more natural to work with classes \mathcal{F} of real-valued functions on $\{1, \dots, n\}$ than with geometric sets $K \subseteq \mathbb{R}^n$. (We identify a vector in \mathbb{R}^n with a function on $\{1, \dots, n\}$.) The Gaussian mean width serves as a measure of complexity of a function class in statistical learning theory, see [42]. It is sometimes called *Gaussian complexity* and is usually denoted $\gamma_2(\mathcal{F})$.

To get a better feeling of mean width as complexity, assume that K lies in the unit Euclidean ball B_2^n . The square of the mean width, $w(K)^2$, may be interpreted as the *effective dimension* of K . By Example 3.7, the effective dimension is always bounded by the linear algebraic dimension. However, unlike algebraic dimension, the effective dimension is *robust* – a small perturbation of K leads to a small change in $w(K)^2$.

3.6. Random sections of small codimension: M^* bound. Let us return to the problem we posed in Section 3.4 – bounding the diameter of random sections $K \cap E$ where E is a high-dimensional subspace. The following important result in asymptotic convex geometry gives a good answer to this question.

Theorem 3.12 (M^* bound). *Let K be a bounded subset of \mathbb{R}^n . Let E be a random subspace of \mathbb{R}^n of a fixed codimension m , drawn from the Grassmannian $G_{n,n-m}$ according to the Haar measure. Then*

$$\mathbb{E} \text{diam}(K \cap E) \leq \frac{Cw(K)}{\sqrt{m}}.$$

We will prove a stronger version of this result in Section 5. The first variant of M^* bound was found by V. Milman [46, 47]; its present form is due to A. Pajor and N. Tomczak-Jaegermann [52]; an alternative argument which yields tight constants was given by Y. Gordon [28]; an exposition of M^* bound can be found in [53, 40].

To understand the M^* bound better, it is helpful to recall from Section 3.5.1 that $w(K)/\sqrt{n}$ is equivalent to the *spherical* mean width of K . Heuristically, the spherical mean width measures the *size of the bulk* of K .

For subspaces E of not very high dimension, where $m = \Omega(n)$, the M^* bound states that the size of the random section $K \cap E$ is bounded by the spherical mean width of K . In other words, subspaces E of proportional dimension *passes through the bulk* of K and ignores the outliers (“tentacles”), just as Figure 3 illustrates. But when the dimension of the subspace E grows toward n (so the codimension m becomes small), the diameter of $K \cap E$ also grows by a factor of $\sqrt{n/m}$. This gives a precise control of how E in this case *interferes with the outliers* of K .

4. FROM GEOMETRY TO ESTIMATION: LINEAR OBSERVATIONS

Having completed the excursion into geometry, we can now return to the high-dimensional estimation problems that we started to discuss in Section 2. To recall, our goal is to estimate an unknown vector

$$\mathbf{x} \in K \subseteq \mathbb{R}^n$$

that lies in a known feasible set K , from a random observation vector

$$\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m,$$

whose coordinates y_i are random i.i.d. observations of x .

So far, we have not been clear about possible distributions of the observations y_i . In this section, we will study perhaps the simplest model – *Gaussian linear observations*. Consider i.i.d. standard Gaussian vectors

$$\mathbf{a}_i \sim N(0, I_n)$$

and define

$$y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle, \quad i = 1, \dots, m.$$

Thus the observation vector \mathbf{y} depends linearly on \mathbf{x} . This is best expressed in a matrix form:

$$\mathbf{y} = A\mathbf{x}.$$

Here A is an $m \times n$ Gaussian random matrix, which means that the entries of A are i.i.d. $N(0, 1)$ random variables; the vectors \mathbf{a}_i form the rows of A .

The interesting regime is when the number of observations is smaller than the dimension, i.e. when $m < n$. In this regime, the problem of estimating $\mathbf{x} \in \mathbb{R}^n$ from $\mathbf{y} \in \mathbb{R}^m$ is ill posed. (In the complementary regime, where $m \geq n$, the linear system $\mathbf{y} = A\mathbf{x}$ is well posed, so the solution is trivial.)

4.1. Estimation based on M^* bound. Recall that we know two pieces of information about \mathbf{x} :

1. \mathbf{x} lies in a known random affine subspace $\{\mathbf{x}' : A\mathbf{x}' = \mathbf{y}\}$;
2. \mathbf{x} lies in a known set K .

Therefore, a good estimator of \mathbf{x} can be obtained by picking any vector $\hat{\mathbf{x}}$ from the *intersection of these two sets*; see Figure 5. Moreover, since just these two pieces of information about \mathbf{x} are available, such estimator is best possible in some sense.

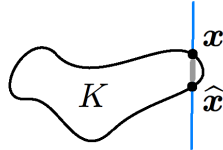


Figure 5. Estimating \mathbf{x} by any vector $\hat{\mathbf{x}}$ in the intersection of K with the affine subspace $\{\mathbf{x}' : A\mathbf{x}' = \mathbf{y}\}$

How good is such estimate? The maximal error is, of course, the distance between two farthest points in the intersection of K with the affine subspace $\{\mathbf{x}' : A\mathbf{x}' = \mathbf{y}\}$. This distance in turn equals the diameter of the section of K by this random subspace. But this diameter is controlled by M^* bound, Theorem 3.12. Let us put together this argument more rigorously.

In the following theorem, the setting is the same as above: $K \subset \mathbb{R}^n$ is a bounded subset, $\mathbf{x} \in K$ is an unknown vector and $\mathbf{y} = A\mathbf{x}$ is the observation vector, where A is an $m \times n$ Gaussian matrix.

Theorem 4.1 (Estimation from linear observations: feasibility program). *Choose $\hat{\mathbf{x}}$ to be any vector satisfying*

$$\hat{\mathbf{x}} \in K \quad \text{and} \quad A\hat{\mathbf{x}} = \mathbf{y}. \quad (4.1)$$

Then

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{Cw(K)}{\sqrt{m}}.$$

Proof. We apply the M^* bound, Theorem 3.12, for the set $K - K$ and the subspace $E = \ker(A)$. Rotation invariance of Gaussian distribution implies that E is uniformly distributed in the Grassmanian $G_{n,n-m}$, as required by the M^* bound. Moreover, it is straightforward to check that $w(K - K) \leq 2w(K)$. It follows that

$$\mathbb{E} \text{diam}((K - K) \cap E) \leq \frac{Cw(K)}{\sqrt{m}}.$$

It remains to note that since $\hat{\mathbf{x}}, \mathbf{x} \in K$ and $A\hat{\mathbf{x}} = A\mathbf{x} = \mathbf{y}$, we have $\hat{\mathbf{x}} - \mathbf{x} \in (K - K) \cap E$. \square

The argument we just described was first suggested by S. Mendelson, A. Pajor and N. Tomczak-Jaegermann [49].

4.2. Estimation as an optimization problem. Let us make one step forward and replace the feasibility program (4.1) by a more flexible *optimization* program.

For this, let us make an additional (but quite mild) assumption that K has non-empty interior and is *star-shaped*. Being star-shaped means that together with each point, the set K contains the segment joining that point to the origin; in other words,

$$tK \subseteq K \quad \text{for all } t \in [0, 1].$$

For such set K , let us revise the feasibility program (4.1). Instead of intersecting a fixed set K with the affine subspace $\{\mathbf{x}' : A\mathbf{x}' = \mathbf{y}\}$, we may *blow up* K (i.e. consider a dilate tK with increasing $t \geq 0$) until it touches that subspace. Choose $\hat{\mathbf{x}}$ to be the touching point, see Figure 6. The fact that K is star-shaped implies that $\hat{\mathbf{x}}$ still belongs to K and (obviously) the affine subspace; thus $\hat{\mathbf{x}}$ satisfies the same error bound as in Theorem 4.1.

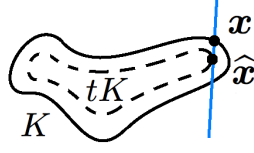


Figure 6. Estimating \mathbf{x} by blowing up K until it touches the affine subspace $\{\mathbf{x}' : A\mathbf{x}' = \mathbf{y}\}$

To express this estimator analytically, it is convenient to use the notion of *Minkowski functional* of K , which associates to each point $\mathbf{x} \in \mathbb{R}^n$ a non-negative number $\|\mathbf{x}\|_K$ defined by the rule

$$\|\mathbf{x}\|_K = \inf \{ \lambda > 0 : \lambda^{-1} \mathbf{x} \in K \}.$$

A simple situation to think of is when K is an compact and origin-symmetric convex set with non-empty interior; then $\|\mathbf{x}\|_K$ is a *norm* on \mathbb{R}^n . The closed unit ball corresponding to this norm is K .

Let us now accurately state an optimization version of Theorem 4.1. It is valid for an arbitrary bounded star-shaped set K with non-empty interior.

Theorem 4.2 (Estimation from linear observations: optimization program). *Choose $\hat{\mathbf{x}}$ to be a solution of the program*

$$\text{minimize } \|\mathbf{x}'\|_K \quad \text{subject to} \quad A\mathbf{x}' = \mathbf{y}. \quad (4.2)$$

Then

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{Cw(K)}{\sqrt{m}}.$$

Proof. It suffices to check that $\hat{\mathbf{x}} \in K$; the conclusion would then follow from Theorem 4.1. Both $\hat{\mathbf{x}}$ and \mathbf{x} satisfy the linear constraint $A\mathbf{x}' = \mathbf{y}$. Therefore, by choice of $\hat{\mathbf{x}}$, we have

$$\|\hat{\mathbf{x}}\|_K \leq \|\mathbf{x}\|_K \leq 1;$$

the last inequality is nothing else than our assumption that $\mathbf{x} \in K$. Thus $\hat{\mathbf{x}} \in K$ as claimed. \square

4.3. Algorithmic aspects: convex programming. What does it take to solve the optimization problem (4.2) algorithmically? If the feasible set K is convex, then (4.2) is a *convex program*. In this case, to solve this problem numerically, we may use one of the array of convex solvers. Further, if K is a polytope, then (4.2) can be cast as a *linear program*, which widens an array of algorithmic possibilities even further. For a quick preview, let us mention that examples of the latter kind will be discussed in detail in Section 7, where we will use K to enforce *sparsity*. We will thus choose K to be a ball of ℓ_1 norm in \mathbb{R}^n , so the program (4.2) will minimize $\|\mathbf{x}'\|_1$ subject to $A\mathbf{x}' = \mathbf{y}$. This is a typical linear program in the area of compressed sensing.

If K is *not* convex, then we can convexify it, thereby replacing K with its convex hull $\text{conv}(K)$. Convexification does not change the mean width,

according to the remarkable property (3.4). Therefore, the generally non-convex problem (4.2) can be relaxed to the convex program

$$\text{minimize } \|\mathbf{x}'\|_{\text{conv}(K)} \quad \text{subject to } A\mathbf{x}' = \mathbf{y}, \quad (4.3)$$

without compromising the guarantee of estimation stated in Theorem 4.2. The solution $\hat{\mathbf{x}}$ of the convex program (4.3) still satisfies

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{Cw(K)}{\sqrt{m}}.$$

Summarizing, we see that in any case, whether K is convex or not, the estimation problem reduces to solving an algorithmically tractable convex program. Of course, for this one needs to be able to compute $\|\mathbf{z}\|_{\text{conv}(K)}$ algorithmically for a given vector $\mathbf{z} \in \mathbb{R}^n$. This is possible for many (but not all) feasible sets K .

4.4. Information-theoretic aspects: effective dimension. If we fix a desired error level, for example if we aim for

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq 0.01,$$

then

$$m \sim w(K)^2$$

observations will suffice. The implicit constant factor here is determined by the desired error level.

Notice that this result is *uniform*, in the sense that with high probability in A (which determines the observation model) the estimation is accurate simultaneously for all vectors $\mathbf{x} \in K$.

The square of the mean width, $w(K)^2$, can be thought of an *effective dimension* of the feasible set K , as we pointed out in Section 3.5.6.

We can summarize our findings as follows.

Using convex programming, one can estimate a vector \mathbf{x} in a general feasible set K from m random linear observations. A sufficient number of observations m is the same as the effective dimension of K (the mean width squared), up to a constant factor.

5. HIGH DIMENSIONAL SECTIONS: PROOF OF A GENERAL M^* BOUND

Let us give a quick proof of the M^* bound, Theorem 3.12. In fact, without much extra work we will be able to derive a more general result from [55]. First, it would allow us to treat noisy observations of the form $\mathbf{y} = A\mathbf{x} + \boldsymbol{\nu}$. Second, it will be generalizable for non-gaussian observations.

Theorem 5.1 (General M^* bound). *Let T be a bounded subset of \mathbb{R}^n . Let A be an $m \times n$ Gaussian random matrix (with i.i.d. $N(0, 1)$ entries). Fix $\varepsilon \geq 0$ and consider the set*

$$T_\varepsilon := \left\{ \mathbf{u} \in T : \frac{1}{m} \|A\mathbf{u}\|_1 \leq \varepsilon \right\}. \quad (5.1)$$

Then

$$\mathbb{E} \sup_{\mathbf{u} \in T_\varepsilon} \|\mathbf{u}\|_2 \leq \sqrt{\frac{2\pi}{m}} \mathbb{E} \sup_{\mathbf{u} \in T} |\langle \mathbf{g}, \mathbf{u} \rangle| + \sqrt{\frac{\pi}{2}} \varepsilon, \quad (5.2)$$

where $\mathbf{g} \sim N(0, I_n)$ is a standard Gaussian random vector in \mathbb{R}^n .

To see that this result contains the classical M^* bound, Theorem 3.12, we can apply it for $T = K - K$, $\varepsilon = 0$, and identify $\ker(A)$ with E . In this case,

$$T_\varepsilon = (K - K) \cap E.$$

It follows that $T_\varepsilon \supseteq (K \cap E) - (K \cap E)$, so the left hand side in (5.2) is bounded below by $\text{diam}(K \cap E)$. The the right hand side in (5.2) by symmetry equals $\sqrt{2\pi/m} w(K)$. Thus, we recover Theorem 3.12 with $C = \sqrt{2\pi}$.

Our proof of Theorem 5.1 will be based on two basic tools in the theory of *stochastic processes* – symmetrization and contraction.

A stochastic process is simply a collection of random variables $(Z(t))_{t \in T}$ on the same probability space. The index space T can be arbitrary; it may be a time interval (such as in Brownian motion) or a subset of \mathbb{R}^n (as will be our case). To avoid measurability issues, we can assume that T is finite by discretizing it if necessary.

Proposition 5.2. *Consider a finite collection of stochastic processes $Z_1(t), \dots, Z_m(t)$ indexed by $t \in T$. Let ε_i be independent symmetric Bernoulli random variables, i.e. each ε_i independently takes values -1 and 1 with probabilities $1/2$. Then we have the following.*

(i) (Symmetrization)

$$\mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^m [Z_i(t) - \mathbb{E} Z_i(t)] \right| \leq 2 \mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^m \varepsilon_i Z_i(t) \right|.$$

(ii) (Contraction)

$$\mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^m \varepsilon_i |Z_i(t)| \right| \leq \mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^m \varepsilon_i Z_i(t) \right|.$$

Both statements are relatively easy to prove, and even in greater generality. For example, taking the absolute values of $Z_i(t)$ in the contraction principle can be replaced by applying general Lipschitz functions. Proofs of symmetrization and contraction principles can be found in [40, Lemma 6.3] and [40, Theorem 4.12], respectively.

5.1. Proof of Theorem 5.1. The desired bound (5.2) would follow from the deviation inequality

$$\mathbb{E} \sup_{\mathbf{u} \in T} \left| \frac{1}{m} \sum_{i=1}^m |\langle \mathbf{a}_i, \mathbf{u} \rangle| - \sqrt{\frac{2}{\pi}} \|\mathbf{u}\|_2 \right| \leq \frac{2}{\sqrt{m}} \mathbb{E} \sup_{\mathbf{u} \in T} |\langle \mathbf{g}, \mathbf{u} \rangle|. \quad (5.3)$$

Indeed, if this inequality holds, then same is true if we replace T by the smaller set T_ε . But for $\mathbf{u} \in T_\varepsilon$, we have $\frac{1}{m} \sum_{i=1}^m |\langle \mathbf{a}_i, \mathbf{u} \rangle| = \frac{1}{m} \|A\mathbf{u}\|_1 \leq \varepsilon$, and the bound (5.2) follows by triangle inequality.

The rotation invariance of Gaussian distribution implies that

$$\mathbb{E} |\langle \mathbf{a}_i, \mathbf{u} \rangle| = \sqrt{\frac{2}{\pi}} \|\mathbf{u}\|_2. \quad (5.4)$$

Thus, using symmetrization and then contraction inequalities from Proposition 5.2, we can bound the left side in (5.3) by

$$2 \mathbb{E} \sup_{\mathbf{u} \in T} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle \mathbf{a}_i, \mathbf{u} \rangle \right| = 2 \mathbb{E} \sup_{\mathbf{u} \in T} \left| \left\langle \frac{1}{m} \sum_{i=1}^m \varepsilon_i \mathbf{a}_i, \mathbf{u} \right\rangle \right|. \quad (5.5)$$

Here ε_i are independent symmetric Bernoulli random variables.

Conditioning on ε_i and using rotation invariance, we see that the random vector

$$\mathbf{g} := \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{a}_i$$

has distribution $N(0, I_n)$. Thus (5.5) can be written as

$$\frac{2}{\sqrt{m}} \mathbb{E} \sup_{\mathbf{u} \in T} |\langle \mathbf{g}, \mathbf{u} \rangle|.$$

This proves (5.3) and completes the proof of Theorem 5.1. \square

6. CONSEQUENCES: ESTIMATION FROM NOISY LINEAR OBSERVATIONS

Let us apply the general M^* bound, Theorem 5.1, to estimation problems. This will be even more straightforward than our application of the standard M^* bound in Section 4. Moreover, we will now be able to treat *noisy* observations.

Like before, our goal is to estimate an unknown vector \mathbf{x} that lies in a known feasible set $K \subset \mathbb{R}^n$, from a random observation vector $\mathbf{y} \in \mathbb{R}^m$. This time we assume that, for some known level of noise $\varepsilon \geq 0$, we have

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\nu}, \quad \frac{1}{m} \|\boldsymbol{\nu}\|_1 = \frac{1}{m} \sum_{i=1}^m |\nu_i| \leq \varepsilon. \quad (6.1)$$

Here A is an $m \times n$ Gaussian matrix as before. The noise vector $\boldsymbol{\nu}$ may be unknown and have arbitrary structure. In particular $\boldsymbol{\nu}$ may depend on A , so even adversarial errors are allowed.

The following result is a generalization of Theorem 4.1 for noisy observations (6.1).

Theorem 6.1 (Estimation from noisy linear observations: feasibility program). *Choose $\hat{\mathbf{x}}$ to be any vector satisfying*

$$\hat{\mathbf{x}} \in K \quad \text{and} \quad \frac{1}{m} \|A\hat{\mathbf{x}} - \mathbf{y}\|_1 \leq \varepsilon. \quad (6.2)$$

Then

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \sqrt{2\pi} \left(\frac{w(K)}{\sqrt{m}} + \varepsilon \right).$$

Proof. We apply the general M^* bound, Theorem 5.1, for the set $T = K - K$, and with 2ε instead of ε . It follows that

$$\mathbb{E} \sup_{\mathbf{u} \in T_{2\varepsilon}} \|\mathbf{u}\|_2 \leq \sqrt{\frac{2\pi}{m}} \mathbb{E} \sup_{\mathbf{u} \in T_{2\varepsilon}} |\langle \mathbf{g}, \mathbf{u} \rangle| + \sqrt{2\pi} \varepsilon \leq \sqrt{2\pi} \left(\frac{w(K)}{\sqrt{m}} + \varepsilon \right).$$

The last inequality should be clear once we replace $T_{2\varepsilon}$ by the larger set $T = K - K$ and use the symmetry of T .

To finish the proof, it remains to check that

$$\hat{\mathbf{x}} - \mathbf{x} \in T_{2\varepsilon}. \quad (6.3)$$

To prove this, first note that $\hat{\mathbf{x}}, \mathbf{x} \in K$, so $\hat{\mathbf{x}} - \mathbf{x} \in K - K = T$. Next, by triangle inequality, we have

$$\frac{1}{m} \|A(\hat{\mathbf{x}} - \mathbf{x})\|_1 = \frac{1}{m} \|A\hat{\mathbf{x}} - \mathbf{y} + \boldsymbol{\nu}\|_1 \leq \frac{1}{m} \|A\hat{\mathbf{x}} - \mathbf{y}\|_1 + \frac{1}{m} \|\boldsymbol{\nu}\|_1 \leq 2\varepsilon.$$

The last inequality follows from (6.1) and (6.2). We showed that the vector $\mathbf{u} = \hat{\mathbf{x}} - \mathbf{x}$ satisfies both constraints that define $T_{2\varepsilon}$ in (5.1). Hence (6.3) holds, and the proof of the theorem is complete. \square

And similarly to Theorem 4.2, we can cast estimation as an optimization (rather than feasibility) program.

Theorem 6.2 (Estimation from noisy linear observations: optimization program). *Choose $\hat{\mathbf{x}}$ to be a solution to the program*

$$\text{minimize } \|\mathbf{x}'\|_K \quad \text{subject to} \quad \frac{1}{m} \|A\mathbf{x}' - \mathbf{y}\|_1 \leq \varepsilon. \quad (6.4)$$

Then

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \sqrt{2\pi} \left(\frac{w(K)}{\sqrt{m}} + \varepsilon \right).$$

Proof. It suffices to check that $\hat{\mathbf{x}} \in K$; the conclusion would then follow from Theorem 6.1. Note first that by choice of $\hat{\mathbf{x}}$ we have $\frac{1}{m} \|A\hat{\mathbf{x}} - \mathbf{y}\|_1 \leq \varepsilon$, and by assumption (6.1) we have $\frac{1}{m} \|A\mathbf{x} - \mathbf{y}\|_1 = \frac{1}{m} \|\boldsymbol{\nu}\|_1 \leq \varepsilon$. Thus both $\hat{\mathbf{x}}$ and \mathbf{x} satisfy the constraint in (6.4). Therefore, by choice of $\hat{\mathbf{x}}$, we have

$$\|\hat{\mathbf{x}}\|_K \leq \|\mathbf{x}\|_K \leq 1;$$

the last inequality is nothing else than our assumption that $\mathbf{x} \in K$. It follows $\hat{\mathbf{x}} \in K$ as claimed. \square

The remarks about algorithmic aspects of estimation made in Sections 4.3 and 4.4 apply also to the results of this section. In particular, the estimation from noisy linear observations (6.1) can be formulated as a *convex program*.

7. APPLICATIONS TO SPARSE RECOVERY AND REGRESSION

Remarkable examples of feasible sets K with low complexity come from the notion of *sparsity*. Consider the set K of all unit s -sparse vectors in \mathbb{R}^n . As we mentioned in Example 3.9, the mean width of K is

$$w(K) \sim s \log(n/s).$$

According to the interpretation we discussed in Section 4.4, this means that the *effective dimension* of K is of order $s \log(n/s)$. Therefore,

$$m \sim s \log(n/s)$$

observations should suffice to estimate any s -sparse vector in \mathbb{R}^n . Results of this type form the core of *compressed sensing*, a young area of signal processing, see [16, 36, 12, 32].

In this section we consider a more general model, where an unknown vector \mathbf{x} has a sparse representation *in some dictionary*.

We will specialize Theorem 6.2 to the sparse recovery problem. The convex program will in this case amount to minimizing the ℓ_1 norm of the coefficients. We will note that the notion of sparsity can be relaxed to accommodate approximate, or “effective”, sparsity. Finally, we will observe that the estimate $\hat{\mathbf{x}}$ is most often unique and m -sparse.

7.1. Sparse recovery for general dictionaries. Let us fix a *dictionary* of vectors $\mathbf{d}_1, \dots, \mathbf{d}_N \in \mathbb{R}^n$. The dictionary may be arbitrary and even redundant, i.e. not linearly independent. The choice of a dictionary depends on the application; common examples include unions of orthogonal bases and more generally tight frames (in particular, Gabor frames). See [15, 18, 11, 17] for an introduction to sparse recovery problems with general dictionaries.

Suppose an unknown vector $\mathbf{x} \in \mathbb{R}^n$ is s -sparse *in the dictionary* $\{\mathbf{d}_i\}$. This means that \mathbf{x} can be represented as a linear combination of at most s dictionary elements, i. e.

$$\mathbf{x} = \sum_{i=1}^N \alpha_i \mathbf{d}_i \quad \text{with at most } s \text{ non-zero coefficients } \alpha_i \in \mathbb{R}. \quad (7.1)$$

As in Section 6, our goal is to recover \mathbf{x} from a noisy observation vector $\mathbf{y} \in \mathbb{R}^m$ of the form

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\nu}, \quad \frac{1}{m} \|\boldsymbol{\nu}\|_1 = \frac{1}{m} \sum_{i=1}^m |\nu_i| \leq \varepsilon.$$

Recall that A is a known $m \times n$ Gaussian matrix, and $\boldsymbol{\nu}$ is an unknown noise vector, which can have arbitrary structure (in particular, correlated with A).

Theorem 6.2 will quickly imply the following recovery result.

Theorem 7.1 (Sparse recovery: general dictionaries). *Assume for normalization that all dictionary vectors satisfy $\|\mathbf{d}_i\|_2 \leq 1$. Choose $\hat{\mathbf{x}}$ to be a solution to the convex program*

$$\text{minimize } \|\boldsymbol{\alpha}'\|_1 \text{ such that } \mathbf{x}' = \sum_{i=1}^N \alpha'_i \mathbf{d}_i \text{ satisfies } \frac{1}{m} \|A\mathbf{x}' - \mathbf{y}\|_1 \leq \varepsilon. \quad (7.2)$$

Then

$$\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C \sqrt{\frac{s \log N}{m}} \cdot \|\boldsymbol{\alpha}\|_2 + \sqrt{2\pi} \varepsilon.$$

Proof. Consider the sets

$$\bar{K} := \text{conv}\{\pm \mathbf{d}_i\}_{i=1}^N, \quad K := \|\boldsymbol{\alpha}\|_1 \cdot \bar{K}.$$

Representation (7.1) implies that $\mathbf{x} \in K$, so it makes sense to apply Theorem 6.2 for K .

Let us first argue that the optimization program in Theorem 6.2 can be written in the form (7.2). Observe that we can replace $\|\mathbf{x}'\|_K$ by $\|\mathbf{x}'\|_{\bar{K}}$ in the optimization problem (6.4) without changing its solution. (This is because $\|\mathbf{x}'\|_{\bar{K}} = \|\boldsymbol{\alpha}\|_1 \cdot \|\mathbf{x}'\|_K$ and $\|\boldsymbol{\alpha}\|_1$ is a constant value.) Now, by definition of \bar{K} , we have

$$\|\mathbf{x}'\|_{\bar{K}} = \min \left\{ \|\boldsymbol{\alpha}'\|_1 : \mathbf{x}' = \sum_{i=1}^N \alpha'_i \mathbf{d}_i \right\}.$$

Therefore, the optimization programs (6.4) and (7.2) are indeed equivalent.

Next, to evaluate the error bound in Theorem 6.2, we need to bound the mean width of K . The convexification property (3.4) and Example 3.8 yield

$$w(K) = \|\boldsymbol{\alpha}\|_1 \cdot w(\bar{K}) \leq C \|\boldsymbol{\alpha}\|_1 \cdot \sqrt{\log N}.$$

Putting this into the conclusion of Theorem 6.2, we obtain the error bound

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \sqrt{2\pi} C \sqrt{\frac{\log N}{m}} \cdot \|\boldsymbol{\alpha}\|_1 + \sqrt{2\pi} \varepsilon.$$

To complete the proof, it remains to note that

$$\|\boldsymbol{\alpha}\|_1 \leq \sqrt{s} \cdot \|\boldsymbol{\alpha}\|_2, \quad (7.3)$$

since $\boldsymbol{\alpha}$ is s -sparse, i.e. it has only s non-zero coordinates. \square

7.2. Remarkable properties of sparse recovery. Let us pause to look more closely at the statement of Theorem 7.1.

7.2.1. General dictionaries. Theorem 7.1 is very flexible with respect to the choice of a dictionary $\{\mathbf{d}_i\}$. Note that there are essentially *no restrictions on the dictionary*. (The normalization assumption $\|\mathbf{d}_i\|_2 \leq 1$ can be dispensed of at the cost of increasing the error bound by the factor of $\max_i \|\mathbf{d}_i\|_2$.) In particular, the dictionary may be *linearly dependent*.

7.2.2. Effective sparsity. The reader may have noticed that the proof of Theorem 7.1 used sparsity in a quite mild way, only through inequality (7.3). So the result is still true for vectors \mathbf{x} that are *approximately sparse* in the dictionary. Namely, the Theorem 7.1 will hold if we replace the exact notion of sparsity (the number of nonzero coefficients) by the more flexible notion of *effective sparsity*, defined as

$$\text{effective sparsity}(\boldsymbol{\alpha}) := (\|\boldsymbol{\alpha}\|_1 / \|\boldsymbol{\alpha}\|_2)^2.$$

It is now clear how to extend sparsity in a dictionary (7.1) to approximate sparsity. We can say that a vector \mathbf{x} is *effectively s -sparse in a dictionary* $\{\mathbf{d}_i\}$ if it can be represented as $\mathbf{x} = \sum_{i=1}^N \alpha_i \mathbf{d}_i$ where the coefficient vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$ is effectively s -sparse.

The effective sparsity is clearly bounded by the exact sparsity, and it is robust with respect to small perturbations.

7.2.3. Linear programming. The convex programs (7.2) and (7.5) can be reformulated. This can be done by introducing new variables u_1, \dots, u_N ; instead of minimizing $\|\boldsymbol{\alpha}'\|_1$ in (7.2), we can equivalently minimize the linear function $\sum_{i=1}^N u_i$ subject to the additional linear constraints $-u_i \leq \alpha'_i \leq u_i$, $i = 1, \dots, N$. In a similar fashion, one can replace the convex constraint $\frac{1}{m} \|A\mathbf{x}' - \mathbf{y}\|_1 \leq \varepsilon$ in (7.2) by n linear constraints.

7.2.4. Estimating the coefficients of sparse representation. It is worthwhile to notice that as a result of solving the convex recovery program (7.2), we obtain not only an estimate $\hat{\mathbf{x}}$ of the vector \mathbf{x} , but also an estimate $\hat{\boldsymbol{\alpha}}$ of the *coefficient vector* in the representation $\mathbf{x} = \sum \alpha_i \mathbf{d}_i$. However, only \mathbf{x} can be estimated accurately; it should be clear that $\boldsymbol{\alpha}$ can not be estimated by any method if the dictionary $\{\mathbf{d}_i\}$ is redundant (i.e. linearly dependent).

7.2.5. Sparsity of solution. The solution of the sparse recovery problem (7.2) may not be exact in general, i.e. when $\hat{\mathbf{x}} \neq \mathbf{x}$. This can be due to several factors – the generality of the dictionary, approximate (rather than exact) sparsity of \mathbf{x} in the dictionary, or the noise ν in the observations. But even in this general situation, the solution \mathbf{x} is still m -sparse, in all but degenerate cases.

Proposition 7.2 (Sparsity of solution). *Assume that a given convex recovery program (7.2) has a unique solution $\hat{\boldsymbol{\alpha}}$ for the coefficient vector. Then $\hat{\boldsymbol{\alpha}}$ is m -sparse, and consequently $\hat{\mathbf{x}}$ is m -sparse in the dictionary $\{\mathbf{d}_i\}$. This is true even in presence of noise in observations, and even when no sparsity assumptions on \mathbf{x} are in place.*

Proof. The result follows by simple dimension considerations. First note that the constraint on $\boldsymbol{\alpha}'$ in the optimization problem (7.2) can be written in the form

$$\frac{1}{m} \|AD\boldsymbol{\alpha}' - \mathbf{y}\|_1 \leq \varepsilon, \quad (7.4)$$

where D is the $n \times N$ matrix whose columns are the dictionary vectors \mathbf{d}_i . Since matrix AD has dimensions $m \times N$, the constraint defines a cylinder in \mathbb{R}^N whose infinite directions are formed by the kernel of AD , which has dimension at least $N - m$. Moreover, this cylinder is a polytope (due to the ℓ_1 norm defining it), so it has no faces of dimension smaller than $N - m$.

On the other hand, the level sets of the objective function $\|\boldsymbol{\alpha}'\|_1$ are also polytopes; they are dilates of the unit ℓ_1 ball. The solution $\hat{\boldsymbol{\alpha}}$ of the optimization problem (7.2) is thus a point in \mathbb{R}^N where the smallest dilate of the ℓ_1 ball touches the cylinder. The uniqueness of solution means that a touching point is unique. This is illustrated in Figure 7.

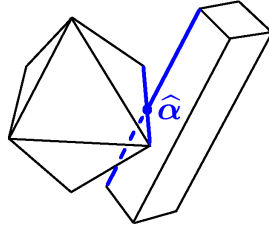


Figure 7. Illustration for the proof of Proposition 7.2. The polytope on the left represents a level set of the ℓ_1 ball. The cylinder on the right represents the vectors $\boldsymbol{\alpha}'$ satisfying the constraint (7.4). The two polytopes touch at point $\hat{\boldsymbol{\alpha}}$.

Consider the faces of these two polytopes of smallest dimensions that contain the touching point; we may call these the touching faces. The touching face of the cylinder has dimension at least $N - m$, as all of its faces do. Then the touching face of the ℓ_1 ball must have dimension at most m , otherwise the two touching faces would intersect by more than one point. This translates into the m -sparsity of the solution $\hat{\boldsymbol{\alpha}}$, as claimed. \square

7.2.6. Uniqueness of solution. In view of Proposition 7.2, we can ask when the solution $\hat{\boldsymbol{\alpha}}$ of the convex program (7.2) is unique. This does not always happen; for example this fails if $\mathbf{d}_1 = \mathbf{d}_2$.

We can get around this problem by making an arbitrarily small generic perturbation of the dictionary elements, such as adding a small independent Gaussian vector to each \mathbf{d}_i . Then one can see that the solution $\hat{\boldsymbol{\alpha}}$ (and therefore $\hat{\mathbf{x}}$ as well) are unique almost surely. Invoking Proposition 7.2 we see that $\hat{\mathbf{x}}$ is m -sparse in the perturbed dictionary.

Note that we do not need the dictionary \mathbf{d}_i to be linearly independent in order for this to happen; the dictionary will always be dependent if $N > n$.

7.3. Sparse recovery for the canonical dictionary. Let us illustrate Theorem 7.1 for the simplest example of a dictionary – the canonical basis of \mathbb{R}^n :

$$\{\mathbf{d}_i\}_{i=1}^n = \{\mathbf{e}_i\}_{i=1}^n.$$

In this case, our assumption is that an unknown vector $\mathbf{x} \in \mathbb{R}^n$ is *s-sparse* in the usual sense, meaning that \mathbf{x} has at most s non-zero coordinates, or *effectively s-sparse* as in Section 7.2.2. Theorem 7.1 then reads as follows.

Corollary 7.3 (Sparse recovery). *Choose $\hat{\mathbf{x}}$ to be a solution to the convex program*

$$\text{minimize } \|\mathbf{x}'\|_1 \text{ subject to } \frac{1}{m} \|A\mathbf{x}' - \mathbf{y}\|_1 \leq \varepsilon. \quad (7.5)$$

Then

$$\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C \sqrt{\frac{s \log n}{m}} \cdot \|\mathbf{x}\|_2 + \sqrt{2\pi} \varepsilon. \quad \square$$

Sparse recovery results like Corollary 7.3 form the core of the area of *compressed sensing*, see [16, 36, 12, 32].

In the noiseless case ($\varepsilon = 0$) and for sparse (rather than effectively sparse) vectors, one may even hope to recover \mathbf{x} *exactly*, meaning that $\hat{\mathbf{x}} = \mathbf{x}$ with high probability. Conditions for exact recovery are now well understood in compressed sensing. We will discuss some exact recovery problems in Section 9.

We can summarize Theorem 7.1 and the discussion around it as follows.

Using linear programming, one can approximately recover a vector \mathbf{x} that is s-sparse (or effectively s-sparse) in a general dictionary of size N , from $m \sim s \log N$ random linear observations.

7.4. Application: linear regression with constraints. The noisy estimation problem (6.1) is equivalent to *linear regression* with constraints. So in this section we will translate the story into the statistical language. We present here just one class of examples out of a wide array of statistical problems; we refer the reader to [66] for a recent review of high dimensional estimation problems from a statistical viewpoint.

Linear regression is a linear model of relationship between one dependent variable and n explanatory variables. The problem is to find the best linear relationship from a sample of p observations of dependent and explanatory variables. Linear regression is usually written as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\nu}.$$

Here X is an $n \times p$ matrix which contains a sample of n observations of p explanatory variables; $\mathbf{y} \in \mathbb{R}^n$ represents a sample of n observations of the dependent variable; $\boldsymbol{\beta} \in \mathbb{R}^p$ is a coefficient vector; $\boldsymbol{\nu} \in \mathbb{R}^n$ is a noise vector. We assume that X and \mathbf{y} are known, while $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are unknown. Our goal is to estimate $\boldsymbol{\beta}$.

We discussed a classical formulation of linear regression. In addition, we often know, believe, or want to enforce some properties about the coefficient vector $\boldsymbol{\beta}$, (for example, sparsity). We can express such extra information as the assumption that

$$\boldsymbol{\beta} \in K$$

where $K \subset \mathbb{R}^p$ is a known feasible set. Such problem may be called a *linear regression with constraints*.

The high dimensional estimation results we have seen so far can be translated into the language of regression in a straightforward way. Let us do this for Theorem 6.2; the interested reader can make a similar translation or other results.

We assume that the explanatory variables are independent $N(0, 1)$, so the matrix X has all i.i.d. $N(0, 1)$ entries. This requirement may be too strong in practice; however see Section 8 on relaxing this assumption. The noise vector $\boldsymbol{\nu}$ is allowed have arbitrary structure (in particular, it can be correlated with X). We assume that its magnitude is controlled:

$$\frac{1}{n} \|\boldsymbol{\nu}\|_1 = \frac{1}{n} \sum_{i=1}^n |\nu_i| \leq \varepsilon$$

for some known noise level ε .

Theorem 7.4 (Linear regression with constraints). *Choose $\hat{\boldsymbol{\beta}}$ to be a solution to the program*

$$\text{minimize } \|\boldsymbol{\beta}'\|_K \quad \text{subject to } \frac{1}{n} \|X\boldsymbol{\beta}' - \mathbf{y}\|_1 \leq \varepsilon.$$

Then

$$\mathbb{E} \sup_{\boldsymbol{\beta} \in K} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \sqrt{2\pi} \left(\frac{w(K)}{\sqrt{n}} + \varepsilon \right). \quad \square$$

8. EXTENSIONS FROM GAUSSIAN TO SUB-GAUSSIAN DISTRIBUTIONS

So far, all our results were stated for Gaussian distributions. Let us show how to relax this assumption. In this section, we will modify the proof of the M^* bound, Theorem 5.1 for general *sub-gaussian* distributions, and indicate the consequences for the estimation problem. A result of this type was proved in [49] with a much more complex argument.

8.1. Sub-gaussian random variables and random vectors. A systematic introduction into sub-gaussian distributions can be found in Sections 5.2.3 and 5.2.5 of [65]; here we briefly mention the basic definitions. According to one of the several equivalent definitions, a random variable X is *sub-gaussian* if

$$\mathbb{E} \exp(X^2/\psi^2) \leq e.$$

for some $\psi > 0$. The smallest ψ is called the *sub-gaussian norm* and is denoted $\|X\|_{\psi_2}$.

The notion of sub-gaussian distribution transfers to higher dimensions as follows. A random vector $\mathbf{X} \in \mathbb{R}^n$ is called sub-gaussian if all one-dimensional marginals $\langle \mathbf{X}, \mathbf{u} \rangle$, $\mathbf{u} \in \mathbb{R}^n$, are sub-gaussian random variables. The sub-gaussian norm of \mathbf{X} is defined as

$$\|\mathbf{X}\|_{\psi_2} := \sup_{\mathbf{u} \in S^{n-1}} \|\langle \mathbf{X}, \mathbf{u} \rangle\|_{\psi_2} \quad (8.1)$$

where, as before, S^{n-1} denotes the Euclidean sphere in \mathbb{R}^n . Recall also that the random vector \mathbf{X} is called *isotropic* if

$$\mathbb{E} \mathbf{X} \mathbf{X}^\top = I_n.$$

Isotropy is a scaling condition; any distribution in \mathbb{R}^n which is not supported in a low-dimensional subspace can be made isotropic by an appropriate linear transformation.

8.2. M^* bound for sub-gaussian distributions. Now we state and prove a version of M^* bound, Theorem 5.1, for general *sub-gaussian* distributions. It is a variant of a result from [49].

Theorem 8.1 (General M^* bound for sub-gaussian distributions). *Let T be a bounded subset of \mathbb{R}^n . Let A be an $m \times n$ matrix whose rows \mathbf{a}_i are i.i.d., mean zero, isotropic and sub-gaussian random vectors in \mathbb{R}^n . Choose $\psi \geq 1$ so that*

$$\|\mathbf{a}_i\|_{\psi_2} \leq \psi, \quad i = 1, \dots, m. \quad (8.2)$$

Fix $\varepsilon \geq 0$ and consider the set

$$T_\varepsilon := \left\{ \mathbf{u} \in T : \frac{1}{m} \|A\mathbf{u}\|_1 \leq \varepsilon \right\}.$$

Then

$$\mathbb{E} \sup_{\mathbf{u} \in T_\varepsilon} \|\mathbf{u}\|_2 \leq C\psi^4 \left(\frac{1}{\sqrt{m}} \mathbb{E} \sup_{\mathbf{u} \in T_\varepsilon} |\langle \mathbf{g}, \mathbf{u} \rangle| + \varepsilon \right),$$

where $\mathbf{g} \sim N(0, I_n)$ is a standard Gaussian random vector in \mathbb{R}^n .

A proof of this result is an extension of the proof of the Gaussian M^* bound, Theorem 5.1. Most of that argument generalizes to sub-gaussian distributions in a standard way. The only non-trivial new step will be based on the deep *comparison theorem for sub-gaussian processes* due to X. Fernique and M. Talagrand, see [64, Section 2.1]. Informally, the result states that any sub-gaussian process is dominated by a Gaussian process with the same (or larger) increments.

Theorem 8.2 (Fernique-Talagrand's comparison theorem). *Let T be an arbitrary set.² Consider a Gaussian random process $(G(\mathbf{t}))_{\mathbf{t} \in T}$ and a sub-gaussian random process $(H(\mathbf{t}))_{\mathbf{t} \in T}$. Assume that $\mathbb{E} G(\mathbf{t}) = \mathbb{E} H(\mathbf{t}) = 0$ for all $\mathbf{t} \in T$. Assume also that for some $M > 0$, the following increment comparison holds:³*

$$\|H(\mathbf{s}) - H(\mathbf{t})\|_{\psi_2} \leq M (\mathbb{E} \|G(\mathbf{s}) - G(\mathbf{t})\|_2^2)^{1/2} \quad \text{for all } \mathbf{s}, \mathbf{t} \in T.$$

Then

$$\mathbb{E} \sup_{\mathbf{t} \in T} H(\mathbf{t}) \leq CM \mathbb{E} \sup_{\mathbf{t} \in T} G(\mathbf{t}).$$

²We can assume T to be finite to avoid measurability complications, and then proceed by approximation; see e.g. [40].

³The increment comparison may look better if we replace the L_2 norm in the right hand side by ψ_2 norm. Indeed, it is easy to see that $\|G(\mathbf{s}) - G(\mathbf{t})\|_{\psi_2} \asymp (\mathbb{E} \|G(\mathbf{s}) - G(\mathbf{t})\|_2^2)^{1/2}$.

This theorem is a combination of a result of X. Fernique [31] that bounds $\mathbb{E} \sup_{t \in T} H(t)$ above by the so-called *majorizing measure* of T , and a result of M. Talagrand [63] that bounds $\mathbb{E} \sup_{t \in T} G(t)$ below by the same majorizing measure of T .

Proof of Theorem 8.1. Let us examine the proof of the Gaussian M^* bound, Theorem 5.1, check where we used Gaussian assumptions, and try to accommodate sub-gaussian assumptions instead.

The first such place is identity (5.4). We claim that a version of it still holds for the sub-gaussian random vector \mathbf{a} , namely

$$\|\mathbf{u}\|_2 \leq C_0 \psi^3 \mathbb{E}_{\mathbf{a}} |\langle \mathbf{a}, \mathbf{u} \rangle| \quad (8.3)$$

where C_0 is an absolute constant.⁴

To check (8.3), we can assume that $\|\mathbf{u}\|_2 = 1$ by dividing both sides by $\|\mathbf{u}\|_2$ if necessary. Then $Z := \langle \mathbf{a}, \mathbf{u} \rangle$ is sub-gaussian random variable, since according to (8.1) and (8.2), we have $\|Z\|_{\psi_2} \leq \|\mathbf{a}\|_{\psi_2} \leq \psi$. Then, since sub-gaussian distributions have moments of all orders (see [65, Lemma 5.5]), we have $(\mathbb{E} Z^3)^{1/3} \leq C_1 \|Z\|_{\psi_2} \leq C_1 \psi$, where C_1 is an absolute constant. Using this together with isotropy and Cauchy-Schwarz inequality, we obtain

$$1 = \mathbb{E} Z^2 = \mathbb{E} Z^{1/2} Z^{3/2} \leq (\mathbb{E} Z)^{1/2} (\mathbb{E} Z^3)^{1/2} \leq (\mathbb{E} Z)^{1/2} (C_1 \psi)^{3/2}.$$

Squaring both sides implies (8.3), since we assumed that $\|\mathbf{u}\|_2 = 1$.

The next steps in the proof of Theorem 5.1 – symmetrization and contraction – go through for sub-gaussian distributions without change. So (5.5) is still valid in our case.

Next, the random vector

$$\mathbf{h} := \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{a}_i$$

is no longer Gaussian as in the proof of Theorem 5.1. Still, \mathbf{h} is sub-gaussian with

$$\|\mathbf{h}\|_{\psi_2} \leq C_2 \psi \quad (8.4)$$

due to the approximate rotation invariance of sub-gaussian distributions, see [65, Lemma 5.9].

In the last step of the argument, we need to replace the sub-gaussian random vector \mathbf{h} by the Gaussian random vector $\mathbf{g} \sim N(0, I_n)$, i.e. prove an inequality of the form

$$\mathbb{E} \sup_{\mathbf{u} \in T_\varepsilon} |\langle \mathbf{h}, \mathbf{u} \rangle| \lesssim \mathbb{E} \sup_{\mathbf{u} \in T_\varepsilon} |\langle \mathbf{g}, \mathbf{u} \rangle|.$$

⁴We should mention that a reverse inequality also holds: by isotropy, one has $\mathbb{E}_{\mathbf{a}} |\langle \mathbf{a}, \mathbf{u} \rangle| \leq (\mathbb{E}_{\mathbf{a}} \langle \mathbf{a}, \mathbf{u} \rangle^2)^{1/2} = \|\mathbf{u}\|_2$. However, this inequality will not be used in the proof.

This can be done by applying the comparison inequality of Theorem 8.2 for the processes

$$H(\mathbf{u}) = \langle \mathbf{h}, \mathbf{u} \rangle \quad \text{and} \quad G(\mathbf{u}) = \langle \mathbf{g}, \mathbf{u} \rangle, \quad \mathbf{u} \in T \cup (-T).$$

To check the increment inequality, we can use (8.4), which yields

$$\|H(\mathbf{u}) - H(\mathbf{v})\|_{\psi_2} = \|\langle \mathbf{h}, \mathbf{u} - \mathbf{v} \rangle\|_{\psi_2} \leq \|\mathbf{h}\|_{\psi_2} \|\mathbf{u} - \mathbf{v}\|_2 \leq C_2 \psi \|\mathbf{u} - \mathbf{v}\|_2.$$

On the other hand,

$$(\mathbb{E} \|G(\mathbf{u}) - G(\mathbf{v})\|_2^2)^{1/2} = \|\mathbf{u} - \mathbf{v}\|_2.$$

Therefore, the increment inequality in Theorem 8.2 holds with $M = C_2 \psi$. It follows that

$$\mathbb{E} \sup_{\mathbf{u} \in T \cup (-T)} \langle \mathbf{h}, \mathbf{u} \rangle \leq C_3 \psi \mathbb{E} \sup_{\mathbf{u} \in T \cup (-T)} \langle \mathbf{g}, \mathbf{u} \rangle.$$

This means that

$$\mathbb{E} \sup_{\mathbf{u} \in T} |\langle \mathbf{h}, \mathbf{u} \rangle| \leq C_3 \psi \mathbb{E} \sup_{\mathbf{u} \in T} |\langle \mathbf{g}, \mathbf{u} \rangle|$$

as claimed.

Replacing all Gaussian inequalities by their sub-gaussian counterparts discussed above, we complete the proof just like in Theorem 5.1. \square

Remark 8.3 (Dependence on sub-gaussian norm). The dependence on ψ in Theorem 8.1 is not optimal. To make the argument transparent, we have not tried to optimize this dependence; the interested reader is encouraged to do so.

8.3. Estimation from sub-gaussian linear observations. It is now straightforward to generalize all recovery results we developed before from Gaussian to sub-gaussian observations. So our observations are now

$$y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle + \nu_i, \quad i = 1, \dots, m$$

where \mathbf{a}_i are i.i.d., mean zero, isotropic and *sub-gaussian* random vectors in \mathbb{R}^n . As in Theorem 8.1, we control the sub-gaussian norm with the parameter $\psi > 1$, choosing it so that

$$\|\mathbf{a}_i\|_{\psi_2} \leq \psi, \quad i = 1, \dots, m.$$

We can write observations in the matrix form as in (6.1), i.e.

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\nu},$$

where A is the $m \times n$ matrix with rows \mathbf{a}_i . As before, we assume some control on the error:

$$\frac{1}{m} \|\boldsymbol{\nu}\|_1 = \frac{1}{m} \sum_{i=1}^m |\nu_i| \leq \varepsilon.$$

Let us state a version of Theorem 6.1 for sub-gaussian observations. Its proof is the same, except we use the sub-gaussian M^* bound, Theorem 8.1 where previously a Gaussian M^* bound was used.

Theorem 8.4 (Estimation from sub-gaussian observations). *Choose $\hat{\mathbf{x}}$ to be any vector satisfying*

$$\hat{\mathbf{x}} \in K \quad \text{and} \quad \frac{1}{m} \|A\hat{\mathbf{x}} - \mathbf{y}\|_1 \leq \varepsilon.$$

Then

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C\psi^4 \left(\frac{w(K)}{\sqrt{m}} + \varepsilon \right). \quad \square$$

In a similar fashion, one can generalize all other estimation results established before to sub-gaussian observations. We leave this to the interested reader.

9. EXACT RECOVERY

In some situations, one can hope to estimate vector $\mathbf{x} \in K$ from \mathbf{y} *exactly*, without any error. Such results form the core of the area of *compressed sensing*. [16, 36, 32]. Here we will present an approach to exact recovery based on Y. Gordon’s “escape through a mesh” theorem [28]. This argument goes back to [60] for the set of sparse vectors; it was put in a general context in [13].

We will work here with Gaussian observations

$$\mathbf{y} = A\mathbf{x},$$

where A is an $m \times n$ Gaussian random matrix. This is the same model as we considered in Section 4.

9.1. Exact recovery condition and the descent cone. When can \mathbf{x} be inferred from \mathbf{y} exactly? Recall that we only know two things about \mathbf{x} – that it lies in the feasible set K and in the affine subspace

$$E_{\mathbf{x}} := \{\mathbf{x}' : A\mathbf{x}' = \mathbf{y}\}.$$

This two pieces of information determine \mathbf{x} uniquely if and only if these two sets intersect at the single point \mathbf{x} :

$$K \cap E_{\mathbf{x}} = \{\mathbf{x}\}. \tag{9.1}$$

Notice that this situation would go far beyond the M^* bound on the diameter of $K \cap E$ (see Theorem 3.12) – indeed, in this case the diameter would equal zero!

How can this be possible? Geometrically, the exact recovery condition (9.1) states that *the affine subspace $E_{\mathbf{x}}$ is tangent to the set K at the point \mathbf{x}* ; see Figure 8a for illustration.

This condition is local. Assuming that K is convex for better understanding, we see that the tangency condition depends on the shape of K in an infinitesimal neighborhood of \mathbf{x} , while the global geometry of K is irrelevant. So we would not lose anything if we replace K by the *descent*

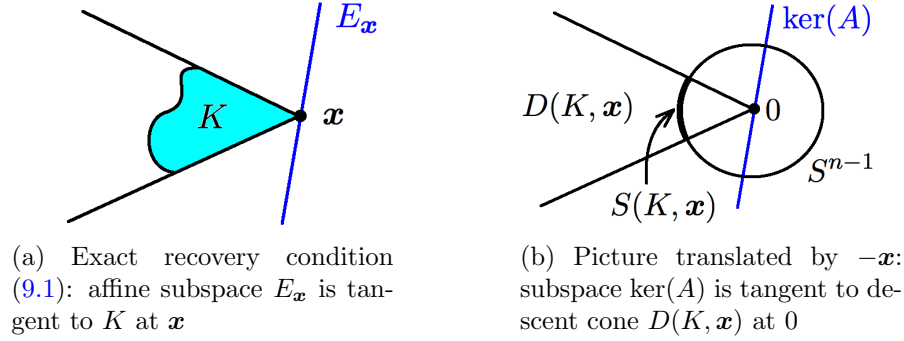


Figure 8. Illustration of the exact recovery condition (9.1)

cone at point x , see Figure 8b. This set is formed by the rays emanating from x into directions of points from K :

$$D(K, x) := \{t(z - x) : z \in K, t \geq 0\}.$$

Translating by $-x$, can we rewrite the exact recovery condition (9.1) as

$$(K - x) \cap (E_x - x) = \{0\}$$

Replacing $K - x$ by the descent cone (a bigger set) and noting that $E_x - x = \ker(A)$, we rewrite this again as

$$D(K, x) \cap \ker(A) = \{0\}.$$

The descent cone can be determined by its intersection with the unit sphere, i.e. by

$$S(K, x) := D(K, x) \cap S^{n-1} = \left\{ \frac{z - x}{\|z - x\|_2} : z \in K \right\}. \quad (9.2)$$

Thus we arrive at the following equivalent form of the exact recovery condition (9.1):

$$S(K, x) \cap \ker(A) = \emptyset;$$

see Figure 8b for an illustration.

9.2. Escape through a mesh, and implications for exact recovery.

It remains to understand under what conditions the random subspace $\ker A$ misses a given subset $S = S(K, x)$ of the unit sphere. There is a remarkably sharp result in asymptotic convex geometry that answers this question for general subsets S . This is the theorem on *escape through a mesh*, which is due to Y. Gordon [28]. Similarly to the other results we saw before, this theorem depends on the *mean width* of S , defined as⁵

$$\bar{w}(S) = \mathbb{E} \sup_{u \in S} \langle g, u \rangle, \quad \text{where } g \sim N(0, I_n).$$

⁵The only (minor) difference with our former definition (3.2) of the mean width is that we take supremum over S instead of $S - S$, so $\bar{w}(S)$ is a smaller quantity. The reason we do not need to consider $S - S$ because we already subtracted x in the definition of the descent cone.

Theorem 9.1 (Escape through a mesh). *Let S be a fixed subset of S^{n-1} . Let E be a random subspace of \mathbb{R}^n of a fixed codimension m , drawn from the Grassmanian $G_{n,n-m}$ according to the Haar measure. Assume that*

$$\bar{w}(S) < \sqrt{m}.$$

Then

$$S \cap E = \emptyset$$

with high probability, namely $1 - 2.5 \exp[-(m/\sqrt{m+1} - \bar{w}(S))^2/18]$.

Applying this result for $S = S(K, x)$ and $E = \ker(A)$, we conclude by the argument above that the exact recovery condition (9.1) holds with high probability if

$$m > \bar{w}(S)^2.$$

How can we *algorithmically* recover \mathbf{x} in these circumstances? We can do the same as in Section 4.1, either using the feasibility program (4.1) or, better yet, the optimization program (4.2). The only difference is that the diameter of the intersection is now zero, so the recovery is exact. The following is an exact version of Theorem 4.2.

Theorem 9.2 (Exact recovery from linear observations). *Choose $\hat{\mathbf{x}}$ to be a solution of the program*

$$\text{minimize } \|\mathbf{x}'\|_K \quad \text{subject to } A\mathbf{x}' = \mathbf{y}.$$

Assume that the number of observations satisfies

$$m > \bar{w}(S)^2 \tag{9.3}$$

where $S = S(K, x)$ is the spherical part of the descent cone of K , defined in (9.2). Then

$$\hat{\mathbf{x}} = \mathbf{x}$$

with high probability (the same as in Theorem 9.1). □

Note the familiar condition (9.3) on m which we have seen before, see e.g. Section 4.3. Informally, it states the following:

Exact recovery is possible when the number of measurements exceeds the effective dimension of the descent cone.

Remarkably, the condition (9.3) does not have absolute constant factors which we had in results before.

9.3. Application: exact sparse recovery. Let us illustrate how Theorem 9.2 works for *exact sparse recovery*. Assume that \mathbf{x} is s -sparse, i.e. it has at most s non-zero coefficients. For the feasible set, we can choose $K := \|\mathbf{x}\|_1 B_1^n = \{\mathbf{x}' : \|\mathbf{x}'\|_1 \leq \|\mathbf{x}\|_1\}$. One can write down accurately an expression for the descent cone, and derive a familiar bound on the mean width of $S = S(K, x)$:

$$\bar{w}(S) \leq C\sqrt{s \log n};$$

see [60] for details. We plug this into Theorem 9.2, where we replace $\|\mathbf{x}'\|_K$ in the optimization problem by the proportional quantity $\|\mathbf{x}'\|_1$. This leads to the following exact version of Corollary 7.3:

Theorem 9.3 (Exact sparse recovery). *Assume that an unknown vector $\mathbf{x} \in \mathbb{R}^n$ is s -sparse. Choose $\hat{\mathbf{x}}$ to be a solution to the convex program*

$$\text{minimize } \|\mathbf{x}'\|_1 \quad \text{subject to } A\mathbf{x}' = \mathbf{y}.$$

Assume that the number of observations satisfies $m > Cs \log n$. Then

$$\hat{\mathbf{x}} = \mathbf{x}$$

with high probability, namely $1 - 2.5e^{-m}$. \square

Due to the remarkable sharpness of Gordon's theorem, one may hope to obtain *sharp* conditions on the number of observations m (without absolute constants). This was done in [19] for the sparse recovery problem, and more recently in [3] for general feasible cones. The latter paper proves a variant of Gordon's theorem with a slightly different (but still closely related) version of mean width.

10. LOW-RANK MATRIX RECOVERY AND MATRIX COMPLETION

10.1. Background: matrix norms. The theory we developed so far concerns estimation of *vectors* in \mathbb{R}^n . It should not be surprising that this theory can also be applied for *matrices*. Matrix estimation problems were studied recently in particular in [8, 9, 34, 10, 59].

Let us recall some basic facts about matrices and their norms. We can identify $d_1 \times d_2$ matrices with vectors in $\mathbb{R}^{d_1 \times d_2}$. The ℓ_2 norm in $\mathbb{R}^{d_1 \times d_2}$ is then nothing else than *Frobenius* (or Hilbert-Schmidt) norm of matrices:

$$\|X\|_F = \left(\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |X_{ij}|^2 \right)^{1/2}.$$

The inner product in $\mathbb{R}^{d_1 \times d_2}$ can be written in matrix form as follows:

$$\langle X, Y \rangle = \text{tr}(X^\top Y).$$

Denote $d = \min(d_1, d_2)$. Let

$$s_1(X) \geq s_2(X) \geq \cdots \geq s_d(X) \geq 0$$

denote the *singular values* of X . Then Frobenius norm has the following spectral representation:

$$\|X\|_F = \left(\sum_{i=1}^d s_i(X)^2 \right)^{1/2}.$$

Recall also the *operator norm* of X , which is

$$\|X\| = \max_{\mathbf{u} \in \mathbb{R}^n \setminus \{0\}} \frac{\|X\mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \max_{i=1, \dots, d} s_i(X).$$

Finally, the *nuclear norm* of X is defined as

$$\|X\|_* = \sum_{i=1}^d s_i(X).$$

Spectrally, i.e. on the level of singular values, the nuclear norm is a version of ℓ_1 norm for matrices, the Frobenius norm is a version of ℓ_2 norm for matrices, and the operator norm is a version of ℓ_∞ norm for matrices. In particular, the following inequality holds:

$$\|X\| \leq \|X\|_F \leq \|X\|_*.$$

The reader should be able to derive many other useful inequalities in a similar way, for example

$$\|X\|_* \leq \sqrt{\text{rank}(X)} \cdot \|X\|_F, \quad \|X\|_F \leq \sqrt{\text{rank}(X)} \cdot \|X\| \quad (10.1)$$

and

$$\langle X, Y \rangle \leq \|X\| \cdot \|Y\|_*. \quad (10.2)$$

10.2. Low-rank matrix recovery. We are ready to formulate a matrix version of the sparse recovery problem from Section 7. Our goal is to estimate an unknown $d_1 \times d_2$ matrix X from m linear observations given by

$$y_i = \langle A_i, X \rangle, \quad i = 1, \dots, m. \quad (10.3)$$

Here A_i are independent $d_1 \times d_2$ Gaussian matrices with all i.i.d. $N(0, 1)$ entries.

There are two natural matrix versions of sparsity. The first version is the sparsity of entries. We will be concerned with the other, spectral, type of sparsity, where there are only a few non-zero singular values. This simply means that the matrix has *low rank*. So let us assume that the unknown matrix X satisfies

$$\text{rank}(X) \leq r \quad (10.4)$$

for some fixed (and possibly unknown) $r \leq n$.

The following is a matrix version of Corollary 7.3; for simplicity we are stating it in a noise-free setting ($\varepsilon = 0$).

Theorem 10.1 (Low-rank matrix recovery). *Choose \hat{X} to be a solution to the convex program*

$$\text{minimize } \|X'\|_* \text{ subject to } \langle A_i, X' \rangle = y_i, \quad i = 1, \dots, m. \quad (10.5)$$

Then

$$\mathbb{E} \sup_X \|\hat{X} - X\|_F \leq 4\sqrt{\pi} \sqrt{\frac{r(d_1 + d_2)}{m}} \cdot \|X\|_F.$$

Here the supremum is taken over all $d_1 \times d_2$ matrices X of rank at most r .

The proof of Theorem 10.1 will closely follow its vector prototype, that of Theorem 7.1; we will just need to replace the ℓ_1 norm by the nuclear norm. The only real difference will be in the computation of the *mean width of the unit ball of the nuclear norm*. This computation will be based on Y. Gordon's bound on the operator norm of Gaussian random matrices, see Theorem 5.32 in [65].

Theorem 10.2 (Gordon's bound for Gaussian random matrices). *Let G be an $d_1 \times d_2$ matrix whose entries are i.i.d. mean zero random variables. Then*

$$\mathbb{E} \|G\| \leq \sqrt{d_1} + \sqrt{d_2}.$$

Proposition 10.3 (Mean width of the unit ball of nuclear norm). *Consider the unit ball in the space of $d_1 \times d_2$ matrices corresponding to the nuclear norm:*

$$B := \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_* \leq 1\}.$$

Then

$$w(B) \leq 2(\sqrt{d_1} + \sqrt{d_2}).$$

Proof. By definition and symmetry of B , we have

$$w(B) = \mathbb{E} \sup_{X \in B-B} \langle G, X \rangle = 2 \mathbb{E} \sup_{X \in B} \langle G, X \rangle,$$

where G is a $d_1 \times d_2$ Gaussian random matrix with $N(0, 1)$ entries. Using inequality (10.2) and definition of B , we obtain we obtain

$$w(B) \leq 2 \mathbb{E} \sup_{X \in B} \|G\| \cdot \|X\|_* \leq 2 \mathbb{E} \|G\|.$$

To complete the proof, it remains to apply Theorem 10.2. \square

Let us mention an immediate consequence of Proposition 10.3, although it will not be used in the proof of Theorem 10.1.

Proposition 10.4 (Mean width of the set of low-rank matrices). *Let*

$$D = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F = 1, \text{rank}(X) \leq r\}.$$

Then

$$w(D) \leq C\sqrt{r(d_1 + d_2)}.$$

Proof of Proposition 10.4. The bound follows immediately from Proposition 10.3 and the first inequality in (10.1), which implies that $D \subset \sqrt{r} \cdot B$. \square

Proof of Theorem 10.1. The argument is a matrix version of the proof of Theorem 7.1. We consider the following subsets of $d_1 \times d_2$ matrices:

$$\bar{K} := \{X' : \|X'\|_* \leq 1\}, \quad K := \|X\|_* \cdot \bar{K}.$$

Then obviously $X \in K$, so it makes sense to apply Theorem 6.2 (with $\varepsilon = 0$) for K . It should also be clear that the optimization program in Theorem 6.2 can be written in the form (10.5).

Applying Theorem 6.2, we obtain

$$\mathbb{E} \sup_X \|\hat{X} - X\|_F \leq \sqrt{2\pi} \cdot \frac{w(K)}{\sqrt{m}}.$$

Recalling the definition of K and using Proposition 10.3 to bound its mean width, we have

$$w(K) = w(\bar{K}) \cdot \|X\|_* \leq 2\sqrt{2} \sqrt{d_1 + d_2} \cdot \|X\|_*.$$

It follows that

$$\mathbb{E} \sup_X \|\hat{X} - X\|_F \leq 4\sqrt{\pi} \sqrt{\frac{d_1 + d_2}{m}} \cdot \|X\|_*.$$

It remains to use the low-rank assumption (10.4). According to the first inequality in (10.1), we have

$$\|X\|_* \leq \sqrt{r} \|X\|_F.$$

This completes the proof of Theorem 10.1. \square

10.3. Low-rank matrix recovery: some extensions.

10.3.1. *From exact to effective low rank.* The exact low rank assumption (10.4) can be replaced by approximate low rank assumption. This is a matrix version of a similar observation about sparsity which we made in Section 7.2.2. Indeed, our argument shows that Theorem 10.1 will hold if we replace the rank by the more flexible *effective rank*, defined for a matrix X as

$$r(X) = (\|X\|_* / \|X\|_F)^2.$$

The effective rank is clearly bounded by the algebraic rank, and it is robust with respect to small perturbations.

10.3.2. *Noisy and sub-gaussian observations.* Our argument makes it easy to allow noise in the observations (10.3), i.e. consider observations of the form $y_i = \langle A_i, X \rangle + \nu_i$. We leave details to the interested reader.

Further, just like in Section 8, we can relax the requirement that A_i be Gaussian random matrices, replacing it with a *sub-gaussian* assumption. Namely, it is enough to assume that the columns of A_i are i.i.d., mean zero, isotropic and sub-gaussian random vectors in \mathbb{R}^{d_1} , with a common bound on the sub-gaussian norm. We again leave details to the interested reader.

We can summarize the results about low-rank matrix recovery as follows.

Using convex programming, one can approximately recover a $d_1 \times d_2$ matrix which has rank (or effective rank) r , from $m \sim r(d_1 + d_2)$ random linear observations.

To understand this number of observations better, note that it is of the same order as the number of degrees of freedom in the set of $d_1 \times d_2$ matrices or rank r .

10.4. Matrix completion. Let us now consider a different, and perhaps more natural, model of observations of matrices. Assume that we are given a *small random sample of entries* of an unknown matrix X . Our goal is to estimate X from this sample. As before, we assume that X has low rank. This is called a *matrix completion problem*, and it was extensively studied recently [8, 9, 34, 59].

We can not apply the previously developed theory for such observations. While sampling of entries is a linear operation, such observations are not Gaussian or sub-gaussian (more accurately, we should say that the sub-gaussian norm of such observations is too large).

Nevertheless, it is possible able to derive a matrix completion result in this setting. Our exposition will be based on a direct argument and simple from [57].

Let us formalize the process of sampling the entries of X . First, we fix the average size m of the sample. Then we generate selectors $\delta_{ij} \in \{0, 1\}$ for each entry of X . Those are i.i.d. random variables with

$$\mathbb{E} \delta_{ij} = \frac{m}{d_1 d_2} =: p.$$

Our observations are given as the $d_1 \times d_2$ matrix Y whose entries are

$$Y_{ij} = \delta_{ij} X_{ij}.$$

Therefore, the observations are randomly and independently sampled entries of X along with the indices of these entries; the average sample size is fixed and equals m . We will require that

$$m \geq d_1 \log d_1, \quad m \geq d_2 \log d_2. \quad (10.6)$$

These restrictions ensure that, with high probability, the sample contains at least one entry from each row and each column of X (recall the classical coupon collector's problem).

As before, we assume that

$$\text{rank}(X) \leq r.$$

The next result that X can be estimated from Y using low-rank approximation.

Theorem 10.5 (Matrix completion). *Choose \hat{X} to be best rank- r approximation⁶ of $p^{-1}Y$. Then*

$$\mathbb{E} \frac{1}{\sqrt{d_1 d_2}} \|\hat{X} - X\|_F \leq C \sqrt{\frac{r(d_1 + d_2)}{m}} \|X\|_\infty, \quad (10.7)$$

where $\|X\|_\infty = \max_{i,j} |X_{ij}|$.

⁶Formally, consider the singular value decomposition $p^{-1}Y = \sum_i s_i \mathbf{u}_i \mathbf{v}_i^\top$ with non-increasing singular values s_i . We define \hat{X} by retaining the r leading terms of this decomposition, i.e. $\hat{X} = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$.

To understand the form of this estimate, note that the left side of (10.7) measures the *average error per entry* of X :

$$\frac{1}{\sqrt{d_1 d_2}} \|\hat{X} - X\|_F = \left(\frac{1}{d_1 d_2} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |\hat{X}_{ij} - X_{ij}|^2 \right)^{1/2}.$$

So, Theorem 10.5 allows to make the average error per entry arbitrarily smaller than the maximal entry of the matrix. Such estimation succeeds with a sample of $m \sim r(d_1 + d_2)$ entries of X .

The proof of Theorem 10.5 will be based on a known bound on the operator norm of random matrices, which is more general than Y. Gordon's Theorem 10.2. There are several ways to obtain general bounds; see [65] for a systematic treatment of this topic. We will use one such result due to Y. Seginer [61].

Theorem 10.6 (Seginer's bound for general random matrices). *Let G be an $d_1 \times d_2$ matrix whose entries are i.i.d. mean zero random variables. Then*

$$\mathbb{E} \|G\| \leq C \left(\mathbb{E} \max_i \|G_i\|_2 + \mathbb{E} \max_j \|G^j\|_2 \right)$$

where the maxima are taken over all rows G_i and over all columns G^j of G , respectively.

Proof of Theorem 10.5. We shall first control the error in the operator norm. By triangle inequality,

$$\|\hat{X} - X\| \leq \|\hat{X} - p^{-1}Y\| + \|p^{-1}Y - X\|. \quad (10.8)$$

Since \hat{X} is the best rank- r approximation to $p^{-1}Y$, and both X and \hat{X} are rank- r matrices, the first term in (10.8) is bounded by the second term. Thus

$$\|\hat{X} - X\| \leq 2\|p^{-1}Y - X\| = \frac{2}{p}\|Y - pX\|. \quad (10.9)$$

The matrix $Y - pX$ has independent mean zero entries, namely

$$(Y - pX)_{ij} = (\delta_{ij} - p)X_{ij}.$$

So we can apply Y. Seginer's Theorem 10.6, which yields

$$\mathbb{E} \|Y - pX\| \leq C \left(\mathbb{E} \max_{i \leq d_1} \|(Y - pX)_i\|_2 + \mathbb{E} \max_{j \leq d_2} \|(Y - pX)^j\|_2 \right). \quad (10.10)$$

It remains to bound the ℓ_2 norms of rows and columns of $Y - pX$. Let us do this for rows; a similar argument would control the columns. Since

$$\|(Y - pX)_i\|_2^2 = \sum_{j=1}^{d_2} (\delta_{ij} - p)^2 |X_{ij}|^2 \leq \sum_{j=1}^{d_2} (\delta_{ij} - p)^2 \cdot \|X\|_\infty,$$

this quantity can be bounded using concentration inequalities for sums of independent random variables. In particular, we can use Bernstein's inequality (see [6]), which yields

$$\mathbb{P} \left\{ \sum_{j=1}^{d_2} (\delta_{ij} - p)^2 > pd_2 t \right\} \leq \exp(-cpd_2 t), \quad t \geq 2.$$

The first restriction in (10.6) guarantees that $pd_2 \geq \log d_1$. This enables us to use the union bound over $i \leq d_1$, which yields

$$\mathbb{E} \max_{i \leq d_1} \left[\sum_{j=1}^{d_2} (\delta_{ij} - p)^2 \right]^{1/2} \leq C_1 pd_2.$$

This translates into the following bound for the rows of $Y - PX$:

$$\mathbb{E} \max_{i \leq d_1} \|(Y - pX)_i\|_2 \leq C_1 \sqrt{pd_2} \|X\|_\infty.$$

Repeating this argument for columns and putting the two bounds into (10.10), we obtain

$$\|Y - pX\| \leq C_2 \sqrt{p(d_1 + d_2)} \|X\|_\infty.$$

Substituting into (10.9), we conclude that

$$\|\hat{X} - X\| \leq C_3 \sqrt{\frac{d_1 + d_2}{p}} \|X\|_\infty. \quad (10.11)$$

It remains to pass to the Frobenius norm. This is where we use the low rank assumption on X . Since both X and \hat{X} have ranks bounded by r , we have $\text{rank}(\hat{X} - X) \leq 2r$. Then, according to the second inequality in (10.1),

$$\|\hat{X} - X\|_F \leq \sqrt{2r} \|\hat{X} - X\|.$$

Combining this with (10.11) and recalling that $p = m/(d_1 d_2)$ by definition, we arrive at the desired bound (10.7). \square

Remark 10.7 (Noisy observations). One can easily extend Theorem 10.5 for noisy sampling, where every observed entry of X is independently corrupted by a mean-zero noise. Formally, we assume that the entries of the observation matrix Y are

$$Y_{ij} = \delta_{ij}(X_{ij} + \nu_{ij})$$

where ν_{ij} are independent and mean zero random variables. Let us further assume that $|\nu_{ij}| \leq M$ almost surely. Then a slight modification of the proof of Theorem 10.5 yields the following error bound:

$$\mathbb{E} \frac{1}{\sqrt{d_1 d_2}} \|\hat{X} - X\|_F \leq C \sqrt{\frac{r(d_1 + d_2)}{m}} (\|X\|_\infty + M).$$

We leave details to the interested reader.

11. SINGLE-BIT OBSERVATIONS VIA HYPERPLANE TESSELLATIONS

It may perhaps be surprising that a theory of similar strength can be developed for estimation problems with *non-linear* observations, in which the observation vector $\mathbf{y} \in \mathbb{R}^m$ depends non-linearly on the unknown vector $\mathbf{x} \in \mathbb{R}^n$.

In this and next sections we explore an example of extreme non-linearity – the one given by the sign function. In Section 13, we will extend the theory to completely general non-linearities.

11.1. Single-bit observations. As before, our goal is to estimate an unknown vector \mathbf{x} that lies in a known feasible set $K \subset \mathbb{R}^n$, from a random observation vector $\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$. This time, we will work with *single-bit observations* $y_i \in \{-1, 1\}$. So, we assume that

$$y_i = \text{sign} \langle \mathbf{a}_i, \mathbf{x} \rangle, \quad i = 1, \dots, m, \quad (11.1)$$

where \mathbf{a}_i are standard Gaussian random vectors, i.e. $\mathbf{a}_i \sim N(0, I_n)$. We can represent the model in a matrix form:

$$\mathbf{y} = \text{sign}(A\mathbf{x}),$$

where A is an $m \times n$ Gaussian random matrix with rows \mathbf{a}_i , and where our convention is that the sign function is applied to each coordinate of the vector $A\mathbf{x}$.

The single-bit model represents an extreme *quantization* of the linear model we explored before, where $\mathbf{y} = A\mathbf{x}$. Only one bit is retained from each linear observation y_i . Yet we hope to estimate \mathbf{x} as accurately as if all bits were available.

The model of single-bit observations was first studied in this context in [7]. Our discussion will follow [55].

11.2. Hyperplane tessellations. Let us try to understand single-bit observations y_i from a geometric perspective. Each $y_i \in \{-1, 1\}$ represents the orientation of the vector \mathbf{x} with respect to the hyperplane with normal \mathbf{a}_i . There are m such hyperplanes. The observation vector $\mathbf{y} = (y_1, \dots, y_m)$ represents orientation of \mathbf{x} with respect to all these hyperplanes.

Geometrically, the m hyperplanes induce a *tessellation* of \mathbb{R}^n by *cells*. A cell is a set of points that have the same orientation with respect to all hyperplanes; see Figure 9. Knowing \mathbf{y} is the same as knowing the cell where \mathbf{x} lies.

How can we estimate \mathbf{x} ? Recall that we know two pieces of information about \mathbf{x} :

1. \mathbf{x} lies in a known cell of the hyperplane tessellation;
2. \mathbf{x} lies in a known set K .

Therefore, a good estimator of \mathbf{x} can be obtained by picking any vector $\hat{\mathbf{x}}$ from the *intersection of these two sets*. Moreover, since just these two

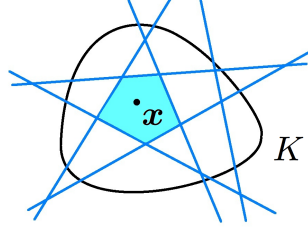


Figure 9. A tessellation of the feasible set K by hyperplanes. The cell containing \mathbf{x} is highlighted.

pieces of information about \mathbf{x} are available, such estimator is best possible in some sense.

11.3. M^* bound for random tessellations. How good is such estimate? The maximal error is of course the diameter of the intersection of the cell with K . So in order to bound the error, we need to prove that this diameter is small.

Note that our strategy is parallel to what we have done for linear observations in Section 4.1. The only piece we are missing is a version of M^* bound for random tessellations instead of random subspaces. Informally, we need a result about the following question:

Question 11.1 (Pizza cutting). *How many random hyperplanes would cut a given set K into pieces that are at most ε in size?*

A result about this problem was proved in [55].

Theorem 11.2 (M^* bound for random tessellations). *Consider a set $K \subseteq S^{n-1}$ and m independent random hyperplanes drawn uniformly from the Grassmanian $G_{n,n-1}$. Then*

$$\mathbb{E} \max_{\mathcal{C}} \text{diam}(K \cap \mathcal{C}) \leq \left[\frac{Cw(K)}{\sqrt{m}} \right]^{1/2},$$

where the maximum is taken over all cells \mathcal{C} of the hyperplane tessellation.

Apart from the exponent $1/2$ which is unlikely to be optimal, this result is indeed a version of the M^* bound, Theorem 3.12. To further highlight the similarity, note that when $m < n$, the intersection of the m random hyperplanes is a random linear subspace E of codimension m . This subspace lies in each cell of the tessellation. So in particular, Theorem 11.2 controls the quantity $\mathbb{E} \text{diam}(K \cap E)$ appearing in the standard M^* bound, Theorem 3.12.

11.4. Estimation based on M^* bound for random tessellations. Now we can apply Theorem 11.2 for the estimation problem. Based on our discussion in Section 11.2, this result immediately implies the following.

Theorem 11.3 (Estimation from single-bit observations: feasibility program). *Assume the unknown vector \mathbf{x} lies in some known set $K \subseteq S^{n-1}$, and the single-bit observation vector \mathbf{y} is given by (11.1). Choose $\hat{\mathbf{x}}$ to be any vector satisfying*

$$\hat{\mathbf{x}} \in K \quad \text{and} \quad \text{sign}(A\hat{\mathbf{x}}) = \mathbf{y}. \quad (11.2)$$

Then

$$\mathbb{E} \sup_{\mathbf{x} \in K} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \left[\frac{Cw(K)}{\sqrt{m}} \right]^{1/3}. \quad \square$$

We assumed in this result that feasible set K lies on the unit sphere. This is because the magnitude $\|\mathbf{x}\|_2$ is obviously lost in the single-bit observations. So we can only hope to estimate the direction of \mathbf{x} , which is the vector $\mathbf{x}/\|\mathbf{x}\|_2$ on the unit sphere.

A good news is that estimation can be made from $m \sim w(K)^2$ single-bit observations, the same as for linear observations. So, perhaps surprisingly, the essential information about \mathbf{x} is contained in a single bit of each observation.

A bad news is that the feasibility program (11.2) is *not convex*. When K is restricted to lie on the sphere, it can never be convex or be convexified. One can get around this issue, for example, by lifting the restriction; see [55] for pizza-cutting of general sets in \mathbb{R}^n .

But a better idea will be to replace the feasibility problem (11.2) by an optimization problem – just like we did in Section 4.2 – which will work for general sets K in the unit ball B_2^n rather than the unit sphere. Such sets can be convexified. We will do this in the next section.

12. SINGLE-BIT OBSERVATIONS VIA OPTIMIZATION, AND APPLICATIONS TO LOGISTIC REGRESSION

Our goal remains the same as we described in Section 11.1. We would like to estimate a vector \mathbf{x} that lies in a known feasible set $K \subset \mathbb{R}^n$, from single-bit observations given as

$$\mathbf{y} = \text{sign}(A\mathbf{x}) \in \{-1, 1\}^m.$$

Instead of formulating estimation as a feasibility problem (11.2), we will now state it as an *optimization* problem, as follows:

$$\text{maximize } \langle A\mathbf{x}', \mathbf{y} \rangle \text{ subject to } \mathbf{x}' \in K. \quad (12.1)$$

This program tries to fit linear observations $A\mathbf{x}'$ to the single-bit observations \mathbf{y} . It does so by maximizing the correlation between linear and single-bit observations while searching inside the feasible set K .

If K is a convex set, (12.1) is a convex program. Otherwise one can convexify K as we did several times before.

The following result from [56] provides a guarantee for such estimator.

Theorem 12.1 (Estimation from single-bit observations: optimization program). *Assume the unknown vector $\mathbf{x} \in \mathbb{R}^n$ satisfies $\|\mathbf{x}\|_2 = 1$ and \mathbf{x} lies in some known set $K \subseteq B_2^n$. Choose $\hat{\mathbf{x}}$ to be a solution to the program (12.1). Then*

$$\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{Cw(K)}{\sqrt{m}}.$$

Here $C = \sqrt{8\pi} \approx 5.01$.

Our proof of Theorem 12.1 will be based on properties of the *loss function*, which we define as

$$L_{\mathbf{x}}(\mathbf{x}') = -\frac{1}{m} \langle A\mathbf{x}', \mathbf{y} \rangle = -\frac{1}{m} \sum_{i=1}^m y_i \langle \mathbf{a}_i, \mathbf{x}' \rangle.$$

The index \mathbf{x} indicates that the loss function depends on \mathbf{x} through \mathbf{y} . The negative sign is chosen so that program (12.1) minimizes the loss function over K .

We will now compute the expected value and the deviation of the loss function for fixed \mathbf{x} and \mathbf{x}' .

Lemma 12.2 (Expectation of loss function). *Let $\mathbf{x} \in S^{n-1}$ and $\mathbf{x}' \in \mathbb{R}^n$. Then*

$$\mathbb{E} L_{\mathbf{x}}(\mathbf{x}') = -\sqrt{\frac{2}{\pi}} \langle \mathbf{x}, \mathbf{x}' \rangle.$$

Proof. We have

$$\mathbb{E} L_{\mathbf{x}}(\mathbf{x}') = -\mathbb{E} y_1 \langle \mathbf{a}_1, \mathbf{x}' \rangle = -\mathbb{E} \text{sign}(\langle \mathbf{a}_1, \mathbf{x} \rangle) \langle \mathbf{a}_1, \mathbf{x}' \rangle.$$

It remains to note that $\langle \mathbf{a}_1, \mathbf{x} \rangle$ and $\langle \mathbf{a}_1, \mathbf{x}' \rangle$ is a pair of normal random variables with zero mean, variances $\|\mathbf{x}\|_2^2 = 1$ and $\|\mathbf{x}'\|_2^2$ respectively, and covariance $\langle \mathbf{x}, \mathbf{x}' \rangle$. A simple calculation renders the expectation above as $-\langle \mathbf{x}, \mathbf{x}' \rangle \cdot \mathbb{E} \text{sign}(g)g$ where $g \sim N(0, 1)$. It remains to recall that $\mathbb{E} \text{sign}(g)g = \mathbb{E} |g| = \sqrt{2/\pi}$. \square

Lemma 12.3 (Uniform deviation of loss function). *We have*

$$\mathbb{E} \sup_{\mathbf{u} \in K-K} |L_{\mathbf{x}}(\mathbf{u}) - \mathbb{E} L_{\mathbf{x}}(\mathbf{u})| \leq \frac{2w(K)}{\sqrt{m}}. \quad (12.2)$$

Proof. Due to the form of loss function, we can apply the symmetrization inequality of Proposition 5.2, which bounds the left side of (12.2) by

$$\frac{2}{m} \mathbb{E} \sup_{\mathbf{u} \in K-K} \left| \sum_{i=1}^m \varepsilon_i y_i \langle \mathbf{a}_i, \mathbf{u} \rangle \right| = \frac{2}{m} \mathbb{E} \sup_{\mathbf{u} \in K-K} \left| \left\langle \sum_{i=1}^m \varepsilon_i y_i \mathbf{a}_i, \mathbf{u} \right\rangle \right|. \quad (12.3)$$

By symmetry and since $y_i \in \{-1, 1\}$, the random vectors $\{\varepsilon_i y_i \mathbf{a}_i\}$ are distributed identically with $\{\mathbf{a}_i\}$. In other words, we can remove $\varepsilon_i y_i$ from (12.3) without changing the value of the expectation.

Next, by rotation invariance, $\sum_{i=1}^m \mathbf{a}_i$ is distributed identically with $\sqrt{m} \mathbf{g}$, where $\mathbf{g} \sim N(0, I_n)$. Therefore, the quantity in (12.3) equals

$$\frac{2}{\sqrt{m}} \mathbb{E} \sup_{\mathbf{u} \in K-K} |\langle \mathbf{g}, \mathbf{u} \rangle| = \frac{2w(K)}{\sqrt{m}}.$$

This completes the proof. \square

Proof of Theorem 12.1. Fix $\mathbf{x}' \in K$. Let us try to bound $\|\mathbf{x} - \mathbf{x}'\|_2$ in terms of $L_{\mathbf{x}}(\mathbf{x}) - L_{\mathbf{x}}(\mathbf{x}')$. By linearity of the loss function, we have

$$L_{\mathbf{x}}(\mathbf{x}) - L_{\mathbf{x}}(\mathbf{x}') = L_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') = \mathbb{E} L_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') + D_{\mathbf{x}} \quad (12.4)$$

where the deviation

$$D_{\mathbf{x}} := \sup_{\mathbf{u} \in K-K} |L_{\mathbf{x}}(\mathbf{u}) - \mathbb{E} L_{\mathbf{x}}(\mathbf{u})|$$

will be controlled using Lemma 12.3 a bit later.

To compute the expected value in (12.4), we can use Lemma 12.2 along with the conditions $\|\mathbf{x}\|_2 = 1$, $\|\mathbf{x}'\|_2 \leq 1$ (the latter holds since $\mathbf{x}' \in K \subseteq B_2^n$). This way we obtain

$$\mathbb{E} L_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') = -\sqrt{\frac{2}{\pi}} \langle \mathbf{x}, \mathbf{x} - \mathbf{x}' \rangle \leq -\frac{1}{2} \sqrt{\frac{2}{\pi}} \|\mathbf{x} - \mathbf{x}'\|_2^2.$$

Putting this into (12.4), we conclude that

$$L_{\mathbf{x}}(\mathbf{x}) - L_{\mathbf{x}}(\mathbf{x}') \leq -\frac{1}{\sqrt{2\pi}} \|\mathbf{x} - \mathbf{x}'\|_2^2 + D_{\mathbf{x}}. \quad (12.5)$$

This bound holds for any fixed $\mathbf{x}' \in K$ and for any point in the probability space (i.e. for any realization of the random variables appearing in this bound). Therefore (12.5) must hold for the random vector $\mathbf{x}' = \hat{\mathbf{x}}$, again for any point in the probability space.

The solution $\hat{\mathbf{x}}$ was chosen to minimize the loss function, thus $L_{\mathbf{x}}(\hat{\mathbf{x}}) \leq L_{\mathbf{x}}(\mathbf{x})$. This means that for $\mathbf{x}' = \hat{\mathbf{x}}$, the left hand side of (12.5) is non-negative. Rearranging the terms, we obtain

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq \sqrt{2\pi} D_{\mathbf{x}}.$$

It remains to take expectation of both sides and use Lemma 12.3. This yields

$$\mathbb{E} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq \sqrt{2\pi} \frac{2w(K)}{\sqrt{m}}.$$

This completes the proof of Theorem 12.1. \square

12.1. Single-bit observations with general non-linearities. The specific non-linearity of observations that we considered so far – the one given by sign function – did not play a big role in our argument in the last section. The same argument, and surprisingly, the same optimization program (12.1), can serve any non-linearity in the observations.

So let us consider a general model of single-bit observations $\mathbf{y} = (y_1, \dots, y_m) \in \{-1, 1\}^m$, which satisfy

$$\mathbb{E} y_i = \theta(\langle \mathbf{a}_i, \mathbf{x} \rangle), \quad i = 1, \dots, m \quad (12.6)$$

Here $\theta : \mathbb{R} \rightarrow \mathbb{R}$ is some *link function*, which describes non-linearity of observations. We assume that y_i are independent given \mathbf{a}_i , which are standard Gaussian random vectors as before. The matrix form of this model can be written as

$$\mathbb{E} \mathbf{y} = \theta(A\mathbf{x}),$$

where A is an $m \times n$ Gaussian random matrix with rows \mathbf{a}_i , and where our convention is that the θ is applied to each coordinate of the vector $A\mathbf{x}$.

To estimate \mathbf{x} , an unknown vector in a known feasible set K , we will try to use the same optimization program (12.1) in the last section. This may be surprising since *the program does not even need to know the non-linearity θ* , nor does it attempt to estimate θ . Yes, this idea works in general as nicely as for the specific sign function. The following result from [56] is a general version of Theorem 12.1.

Theorem 12.4 (Estimation from single-bit observations with general non-linearity). *Assume the unknown vector $\mathbf{x} \in \mathbb{R}^n$ satisfies $\|\mathbf{x}\|_2 = 1$ and \mathbf{x} lies in some known set $K \subseteq B_2^n$. Choose $\hat{\mathbf{x}}$ to be a solution to the program (12.1). Then*

$$\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{4w(K)}{\lambda\sqrt{m}}.$$

Here we assume that

$$\lambda := \mathbb{E} \theta(g)g > 0 \quad \text{for } g \sim N(0, 1). \quad (12.7)$$

Proof. The argument follows very closely the proof of Theorem 12.1. The only different place is the computation of expected loss function in Lemma 12.2. When the sign function is replaced by a general non-linearity θ , one easily checks that the expected value becomes

$$\mathbb{E} L_{\mathbf{x}}(\mathbf{x}') = -\lambda \langle \mathbf{x}, \mathbf{x}' \rangle.$$

The rest of the argument is the same. \square

For $\theta(z) = \text{sign}(z)$, Theorem 12.4 is identical with Theorem 12.1. However, the new result is much more general. *Virtually no restrictions are imposed on the non-linearity θ* . In particular, θ needs not be continuous or one-to-one.

The parameter λ simply measures the information content retained through the non-linearity. It might be useful to express λ as

$$\lambda = \mathbb{E} \theta(\langle \mathbf{a}_i, \mathbf{x} \rangle) \langle \mathbf{a}_i, \mathbf{x} \rangle,$$

so λ measures how much the non-linear observations $\theta(\langle \mathbf{a}_i, \mathbf{x} \rangle)$ are correlated with linear observations $\langle \mathbf{a}_i, \mathbf{x} \rangle$.

The assumption that $\lambda > 0$ is made for convenience; if $\lambda < 0$ we can switch the sign of θ . However, if $\lambda = 0$, the non-linear and linear measurements are uncorrelated, and often no estimation is possible. An extreme example of the latter situation occurs when θ is a constant function, which clearly carries no information about \mathbf{x} .

12.2. Logistic regression, and beyond. For the link function $\theta(z) = \tanh(z/2)$, the estimation problem (12.6) is equivalent to *logistic regression with constraints*. In the usual statistical notation explained in Section 7.4, logistic regression takes the form

$$\mathbb{E} \mathbf{y} = \tanh(X\boldsymbol{\beta}/2).$$

The coefficient vector $\boldsymbol{\beta}$ is constrained to lie in some known feasible set K . We will leave it to the interested reader to translate Theorem 12.4 into the language of logistic regression, just like we did in Section 7.4 for linear regression.

The fact that Theorem 12.4 applies for general and unknown link function should be important in statistics. It means that one *does not need to know the non-linearity of the model (the link function) to make inference*. Be it the tanh function specific to logistic regression or (virtually) any other non-linearity, the estimator $\hat{\boldsymbol{\beta}}$ is the same.

13. GENERAL NON-LINEAR OBSERVATIONS VIA METRIC PROJECTION

Finally, we pass to the most general model of observations $\mathbf{y} = (y_1, \dots, y_m)$, which are not necessarily linear or single-bit. In fact, we will not even specify a dependence of y_i on \mathbf{x} . Instead, we only require that y_i be i.i.d.random variables, and

$$\text{each observation } y_i \text{ may depend on } \mathbf{a}_i \text{ only through } \langle \mathbf{a}_i, \mathbf{x} \rangle. \quad (13.1)$$

Technically, the latter requirement means that, given $\langle \mathbf{a}_i, \mathbf{x} \rangle$, the observation y_i is independent from \mathbf{a}_i . This type of observation models are called *single-index models* in statistics.

How can we estimate $\mathbf{x} \in K$ from such general observation vector \mathbf{y} ? Let us look again at the optimization problem (12.1), writing it as follows:

$$\text{maximize } \langle \mathbf{x}', A^\top \mathbf{y} \rangle \text{ subject to } \mathbf{x}' \in K.$$

It might be useful to imagine solving this program as a sequence of two steps: (a) compute a *linear estimate* of \mathbf{x} , which is

$$\hat{\mathbf{x}}_{\text{lin}} = \frac{1}{m} A^\top \mathbf{y} = \frac{1}{m} \sum_{i=1}^m y_i \mathbf{a}_i, \quad (13.2)$$

and then (b) *fitting* $\hat{\mathbf{x}}_{\text{lin}}$ to the feasible set K , which is done by choosing a point in K that is most correlated with $\hat{\mathbf{x}}_{\text{lin}}$.

Surprisingly, almost the same estimation procedure succeeds for the general single-index model (13.1). We just need to adjust the second, fitting, step. Instead of maximizing the correlation, let us metrically *project* $\hat{\mathbf{x}}_{\text{lin}}$ onto the feasible set K , thus choosing $\hat{\mathbf{x}}$ to be a solution of the program

$$\text{minimize } \|\mathbf{x}' - \hat{\mathbf{x}}_{\text{lin}}\| \text{ subject to } \mathbf{x}' \in K. \quad (13.3)$$

Just like in the previous section, it may be surprising that this estimator does not need to know the nature non-linearity in observations \mathbf{y} . To get a heuristic evidence of why this knowledge may not be needed, one can quickly check (using rotation invariance) that

$$\mathbb{E} \hat{\mathbf{x}}_{\text{lin}} = \mathbb{E} y_1 \mathbf{a}_1 = \lambda \bar{\mathbf{x}}, \quad \text{where } \bar{\mathbf{x}} = \mathbf{x} / \|\mathbf{x}\|_2, \quad \lambda = \mathbb{E} y_1 \langle \mathbf{a}_1, \bar{\mathbf{x}} \rangle.$$

So despite not knowing the non-linearity, $\hat{\mathbf{x}}_{\text{lin}}$ already provides an *unbiased estimate* of \mathbf{x} , up to scaling.

A result from [57] provides a guarantee for the two-step estimator (13.2), (13.3). Let us state this result in a special case where K is a *cone*, i.e. $tK = K$ for all $t \geq 0$. A version for general sets K is not much more difficult, see [57] for details.

Since cones are unbounded sets, the standard mean width (as defined in (3.2)) would be infinite. To get around this issue, we should consider a *local* version of mean width, which we can define as

$$w_1(K) = \mathbb{E} \sup_{\mathbf{u} \in (K-K) \cap B_2^n} \langle \mathbf{g}, \mathbf{u} \rangle, \quad \mathbf{g} \sim N(0, 1).$$

Theorem 13.1 (Estimation from non-linear observations). *Assume the unknown vector \mathbf{x} lies in a known closed cone K in \mathbb{R}^n . Choose $\hat{\mathbf{x}}$ to be a solution to the program (13.3). Let $\bar{\mathbf{x}} = \mathbf{x} / \|\mathbf{x}\|_2$. Then*

$$\mathbb{E} \hat{\mathbf{x}} = \lambda \bar{\mathbf{x}} \quad \text{and} \quad \mathbb{E} \|\hat{\mathbf{x}} - \lambda \bar{\mathbf{x}}\|_2 \leq \frac{M w_1(K)}{\sqrt{m}}.$$

Here we assume that

$$\lambda = \mathbb{E} y_1 \langle \mathbf{a}_1, \bar{\mathbf{x}} \rangle > 0 \quad \text{and} \quad M = \sqrt{2\pi} [\mathbb{E} y_1^2 + \text{Var}(y_1 \langle \mathbf{a}_1, \bar{\mathbf{x}} \rangle)]^{1/2}.$$

The proof of Theorem 13.1 is given in [57]. It is not difficult, and is close in spirit to the arguments we saw here; we will not reproduce it.

The role of parameters λ and M is to determine the correct magnitude and deviation of the estimator; one can think of them as *constants* that are usually easy to compute or estimate. By rotation invariance, λ and

M depend on the *magnitude* $\|\mathbf{x}\|_2$ (through y_1) but not on the direction $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$ of the unknown vector \mathbf{x} .

We can summarize results of this and previous section as follows.

Using convex programming, one can estimate a vector \mathbf{x} in a general feasible set K from $m \sim w(K)^2$ random non-linear observations, even if the non-linearity is not known.

13.1. Examples of observations. To give a couple concrete examples, consider *noisy linear observations*

$$y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle + \nu_i.$$

We already explored this model in Section 6, where ν_i were arbitrary numbers representing noise. This time, let us assume ν_i are independent random variables with zero mean and variance σ^2 . A quick computation gives

$$\lambda = \|\mathbf{x}\|_2, \quad M = C(\|\mathbf{x}\|_2 + \sigma).$$

Theorem 13.1 then yields the following error bound:

$$\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{Cw_1(K)}{\sqrt{m}} (\|\mathbf{x}\|_2 + \sigma).$$

Let us give one more example, for the *single-bit observations*

$$y_i = \text{sign} \langle \mathbf{a}_i, \mathbf{x} \rangle.$$

We explored this model in Sections 11 and 12. A quick computation gives

$$\lambda = \sqrt{\frac{2}{\pi}}, \quad M = C.$$

Theorem 13.1 then yields the following error bound:

$$\mathbb{E} \left\| \hat{\mathbf{x}} - \sqrt{\frac{2}{\pi}} \mathbf{x} \right\|_2 \leq \frac{Cw_1(K)}{\sqrt{m}}.$$

13.2. Examples of feasible cones. To give a couple of concrete examples of feasible cones, consider the set K of *s-sparse vectors* in \mathbb{R}^n , those with at most s non-zero coordinates. A quick computation yields

$$w_1(K) \leq C\sqrt{s \log n}.$$

Further, solving the program (13.3) (i.e. computing the metric projection of $\hat{\mathbf{x}}_{\text{lin}}$ onto K) amounts to *hard thresholding* of \mathbf{x}' . The solution $\hat{\mathbf{x}}$ is obtained from $\hat{\mathbf{x}}_{\text{lin}}$ by keeping the s largest coefficients (in absolute value) and zeroing out all other coefficients.

So Theorem 13.1 in this case can be stated informally as follows:

One can estimate an s-sparse vector \mathbf{x} in \mathbb{R}^n from $m \sim s \log n$ non-linear observations \mathbf{y} , even if the non-linearity is not known. The estimation is given by the hard thresholding of $\hat{\mathbf{x}}_{\text{lin}} = m^{-1} A^T \mathbf{y}$.

Another popular example of a feasible cone is a set of *low-rank matrices*. Let K be the set of $d_1 \times d_2$ matrices with rank at most r . Proposition 10.4 implies that

$$w_1(K) \leq C\sqrt{r(d_1 + d_2)}.$$

Further, solving the program (13.3) (i.e. computing the metric projection of \mathbf{x}' onto K) amounts to computing the best rank- r approximation of $\hat{\mathbf{x}}_{\text{lin}}$. This amounts to *hard thresholding of singular values* of $\hat{\mathbf{x}}_{\text{lin}}$, i.e. keeping the leading s terms of the singular value decomposition. This is the same e already came across this hard thresholding of singular values in the matrix completion problem, in Theorem 10.5.

So Theorem 13.1 in this case can be stated informally as follows:

One can estimate an $d_1 \times d_2$ matrix with rank r from $m \sim r(d_1 + d_2)$ non-linear observations, even if the non-linearity is not known. The estimation is given by the hard thresholding of singular values of $\hat{\mathbf{x}}_{\text{lin}}$.

14. SOME EXTENSIONS

14.1. From expectation to overwhelming probability. Many results in this survey were stated in terms of expected value. One can show that they hold also with overwhelming probability. This can be done by using *concentration of measure*, see [39]. For example, we could use Gaussian concentration inequality in the proof of M^* bound in Section 5. We would thus obtain the following high-probability version of Theorem 3.12:

$$\mathbb{P} \left\{ \text{diam}(K \cap E) > \frac{tw(K)}{\sqrt{m}} \right\} \leq \exp \left[\frac{ct^2 w(K)^2}{\text{diam}(K)^2} \right] \leq 2 \exp(-c_1 t^2), \quad t \geq 1.$$

For details, see [55].

Correspondingly, one can state high-probability versions of various estimation results. In particular, all results of this survey can be stated for the *mean-squared error* $\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$ instead of $\mathbb{E} \|\hat{\mathbf{x}} - \mathbf{x}\|_2$.

14.2. From global to local mean width. As we have seen, the concept of Gaussian mean width captures the complexity of a feasible set K quite accurately. Still, it is not exactly the optimal quantity in geometric and estimation results. An optimal quantity is the *local mean width*, which is a function of radius $r > 0$, defined as

$$w_r(K) = \mathbb{E} \sup_{\mathbf{u} \in (K - K) \cap rB_2^n} \langle \mathbf{g}, \mathbf{u} \rangle, \quad \mathbf{g} \sim N(0, I_n).$$

Comparing to Definition 3.4 of the usual mean width, we see that

$$w_r(K) \leq w(K) \quad \text{for all } r.$$

The usefulness of local mean width was noted in asymptotic convex geometry by A. Giannopoulos and V. Milman [23, 24, 25, 27]. They showed

that the function $w_r(K)$ completely describes the diameter of high dimensional sections $K \cap E$, thus proving *two-sided* versions of the M^* bound (Theorem 3.12). An observation of a similar nature was made recently by S. Chatterjee [14] in the context of high dimensional estimation. He noted that a variant of local mean width provides optimal error rates for the *metric projection* onto a feasible set considered in Section 13.

For most results discussed in this survey, one can replace the usual mean width by a local mean width, thus making them stronger. Let us briefly indicate how this can be done for the M^* bound (Theorem 3.12; see [24, 25, 27, 49] for a more detailed discussion).

Such localization is in a sense automatic; it can be done as a “post-processing” of the M^* estimate. The conclusion of the general M^* bound, Theorem 5.1, for $T \cap rB_2^n$, is that

$$\sup_{\mathbf{u} \in T_\varepsilon \cap rB_2^n} \|\mathbf{u}\|_2 \leq C \left(\frac{1}{\sqrt{m}} \mathbb{E} \sup_{\mathbf{u} \in T \cap rB_2^n} |\langle \mathbf{g}, \mathbf{u} \rangle| + \varepsilon \right) \quad (14.1)$$

with high probability (see also Section 14.1.) Let us show that the intersection with the ball rB_2^n can be automatically removed from the left side. Since

$$\sup_{\mathbf{u} \in T_\varepsilon \cap rB_2^n} \|\mathbf{u}\|_2 = \min \left(\sup_{\mathbf{u} \in T_\varepsilon} \|\mathbf{u}\|_2, r \right),$$

it follows that if $\sup_{\mathbf{u} \in T_\varepsilon \cap rB_2^n} \|\mathbf{u}\|_2 < r$ then $\sup_{\mathbf{u} \in T_\varepsilon} \|\mathbf{u}\|_2 \leq r$. Thus, if the right side of (14.1) is smaller than r , then $\sup_{\mathbf{u} \in T_\varepsilon} \|\mathbf{u}\|_2 \leq r$.

When applied to the classical M^* bound, Theorem 3.12, this argument localizes it as follows.

$$\frac{w_r(K)}{r} \leq c\sqrt{m} \quad \text{implies} \quad \text{diam}(K \cap E) \leq r$$

with high probability.

14.3. More general distributions. For simplicity of exposition, the estimation results in this survey were stated for isotropic Gaussian vectors \mathbf{a}_i . We showed in Section 8 how to extend the M^* bound and the corresponding linear estimation results for line for *sub-gaussian* distributions. For more heavy distributions, a version of M^* bound was proved recently in [43]; compressed sensing for such distributions was examined in [37, 38]. However, it seems to be unknown if Gordon’s Theorem 10.2 and the corresponding result for exact recovery in general sets (Theorem 9.2) can be extended to sub-gaussian distributions.

For single-bit observations of Section 12, a generalization for sub-gaussian distributions is discussed in [2]. Some results can be formulated for *anisotropic* Gaussian distributions, where $\mathbf{a}_i \sim N(0, \Sigma)$ with $\Sigma \neq I_n$, see e.g. [56, Section 3.4].

Results for extremely heavy-tailed distributions, such as samples of entries and random Fourier measurements, exist currently only for special cases of feasible sets K . When K consists of sparse vectors, reconstruction of \mathbf{x} from

Fourier measurements (random frequencies of \mathbf{x}) was extensively studied in compressed sensing [16, 36, 12, 32]. Reconstruction of a matrix from a random sample of entries was discussed in Section 10.4 in the context of matrix completion problem.

There are currently no results, for instance, about reconstruction of $\mathbf{x} \in K$ from random Fourier measurements, where K is a general feasible set. It is clear that K needs to be *incoherent* with the Fourier basis of exponentials, but this has yet to be quantified. In the special case where K is a set of sparse vectors, basic results of compressed sensing quantify this incoherence via a *restricted isometry property* [16, 36, 12, 32].

REFERENCES

- [1] R. Adamczak, R. Latała, A. Litvak, K. Oleszkiewicz, A. Pajor, N. Tomczak-Jaegermann, *A short proof of Paouris' inequality*, Canad. Math. Bull. 57 (2014), 3–8.
- [2] A. Ai, A. Lapanowski, Y. Plan and R. Vershynin, One-bit compressed sensing with non-Gaussian measurements, Linear Algebra and Applications 441 (2014), 222–239.
- [3] D. Amelunxen, M. Lotz, M. McCoy, J. A. Tropp, *Living on the edge: A geometric theory of phase transitions in convex optimization*, Inform. Inference, to appear.
- [4] K. Ball, *An elementary introduction to modern convex geometry*. Flavors of geometry, 1–58, Math. Sci. Res. Inst. Publ., 31, Cambridge Univ. Press, Cambridge, 1997.
- [5] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, *Structured sparsity through convex optimization*, Statist. Sci. 27 (2012), 450–468.
- [6] S. Boucheron, O. Bousquet, G. Lugosi, *Concentration inequalities*. In: Advanced Lectures in Machine Learning, edited by Bousquet, O., Luxburg, U. and Rätsch, G. Springer, 2004, pp. 208–240.
- [7] P. Boufounos, R. Baraniuk, *1-Bit compressive sensing*, Conf. on Info. Sciences and Systems (CISS), Princeton, New Jersey, March 2008.
- [8] E. Candès, B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math. 9 (2009), 717–772.
- [9] E. Candès, T. Tao, *The power of convex relaxation: near-optimal matrix completion*, IEEE Trans. Inform. Theory 56 (2010), 2053–2080.
- [10] E. Candès, X. Li, Y. Ma, J. Wright, *Robust principal component analysis?* J. ACM 58 (2011), no. 3, Art. 11, 37 pp.
- [11] E. Candès, Y. Eldar, D. Needell, P. Randall, *Compressed sensing with coherent and redundant dictionaries*, Appl. Comput. Harmon. Anal. 31 (2011), 59–73.
- [12] D. Chafaï, O. Guédon, G. Lecué, A. Pajor, *Interactions between compressed sensing random matrices and high dimensional geometry*. Panoramas et Synthèses, 37. Société Mathématique de France, Paris, 2012.
- [13] V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, *The convex geometry of linear inverse problems*, Found. Comput. Math. 12 (2012), 805–849.
- [14] S. Chatterjee, *A new perspective on least squares under convex constraint*, submitted.
- [15] S. Chen, D. Donoho, M. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. 20 (1998), 33–61.
- [16] M. Davenport, M. Duarte, Y. Eldar, G. Kutyniok, *Introduction to compressed sensing*. Compressed sensing, 1–64, Cambridge Univ. Press, Cambridge, 2012.
- [17] M. Davenport, D. Needell, M. Wakin, *Signal space CoSaMP for sparse recovery with redundant dictionaries*, IEEE Trans. Inform. Theory 59 (2013), 6820–6829.
- [18] D. Donoho, M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization*, Proc. Natl. Acad. Sci. USA 100 (2003), 2197–2202.

- [19] D. Donoho, J. Tanner, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, J. Amer. Math. Soc. 22 (2009), 1–53.
- [20] A. Dvoretzky, *A theorem on convex bodies and applications to Banach spaces*, Proc. Nat. Acad. Sci. U.S.A 45 (1959), 223–226.
- [21] A. Dvoretzky, *Some results on convex bodies and Banach spaces*, in: Proc. Sympos. Linear Spaces, Jerusalem (1961), 123–161.
- [22] A. Giannopoulos, S. Brazitikos, P. Valettas, B.-H. Vritsiou, *Geometry of isotropic convex bodies*. Mathematical Surveys and Monographs 196, AMS, 2014.
- [23] A. Giannopoulos, V. Milman, *How small can the intersection of a few rotations of a symmetric convex body be?* C. R. Acad. Sci. Paris Ser. I Math. 325 (1997), 389–394.
- [24] A. Giannopoulos, V. Milman, *On the diameter of proportional sections of a symmetric convex body*, Internat. Math. Res. Notices 1 (1997), 5–19.
- [25] A. Giannopoulos, V. D. Milman, *Mean width and diameter of proportional sections of a symmetric convex body*, J. Reine Angew. Math. 497 (1998), 113–139.
- [26] A. Giannopoulos, V. Milman, *Asymptotic convex geometry: short overview*. Different faces of geometry, 87–162, Int. Math. Ser. (N. Y.), 3, Kluwer/Plenum, New York, 2004.
- [27] A. Giannopoulos, V. D. Milman, *Asymptotic formulas for the diameter of sections of symmetric convex bodies*, J. Funct. Anal. 223 (2005), 86–108.
- [28] Y. Gordon, *On Milman’s inequality and random subspaces which escape through a mesh in \mathbb{R}^n* , Geometric aspects of functional analysis, Isr. Semin. 1986–87, Lect. Notes Math. 1317, 84–106 (1988).
- [29] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Inform. Theory 57 (2011), 1548–1566.
- [30] O. Guédon, E. Milman, *Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures*, Geom. Funct. Anal. 21 (2011), 1043–1068.
- [31] Fernique, X. *Régularité des trajectoires des fonctions aléatoires gaussiennes*, École d’été de Probabilités de Saint-Flour, IV-1974, pp. 1–96. Lecture Notes in Math., Vol. 480, Springer, Berlin, 1975.
- [32] S. Foucart, H. Rauhut, *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.
- [33] L. Jacques, J. Laska, P. Boufounos, R. Baraniuk, *Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors*, IEEE Trans. Inform. Theory 59 (2013), no. 4, 2082–2102.
- [34] R. Keshavan, A. Montanari, S. Oh, *Matrix completion from a few entries*, IEEE Trans. Inform. Theory 56 (2010), 2980–2998.
- [35] B. Klartag, *Power-law estimates for the central limit theorem for convex sets*, J. Funct. Anal. 245 (2007), 284–310.
- [36] G. Kutyniok, *Theory and applications of compressed sensing*. GAMM-Mitt. 36 (2013), 79–101.
- [37] G. Lecué, S. Mendelson, *Compressed sensing under weak moment assumptions*, submitted.
- [38] G. Lecué, S. Mendelson, *Necessary moment conditions for exact reconstruction via basis pursuit*, submitted.
- [39] M. Ledoux, *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001.
- [40] M. Ledoux, M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes*. Reprint of the 1991 edition. Classics in Mathematics. Springer-Verlag, Berlin, 2011.
- [41] S. Mendelson, *A few notes on Statistical Learning Theory*, In: Advanced Lectures in Machine Learning, (S. Mendelson, A.J. Smola Eds), LNCS 2600, 1–40, Springer 2003.
- [42] , S. Mendelson, *Geometric parameters in learning theory*. Geometric aspects of functional analysis, 193–235, Lecture Notes in Math., 1850, Springer, Berlin, 2004.

- [43] , S. Mendelson, *A remark on the diameter of random sections of convex bodies*, submitted.
- [44] M. Rudelson, R. Vershynin, *Combinatorics of random processes and sections of convex bodies*, Annals of Mathematics 164 (2006), 603–648.
- [45] V. Milman, *New proof of the theorem of Dvoretzky on sections of convex bodies*, Funct. Anal. Appl. 5 (1971), 28–37.
- [46] V. Milman, *Geometrical inequalities and mixed volumes in the Local Theory of Banach spaces*, Astérisque 131 (1985), 373–400.
- [47] V. Milman, *Random subspaces of proportional dimension of finite dimensional normed spaces: approach through the isoperimetric inequality*, Lecture Notes in Mathematics 1166 (1985), 106–115.
- [48] V. Milman, *Surprising geometric phenomena in high-dimensional convexity theory*, European Congress of Mathematics, Vol. II (Budapest, 1996), 73–91, Progr. Math., 169, Birkhäuser, Basel, 1998.
- [49] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, *Reconstruction and subgaussian operators in asymptotic geometric analysis*, Geom. Funct. Anal. 17 (2007), 1248–1282.
- [50] V. Milman, G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*. With an appendix by M. Gromov. Lecture Notes in Mathematics, 1200. Springer-Verlag, Berlin, 1986.
- [51] G. Paouris, *Concentration of mass on convex bodies*, Geom. Funct. Anal. 16 (2006), 1021–1049.
- [52] A. Pajor, N. Tomczak-Jaegermann, *Subspaces of small codimension of finite dimensional Banach spaces*, Proc. Amer. Math. Soc. 97 (1986), 637–642.
- [53] G. Pisier, *The volume of convex bodies and Banach space geometry*. Cambridge Tracts in Mathematics, 94. Cambridge University Press, Cambridge, 1989.
- [54] Y. Plan, R. Vershynin, *One-bit compressed sensing by linear programming*, Communications on Pure and Applied Mathematics 66 (2013), 1275–1297.
- [55] Y. Plan, R. Vershynin, *Dimension reduction by random hyperplane tessellations*, Discrete and Computational Geometry 51 (2014), 438–461.
- [56] Y. Plan, R. Vershynin, *Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach*, IEEE Transactions on Information Theory 59 (2013), 482–494.
- [57] Y. Plan, R. Vershynin, E. Yudovina, *High-dimensional estimation with geometric constraints*, submitted.
- [58] N. Rao, B. Recht, R. Nowak, *Tight Measurement Bounds for Exact Recovery of Structured Sparse Signals*. In: Proceedings of AISTATS, 2012.
- [59] B. Recht, *A simpler approach to matrix completion*, J. Mach. Learn. Res. 12 (2011), 3413–3430.
- [60] M. Rudelson, R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Communications on Pure and Applied Mathematics 61 (2008), 1025–1045.
- [61] Y. Seginer, *The expected norm of random matrices*, Combin. Probab. Comput. 9 (2000), 149–166.
- [62] N. Srebro, N. Alon, T. Jaakkola, *Generalization error bounds for collaborative prediction with low-rank matrices*, Advances in Neural Information Processing Systems (NIPS) 17, 2005.
- [63] M. Talagrand, *Regularity of Gaussian processes*, Acta Math. 159 (1987), 99–149.
- [64] M. Talagrand, *The generic chaining. Upper and lower bounds of stochastic processes*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.
- [65] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed sensing, 210–268, Cambridge Univ. Press, Cambridge, 2012.

- [66] M. Wainwright, *Structured regularizers for high-dimensional problems: Statistical and computational issues*, Annual Review of Statistics and its Applications 1 (2014), 233–253.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MICHIGAN, 530 CHURCH ST., ANN ARBOR, MI 48109, U.S.A.

E-mail address: `romanv@umich.edu`