CS798 Project: Atomic norm minimization for linear inverse problems

Peter A.I. Forsyth

April 17, 2017

1 Introduction

1.1 Inverse Problems

The theory of inverse problems models situations in which an observer knows the general laws governing a phenomenon but does not know certain particulars. The observer wishes to learn these particulars using the general laws and some imperfect observations. To give scientific examples, a radiologist may use knowledge of the general laws governing the attenuation of radiation and observations of the degree to which beams passing through a patient's body attenuate to draw conclusions about the composition of the patient's body; or a geologist may use knowledge of how waves propagate through the ground and a database of seismic measurements to draw conclusions about the makeup of the earth's crust. Many statistical problems are also inverse problems. A statistician may use observations of the independent and dependent variables and knowledge about how different candidate models relate those variables to draw conclusions about which model is correct.

1.2 Main Examples

This essay treats the mathematics of a particular class of inverse problem, in which a linear transformation Φ– representing the known general law– has been applied to a hidden state vector x^*– representing the unknown particulars– to produce a result vector b– representing our observations. We seek to recover x^*, but we cannot do so directly, because Φ is not injective, and b may have been corrupted by error. To assist us, we have some additional knowledge, which tells us that x^* lies in the set $\mathcal{S}$. We hope to use b, Φ, and $\mathcal{S}$ to recover x^*. To summarize

Problem 1. Let X, W be Euclidean spaces, and let $\mathcal{S} \subset X$. Our goal is to recover $x^* \in \mathcal{S}$ from the data (Φ, ϵ, b) where $\Phi : X \to W$ is linear, $\epsilon \geq 0$ and $\|\Phi x^* - b\| \leq \epsilon$.

For general $\mathcal{S}$ there is little we can do to efficiently solve Problem 1. Additional knowledge of the structure of S allows us to make headway. The following examples illustrate the kind of problems in which we are interested.

Example 1. Let $x^* \in \mathbb{R}^n$, and let $\Phi \in \mathbb{R}^{m \times n}$ where $m < n$. We are given $b \in \mathbb{R}^m$ such that $\|b - \Phi x^*\| \leq \epsilon$, and we know that x^* lies in $\mathcal{S}$, the set of vectors with at most k nonzero elements. Such vectors are called k-sparse. We wish to recover x^*.

In statistics, this example and its variants are called the LASSO[18]. The (i, j) entry of Φ is observation i of independent variable j, while the i entry of b is observation i of the dependent variable. The condition $x \in \mathcal{S}$ is a parsimony assumption; we know that the dependent variable depends on only a few independent variables, but we do not know which ones.

This sparse-recovery example also arises in engineering applications. To briefly mention one, a radar device emits an electromagnetic pulse, which interacts with objects in the environment (See chapter 1 of Foucart and Rauhut[6]). The resulting scatter is measured by a receiver. Here each entry of x^* corresponds to a particular location and speed, and is nonzero when there is an object at that location moving at that speed. Φ maps x^* to b, the data measured by the receiver, and we seek to recover x^*. Often, there are only a small number of objects in the environment, and so it is reasonable to assume x^* is sparse.

Example 2. Let $x^* \in \mathbb{R}^{n_1 \times n_2}$, and let $\Phi : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ be linear where $m < n_1 n_2$. We are given $b \in \mathbb{R}^m$ such that $\|\Phi(x^*) - b\| \leq \epsilon$, and we know that $x^* \in \mathcal{S} := \{x \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank} x \leq r\}$.

Low rank matrix recovery has applications in audio-visual data processing, and in bioinformatics, where underlying structure of the signals often makes them representable as low-rank matrices [19].

Examples 1 and 2 admit a unified solution strategy, which we introduce in Section 2.

2 The Atomic Norm Paradigm

Chandrasekaran et al.[3] introduced the approach we describe here. We assume we have a set $\mathcal{A} \subset X$, called the *atoms*. We define for positive integers k the sets

$$\mathcal{A}_k := \left\{ \sum_{i=1}^{k} \alpha_i a_i : \forall i \ \alpha_i \geq 0, a_i \in \mathcal{A} \right\}. \tag{1}$$

That is to say, $\mathcal{A}_k$ consists of the conic combinations of k elements of $\mathcal{A}$. Thus we have[1]

$$\mathcal{A} \subset \mathcal{A}_1 \subset \mathcal{A}_2 \subset \ldots \subset \operatorname{cone} \mathcal{A}. \tag{2}$$

We impose the condition that the $\mathcal{S}$ of Problem 1 satisfies $\mathcal{S} = \mathcal{A}_k$ for some k. Next, we show that examples 1 and 2 satisfy this condition.

Example 1 (Continued)**.** $\mathcal{A} = \{\pm e_1, \ldots, \pm e_n\}$ and $\mathcal{S} = \mathcal{A}_k$, the set of k-sparse vectors.

Example 2 (Continued)**.** $\mathcal{A} = \{xy^T : x \in \mathbb{R}^{n_1}, y \in \mathbb{R}^{n_2}, \|x\| = \|y\| = 1\}$. Since the sum of k rank-1 matrices has rank at most k, and every rank k matrix can be written as the sum of k rank 1 matrices (see section 51 in Halmos[11]), we have $\{A \in \mathbb{R}^{n_1 \times n_2} : \operatorname{rank}(A) \leq k\} = A_k = \mathcal{S}$.

For small k, we expect $\mathcal{A}_k$ to be nonconvex and therefore difficult to work with. It is natural to introduce a convexification to make our task more tractable. To this end, define the function $\gamma : X \to \mathbb{R}$ by

$$\gamma(x) := \inf\{\lambda > 0 : x \in \lambda \operatorname{conv} \mathcal{A}\} \tag{3}$$

[1]Since $\operatorname{cone} \mathcal{A} = \mathbb{R}_+ \operatorname{conv} \mathcal{A}$, by Carathéodory's theorem(III.1.3.6 in Hiriart-Urruty and Lemaréchal[12]) the sequence of sets in (2) does not go on endlessly. Rather, if n is the ambient dimension, then $\mathcal{A}_{n+1} = \operatorname{cone} \mathcal{A}$.

where we use the convention that $\inf \emptyset = \infty$. As the gauge function of a convex set, γ is convex and sublinear (see V.1.12 in Hiriart-Urruty and Lemaréchal[12]). Though γ need not be a norm in general, it is called the *atomic norm*, since it is a norm in examples of interest. Our aim is to approximate the solution of Problem 1 by solving

Problem 2.

$$\text{minimize } \gamma(x) \text{ subject to } \Phi x = b \tag{4}$$

when $\epsilon = 0$ and solving

Problem 3.

$$\text{minimize } \gamma(x) \text{ subject to } \|\Phi x - b\| \leq \epsilon \tag{5}$$

when $\epsilon > 0$.

Example 1 (Continued)**.** Here $\gamma = \|\cdot\|_1$. Problems 2 and 3 become

$$\text{minimize } \|x\|_1 \text{ subject to } \Phi x = b \tag{6}$$
$$\text{minimize } \|x\|_1 \text{ subject to } \|\Phi x - b\| \leq \epsilon. \tag{7}$$

According to Amelunxen et al.[1] this approach to sparse recovery was first considered by Chen et al.[4].

Example 2 (Continued)**.** Let $\|\cdot\|_*$ denote the nuclear norm, which is discussed in somewhat more detail in section B. The nuclear norm of a matrix is the sum of its singular values. In this example, we have

$$\operatorname{conv} \mathcal{A} = \{\sum_{i=1}^{q} \alpha_i x_i y_i^T,\ \|x_i\| = \|y_i\| = 1,\ \alpha_i \geq 0 \,\forall i,\ \sum_{i=1}^{q} \alpha_i = 1,\ q > 0\} \tag{8}$$
$$= \{X \in \mathbb{R}^{n_1 \times n_2} : \|X\|_* \leq 1\} =: B_*. \tag{9}$$

The $\subset$ part of the second equality follows since $\|\sum_{i=1}^{q} \alpha_i x_i y_i^T\|_* \leq \sum_{i=1}^{q} \alpha_i \|x_i y_i^T\|_* = 1$. The $\supset$ part follows since if $X \in B_*$ then the singular value decomposition of X displays X as a convex combination (If the sum of the singular values is less than one, terms of the form $xy^T - xy^T$ can be added in an appropriate quantity.)

Thus we have $\gamma = \|\cdot\|_*$. Problems 2 and 3 become

$$\text{minimize } \|x\|_* \text{ subject to } \Phi x = b \tag{10}$$
$$\text{minimize } \|x\|_* \text{ subject to } \|\Phi x - b\| \leq \epsilon. \tag{11}$$

According to Amelunxen et al.[1] this approach to low-rank recovery was first considered by Recht at el.[15].

That Problems 2 and 3 relate usefully to Problem 1 is not obvious. One might worry that substituting the convex function γ for the nonconvex set $\mathcal{A}_k$ suppresses important information. Presently, we shall provide intuitive motivation for the connection between Problem 1 and its

convexifications. Then, we shall devote the greater part of this essay to a more rigorous analysis of the connection.

To begin, for $x \in X$ define the sets

$$\mathrm{T}(x) := \mathrm{cl}\{\alpha(y-x) : \alpha \geq 0, y \in X, \gamma(y) \leq \gamma(x)\} \tag{12}$$

$$\mathrm{N}(x) := T^{\circ}(x) = \{s \in X : \langle y-x, s\rangle \leq 0 \ \forall y \in X \text{ such that } \gamma(y) \leq \gamma(x)\}. \tag{13}$$

$\mathrm{T}(x)$ is the tangent cone of the set $(\gamma(x)\,\mathrm{conv}\,\mathcal{A})$ at x. $\mathrm{N}(x)$ is the corresponding normal cone. Using $\mathrm{T}(x)$ we can establish optimality conditions for Problems 2 and 3.

Proposition 1 (From Propositions 2.1 and 2.2 in Chandrasekaran et al [3])**.**

1. If $x \in X$ satisfies $\Phi x = b$ and $\mathrm{null}\,\Phi \cap \mathrm{T}(x) = \{0\}$ then x is the unique solution to Problem 2.
2. Let $\hat{x}$ solve Problem 3. If $x \in X$ sastisfies $\|\Phi x - b\| \leq \epsilon$ and if for some $\delta > 0$ all $z \in \mathrm{T}(x)$ satisfy $\|\Phi z\| \geq \delta\|z\|$ then $\|x - \hat{x}\| \leq 2\frac{\epsilon}{\delta}$.

Proof.

1. If $\gamma(y) \leq \gamma(x)$ and $\Phi y = b$ then $y - x \in \mathrm{T}(x)$ and $y - x \in \mathrm{null}\,\Phi$ so that $y = x$.
2. Since $\gamma(\hat{x}) \leq \gamma(x)$, we have that $\hat{x} - x \in \mathrm{T}(x)$. Thus

$$\|\hat{x} - x\| \leq \frac{1}{\delta}\|\Phi(\hat{x} - x)\| \leq \frac{1}{\delta}(\|\Phi\hat{x} - b\| + \|\Phi x - b\|) \leq 2\frac{\epsilon}{\delta} \tag{14}$$

   as desired[2][3][4].

□

Remark 1. In subsequent sections, we will use Proposition 1 to connect Problem 1 to its convexifictions, Problem 2 and Problem 3. We will argue that, in the examples of interest, if $x \in \mathcal{A}_k$ for small k, then $\mathrm{N}(x)$ is large, and consequently its polar $\mathrm{T}(x)$ is small. Because $\mathrm{T}(x)$ is small, the nullspace of a randomly chosen Φ is likely to lie far from it, and so the hypotheses of Proposition 1 are likely to be satisfied. Hence x is likely to solve Problem 2 or approximately solve Problem 3. Note the importance of randomness to this argument. Our key results will hold with high probability, but not with certainty.

Example 1 (Continued)**.** We know from class that $\|\cdot\|_1$ is the support function of the unit ball of its dual norm, $\|\cdot\|_\infty$. Call this ball B_∞ We know from Proposition A.13 that this implies that $\partial\|x\|_1 = \{s \in B_\infty : \langle x, s\rangle \geq \langle x, z\rangle \forall z \in B_\infty\}$. This set is clearly $\{s \in \mathbb{R}^n : s_i = \mathrm{sign}(x_i)$ if $x_i \neq 0$, $s_i \in [-1, 1]$ if $x_i = 0\}$. By Proposition A.19,

$$\mathrm{N}(x) = \mathbb{R}_+\partial\|x\|_1 = \{s \in \mathbb{R}^n : s_i = t\,\mathrm{sign}(x_i) \text{ if } x_i \neq 0 \ , \ s_i \in [-t, t] \text{ if } x_i = 0,\ t \geq 0\}. \tag{15}$$

Thus $\mathrm{N}(x)$ grows larger as x grows sparser.

[2]The first part of Proposition 1 is actually a special case of the second part, since we can use the continuity of Φ and the compactness of the unit sphere to show that if $\mathrm{null}\,\Phi \cap \mathrm{T}(x) = \{0\}$ then there is an appropriate δ.

[3]If $\mathrm{T}(x)$ is replaced by $\{\alpha(y-x) : \alpha \geq 0, y \in X, \gamma(y) \leq \gamma(x)\}$ then the condition in the first part is both necessary and sufficient. This is the approach used in Chandrasekaran et al [3].

[4]The proof depends only on γ being convex, and not on γ being the gauge function of the convex hull of atoms. We assume γ has this later structure because such γ tend to have tangent cones with the properties we desire.

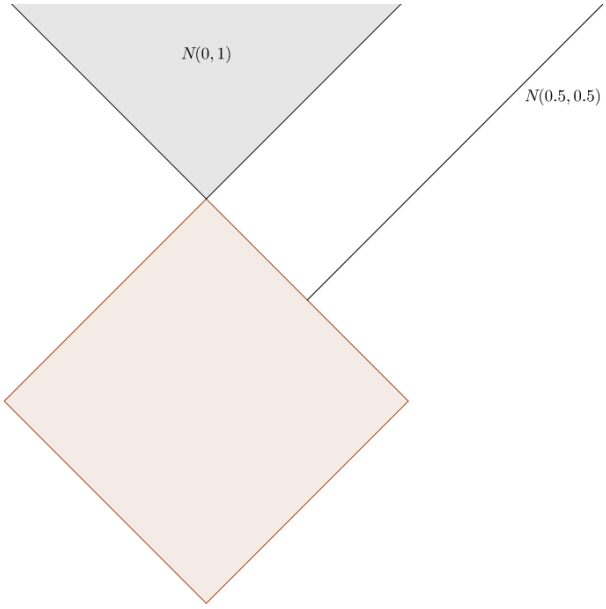

Figure 1: A simple illustration of Example 1. The normal cone of the sublevel set of $\|\cdot\|_1$ is much larger at the 1-sparse point $(0,1)$ than at the 0-sparse point $(0.5,0.5)$.

Example 2 (Continued)**.** Let UDV^T be the singular value decomposition of x, where $U \in \mathbb{R}^{m\times r}, D \in \mathbb{R}^{r\times r}, V \in \mathbb{R}^{n\times r}$ and $\text{rank}(x) = r$. Using Propositions A.19 and B.1, we have

$$\text{N}(x) = \mathbb{R}_+ \partial \|x\|_* = \{tUV^T + W : WV = 0,\, U^T W = 0,\, \|W\| \leq t,\, t \geq 0\}. \tag{16}$$

Here $\text{N}(x)$ grows larger as the rank of x decreases.

3 Gaussian Width

To make rigorous the ideas of Remark 1, we need a useful means of measuring the size of a convex cone. Any convex cone $K \subset X$ corresponds via isomorphism to a *spherically convex*[5] subset P of the unit sphere S_2. The isomorphism and its inverse are defined by $P = K \cap S_2$ and $K = \mathbb{R}_+ P$ respectively. This isomorphism suggests measuring the size of K by measuring P. This is the strategy we will employ, following Chandrasekaran et al.[3].

Definition 1. Let $P \subset X$ be compact. Define the *Gaussian Width* of P to be

$$w(P) := \mathbf{E} \sup_{z \in P} \langle g, z \rangle \tag{17}$$

where g is a standard Gaussian vector.

[5] A subset P of the unit sphere is spherically convex if whenever $x, y \in P$ and the angle between x and y is less than π, the great circle path from x to y also lies in P. See section 6.5 of Schneider and Weil[16].

We will measure the size of a cone K by computing the Gaussian width of $K \cap S_2$. Proposition 3 below alleviates some of the difficulty involved in calculating this width. Before we can prove Proposition 3, we will need an auxiliary result describing the support function of a closed convex cone. In what follows for any set C we denote by σ_C its support function.

Proposition 2 (Example V.2.3.1 Hiriart-Urruty and Lemaréchal[12])**.** Let $K \subset X$ be a closed convex cone. We have

$$\sigma_K(x) = \begin{cases} 0 \text{ if } x \in K^\circ \\ \infty \text{ else} \end{cases} =: I_{K^\circ} \tag{18}$$

the indicator function of the polar cone.

Proof. Given $x \in X$, apply the Morreau decomposition (Proposition A.5) to write

$$x = y + z \tag{19}$$

with $y \in K$, $z \in K^\circ$, $\langle y, z \rangle = 0$. Then

$$\sigma_K(x) = \sup_{s \in K} \langle y + z, s \rangle \geq \sup_{\alpha \geq 0} \langle y + z, \alpha y \rangle = \sup_{\alpha \geq 0} \alpha \|y\|^2. \tag{20}$$

So $\sigma_K(x) = \infty$ for $x \notin K^\circ$. For $x \in K^\circ$, $\langle x, s \rangle \leq 0$ for all $s \in K$ and $\sigma_K(x) = 0$ as desired. □

Proposition 3 (Proposition 3.6 in Chandrasekaran et al.[3]. Proof is significantly modified)**.** Let K be a nonempty closed convex cone. Let g denote a standard Gaussian vector. Then

$$w(K \cap S_2^{n-1}) \leq \mathbf{E}\, d(g, K^\circ) \tag{21}$$

where $d(g, K^\circ)$ is the Euclidean distance between g and K°.

Proof.

$$\begin{aligned}
& w(K \cap S_2) && && (22)\\
&= \mathbf{E} \sup_{z \in K \cap S_2} \langle g, z \rangle && && (23)\\
&\leq \mathbf{E} \sup_{z \in K \cap B_2} \langle g, z \rangle && \text{Since } S_2 \subset B_2 && (24)\\
&= \mathbf{E}\, \sigma_{K \cap B_2}(g) && \text{Def of } \sigma_{K \cap B_2} && (25)\\
&= \mathbf{E}\, \mathrm{cl}\, \mathrm{conv} \min(\sigma_K, \sigma_{B_2})(g) && \text{By Proposition A.9} && (26)\\
&= \mathbf{E}\, \mathrm{cl}\, \mathrm{conv} \min(I_{K^\circ}, \|\cdot\|)(g) && \text{By Proposition 2} && (27)\\
&= \inf\{(1-\alpha)\|x\| : x \in X,\, y \in K^\circ,\, 0 \leq \alpha \leq 1,\, (1-\alpha)x + \alpha y = g\} && \text{By Proposition A.3} && (28)\\
&= \inf\{\|x\| : x \in X,\, y \in K^\circ,\, x + y = g\} && && (29)\\
&= \mathbf{E}\, d(g, K^\circ). && && (30)
\end{aligned}$$

□

4 Concentration of Measure

In section 3 we established Gaussian width as our means for measuring the size of a tangent cone. We next need to connect the Gaussian width of a tangent cone to the likelihood that the hypotheses of Proposition 1 are satisfied for random Φ. To make this connection, we will require two theoretical tools. The first is a concentration of measure result, whose statement and proof come from Tao[17].

Proposition 4 (Concentration of measure for Lipshitz functions of Gaussian random variables)**.** [Theorem 2.1.12 in Tao[17]] Let $F : X \to \mathbb{R}$ be Lipshitz with Lipshitz constant L. Let g be a standard Gaussian vector on X. Then for $\lambda > 0$

$$P(F(g) - \mathbf{E}\, F(g) \geq \lambda) \leq \exp(-C\frac{\lambda^2}{L^2}) \qquad P(F(g) - \mathbf{E}\, F(g) \leq -\lambda) \leq \exp(-C\frac{\lambda^2}{L^2}) \tag{31}$$

for a fixed constant $C > 0$.

Proof. We shall prove this only for continuously differentiable functions with $\|\nabla F\| \leq L$. The proof can be extended to Lipshitz functions by a limiting argument that we shall not include here[6]. We shall only work on the probability $P(F(g) - \mathbf{E}\, F(g) \geq \lambda)$, noting that $P(F(g) - \mathbf{E}\, F(g) \leq -\lambda)$ can be dealt with by considering $-F$. Our approach will be to study the behavior of the moment generating function defined by $\mathbf{E} \exp(tF(g))$ for $t > 0$, and then to relate this behavior to the probability of interest.

We begin by replacing F with $F - \mathbf{E}\, F(g)$ so that the new F satisfies $\mathbf{E}\, F(g) = 0$. It is clearly sufficient to prove the result for this new F. Proposition C.2 (Jensen's Inequality), and the convexity of the exponential imply that for any $t \in \mathbb{R}$,

$$1 = \exp(t\, \mathbf{E}\, F(g)) \leq \mathbf{E} \exp(tF(g)). \tag{32}$$

Let h be an independent identically distributed copy of g. Then

$$\begin{aligned} \mathbf{E} \exp(tF(g)) &\leq \mathbf{E} \exp(tF(g))\, \mathbf{E} \exp(-tF(h)) && \text{by (32)} && (33)\\ &= \mathbf{E} \exp(t(F(g) - F(h))) && \text{by independence} && (34) \end{aligned}$$

From now on we will assume $t > 0$. We next study the quantity $F(g) - F(h)$ by applying the fundamental theorem of calculus along a circular arc:

$$F(g) - F(h) = \int_0^{\frac{\pi}{2}} \frac{d}{d\theta} F(h\cos\theta + g\sin\theta)\, \mathrm{d}\theta. \tag{35}$$

This operation will prove useful later because $g_\theta := h\cos\theta + g\sin\theta$ and its θ-derivative $g'_\theta := -h\sin\theta + g\cos\theta$ are independent identically distributed Gaussian vectors. This follows from

[6] Though I have not worked it out in detail, I believe the limiting argument refereed to above can be constructed using tools from Chapter 8 of Folland[7]. The key idea would be to define a family of continuously differentiable nonegative functions $\{\phi_t\}_{t>0}$ such that $\int \phi_t = 1$, and each ϕ_t is supported on $B_2(0,t)$. This would ensure that the continuously differential convolutions $\phi_t * F$ converge usefully to F as $t \downarrow 0$.

Proposition C.6 and the rotational invariance of the distribution of a standard Gaussian vector. Before making use of this independence, however, we compute

$$\exp(t(F(g) - F(h))) \tag{36}$$

$$= \exp\left(t \int_0^{\frac{\pi}{2}} \langle \nabla F(g_\theta), g'_\theta \rangle \, \mathrm{d}\theta\right) \tag{37}$$

$$= \exp\left(t \int_0^1 \frac{\pi}{2} \langle \nabla F(g_{\frac{\pi}{2}\beta}), g'_{\frac{\pi}{2}\beta} \rangle \, \mathrm{d}\beta\right) \qquad \text{Change of variables } \theta = \frac{\pi}{2}\beta \tag{38}$$

$$\leq \int_0^1 \exp\left(t\frac{\pi}{2} \langle \nabla F(g_{\frac{\pi}{2}\beta}), g'_{\frac{\pi}{2}\beta} \rangle\right) \, \mathrm{d}\beta. \qquad \text{Apply Proposition C.2 (Jensen's inequality)} \tag{39}$$

Next we take the expectation of both sides and use Fubini-Toneli to swap the order of integration (which works since the boundedness of ∇F makes the integral finite). We have

$$\mathbf{E}\left(\exp(t(F(g) - F(h)))\right) \tag{40}$$

$$\leq \int_0^1 \mathbf{E} \exp\left(t\frac{\pi}{2} \langle \nabla F(g_{\frac{\pi}{2}\beta}), g'_{\frac{\pi}{2}\beta} \rangle\right) \, \mathrm{d}\beta. \tag{41}$$

We proceed by using Proposition C.3. For fixed β, we define the function $Q_\beta : X \to \mathbb{R}$ by

$$Q_\beta(x) := \mathbf{E} \exp\left(t\frac{\pi}{2} \langle \nabla F(x), g'_{\frac{\pi}{2}\beta} \rangle\right). \tag{42}$$

For fixed x, if $\nabla F(x) \neq 0$ then by Proposition C.5 $\frac{\pi}{2} \langle \nabla F(x), g'_{\frac{\pi}{2}\beta} \rangle$ is a normal random variable with 0 mean and variance $\frac{\pi^2}{4} \|\nabla F(x)\|^2$. If $\nabla F(x) = 0$, then $\frac{\pi}{2} \langle \nabla F(x), X'_{\frac{\pi}{2}\beta} \rangle$ is a degenerate random variable concentrated at 0. Using Proposition C.7 in the former case and simple algebra in the later, we have

$$Q_\beta(x) \tag{43}$$

$$= \exp(t^2 \frac{\pi^2}{8} \|\nabla F(x)\|^2) \tag{44}$$

$$\leq \exp(t^2 \frac{\pi^2}{8} L^2). \tag{45}$$

Now by Proposition C.3 and by the independence of $g_{\frac{\pi}{2}\beta}$ $g'_{\frac{\pi}{2}\beta}$ we have

$$\mathbf{E} \exp\left(t\frac{\pi}{2} \langle \nabla F(g_{\frac{\pi}{2}\beta}), g'_{\frac{\pi}{2}\beta} \rangle\right) \tag{46}$$

$$= \mathbf{E}\, Q_\beta(g_{\frac{\pi}{2}\beta}) \tag{47}$$

$$\leq \exp(t^2 \frac{\pi^2}{8} L^2). \tag{48}$$

Combining the previous steps, we have proven the bound

$$\mathbf{E} \exp(tF(g)) \leq \exp(t^2 \frac{\pi^2}{8} L^2). \tag{49}$$

It remains only to convert this into a bound on the tail probability of $F(X)$. Assuming $t > 0$ we have

$$
\begin{aligned}
& P(F(X) \geq \lambda) && && (50)\\
& = P\left(\exp(tF(X)) \geq \exp(t\lambda)\right) && \text{by monotonicity of } \exp(t\cdot) && (51)\\
& \leq \exp(-t\lambda)\,\mathbf{E}\exp(tF(X)) && \text{by Prop C.1 (Markov)} && (52)\\
& \leq \exp(t^2\frac{\pi^2}{8}L^2 - t\lambda) && \text{by (49)} && (53)\\
& = \exp(-2\frac{\lambda^2}{\pi^2 L^2}) && \text{by minimizing with respect to } t && (54)
\end{aligned}
$$

Thus we have proven (31) with $C = \frac{2}{\pi^2}$. □

5 Gordon's Comparison Theorem

Having proved our concentration of measure result, we move on to Gordon's theorem, the second theoretical tool required to connect Gaussian width with the hypotheses of Proposition 1. Gordon's theorem bounds the expected minimum of of the function $\|\Phi(\cdot)\|$ on a subset of the unit sphere when Φ is randomized. This is useful because of the importance of $\|\Phi(\cdot)\|$ to Proposition 1.

In what follows, we will let $\lambda_n := \sqrt{2}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$, the expected Euclidean norm of a standard Gaussian vector in n-dimensional Euclidean space (see Proposition C.10).

Proposition 5. [Corollary 1.2 in Gordon[8]] Let Ω be a closed subset of the unit sphere in the n-dimensional Euclidean space X. Let Φ be a random linear transformation from X to $\mathbb{R}^m$ such that its matrix with respect to an orthonormal basis has independent, identically distributed Gaussian entries with zero mean and variance one. Then

$$\mathbf{E}\min_{z\in\Omega}\|\Phi z\| \geq \lambda_m - w(\Omega). \tag{55}$$

Proof. This result follows from Theorem 1.4 in Gordon[10], but that result has an extremely long and involved proof. We omit it. □

5.1 The Value Function

As mentioned above, the quantity $\|\Phi x\|$ links Gordon's theorem (Proposition 5) to Proposition 1. Here we prove a simple result to further describe this quantity. As usual let X be a Euclidean space. Let $\mathcal{M}(X)$ be the linear transformations on X. Let Ω be a closed subset of S_2, the unit sphere in X. Define the function $v : \mathcal{M}(X) \to \mathbb{R}$ by

$$v(\Phi) = \min_{z\in\Omega}\|\Phi z\|. \tag{56}$$

Proposition 6. v is 1-Lipshitz in the operator norm(and thus in the Frobenius norm).

Proof. Let $\Phi_1, \Phi_2 \in \mathcal{M}(X)$. Since Ω is compact there exists $z_1, z_2 \in \Omega$ such that

$$\|\Phi_1 z_1\| = v(\Phi_1) \tag{57}$$
$$\|\Phi_2 z_2\| = v(\Phi_2) \tag{58}$$

So

$$\begin{aligned}
& v(\Phi_1) - v(\Phi_2) && && (59)\\
& \leq \|\Phi_1 z_2\| - v(\Phi_2) && \text{definition of } v(\Phi_1) && (60)\\
& = \|\Phi_1 z_2\| - \|\Phi_2 z_2\| && (58) && (61)\\
& \leq \|(\Phi_1 - \Phi_2) z_2\| && \text{triangle inequality} && (62)\\
& \leq \|\Phi_1 - \Phi_2\| \|z_2\| && \text{def of operator norm} && (63)\\
& = \|\Phi_1 - \Phi_2\| && \Omega \subset S_2. && (64)
\end{aligned}$$

So $v(\Phi_1) - v(\Phi_2) \leq \|\Phi_1 - \Phi_2\|$. Swapping the role of Φ_1 and Φ_2, and combining the result with the above, we have

$$|v(\Phi_1) - v(\Phi_2)| \leq \|\Phi_1 - \Phi_2\|. \tag{65}$$

□

6 Main General Recovery Result

Armed with the concentration of measure result (Proposition 4) and Gordon's Theorem (Proposition 5), we are ready to prove Proposition 7, which relates the probability that a given x will solve Problem 2 or Problem 3 to the size of the corresponding tangent cone.

Proposition 7 (Corollary 3.3 in Chandrasekaran et al.[3])**.** Let X be an n-dimensional Euclidean space. Let Φ be a random linear transformation from X to $\mathbb{R}^m$ whose matrix with respect to an orthonormal basis has independent identically distributed Gaussian entries with variance $\frac{1}{m}$. Let γ be as in Section 2, and assume γ is finite. Let $x^* \in X$. Let $T(x^*)$ be as in Section 2. Let $\Omega = T(x^*) \cap S_2$.

1. Let $b = \Phi x^*$. Assume

$$m \geq w(\Omega)^2 + 1. \tag{66}$$

   Then x^* will be the unique solution to Problem 2 with probability exceeding

$$1 - \exp\left(-C\left(\lambda_m - w(\Omega)\right)^2\right) \tag{67}$$

   where $C > 0$ is as in Proposition 4.

2. Let $b = \Phi x^* + \nu$ where $\|\nu\| \leq \epsilon$ for some $\epsilon > 0$. Let $1 > \delta > 0$. Assume

$$m \geq \frac{w(\Omega)^2 + \frac{3}{2}}{(1-\delta)^2}. \tag{68}$$

Let $\hat{x}$ solve Problem 3. Then $\|x^* - \hat{x}\| \leq \frac{2\epsilon}{\delta}$ with probability exceeding

$$1 - \exp\left(-C\left(\lambda_m - w(\Omega) - \sqrt{m}\delta\right)^2\right) \tag{69}$$

where $C > 0$ is as in Proposition 4.

Proof. Our strategy will be to show that the hypotheses of Proposition 1 hold with high probability.

We first note that in the $\epsilon = 0$ case, the assumption (66) implies $\lambda_m > w(\Omega)$. This follows since

$$\begin{aligned}
&\lambda_m && && (70)\\
&\geq \frac{m}{\sqrt{m+1}} && \text{From Proposition C.10} && (71)\\
&= \frac{\sqrt{m}}{\sqrt{1+\frac{1}{m}}} && && (72)\\
&\geq \frac{\sqrt{w(\Omega)^2+1}}{1+\frac{1}{m}} && \text{By (66)} && (73)\\
&> \frac{\sqrt{w(\Omega)^2+\frac{w(\Omega)^2}{m}}}{1+\frac{1}{m}} && \text{Since by (66) } m > w(\Omega)^2 && (74)\\
&= w(\Omega). && && (75)
\end{aligned}$$

Analogously in the $\epsilon > 0$ case the assumption (68) implies $\lambda_m - \sqrt{m}\delta > w(\Omega)$. This follows

since

$$\lambda_m - \sqrt{m}\delta \tag{76}$$

$$\geq \frac{m}{\sqrt{m+1}} - \sqrt{m}\delta \qquad \text{From Proposition C.10} \tag{77}$$

$$= \frac{m - \sqrt{m}\sqrt{m+1}\delta}{\sqrt{m+1}} \tag{78}$$

$$\geq \frac{m - (m+1)\delta}{\sqrt{m+1}} \tag{79}$$

$$= \frac{(1-\delta)m - \delta}{\sqrt{m+1}} \tag{80}$$

$$= \frac{(1-\delta)\sqrt{m} - \frac{\delta}{\sqrt{m}}}{\sqrt{1+\frac{1}{m}}} \tag{81}$$

$$= \sqrt{\frac{(1-\delta)^2 m - 2(1-\delta)\delta + \frac{\delta^2}{m}}{1+\frac{1}{m}}} \tag{82}$$

$$\geq \sqrt{\frac{(1-\delta)^2 m - \frac{1}{2}}{1+\frac{1}{m}}} \qquad \text{dropping } \frac{\delta^2}{m} \text{ and using } \delta(1-\delta) \leq \frac{1}{4} \text{ because } 0 < \delta < 1 \tag{83}$$

$$\geq \sqrt{\frac{w(\Omega)^2 + 1}{1+\frac{1}{m}}} \qquad \text{by (68)} \tag{84}$$

$$> \sqrt{\frac{w(\Omega)^2 + \frac{w(\Omega)^2}{m}}{1+\frac{1}{m}}} \qquad \text{since } m > w(\Omega)^2 \tag{85}$$

$$= w(\Omega). \tag{86}$$

Defining v by $v(\Psi) = \min_{z \in \Omega} \|\Psi z\|$, we know by Proposition 6 that v is 1-Lipshitz. Furthermore, the transformation $\sqrt{m}\Phi$ is standard Gaussian. Thus by the concentration of measure result Proposition 4, we have for $t \geq 0$

$$P(v(\sqrt{m}\Phi) \geq \mathbf{E}\, v(\sqrt{m}\Phi) - t) \geq 1 - \exp(-Ct^2). \tag{87}$$

Applying Gordon's Theorem and dividing out $\sqrt{m}$, we have

$$P(v(\Phi) \geq \frac{\lambda_m - w(\Omega) - t}{\sqrt{m}}) \geq 1 - \exp(-Ct^2). \tag{88}$$

Setting $\tau = \frac{\lambda_m - w(\Omega) - t}{\sqrt{m}}$ we have the following inequality

$$P(v(\Phi) \geq \tau) \geq 1 - \exp(-C(\lambda_m - w(\Omega) - \sqrt{m}\tau)^2), \tag{89}$$

which holds when $0 \leq t = \lambda_m - w(\Omega) - \sqrt{m}\tau$.

In the $\epsilon = 0$ case, we have by (75) that $\lambda_m - w(\Omega) > 0$. We therefore know that (89) holds for τ less than some $\overline{\tau} > 0$. Thus applying the continuity from below property of measures (see

page 7 of Grimmett and Stirzaker[9]), we have

$$P(v(\Phi) > 0) = P(\bigcup_{\tau \in (0,\overline{\tau}]} \{v(\Phi) \geq \tau\}) = \lim_{\tau \to 0} P(v(\Phi) \geq \tau) \geq 1 - \exp(-C(\lambda_m - w(\Omega))^2). \quad (90)$$

By Proposition 1 this is gives us the desired result.

In the $\epsilon > 0$ case, we have by (86) that $\lambda_m - \sqrt{m}\delta > w(\Omega)$. Substituting $\tau = \delta$ into (89) gives us the desired result by Proposition 1.[7] □

7 Gaussian Widths for Specific problems

We are almost ready to execute the plan set out in Remark 1. With Proposition 7, we can relate the probability a given x^* solves Problem 2 or Problem 3 to a Gaussian width. The next step is to bound this Gaussian width for examples of interest. Each example must be treated separately.

7.1 l_1 Minimization

Proposition 8 (Proposition 3.10 in Chandrasekaran et al.[3])**.** Let $x^* \in \mathbb{R}^n$ be an k-sparse vector and let $\gamma = \|\cdot\|_1$. Let $\mathrm{T}(x^*)$ be as in Section 2. Let S_2 denote the unit sphere in $\mathbb{R}^n$. Then

$$(w(\mathrm{T}(x^*) \cap S_2))^2 \leq 2k \log(\frac{n}{k}) + \left(1 + \frac{1}{\sqrt{\pi}}\right) k. \quad (91)$$

Proof. We shall bound $w(\mathrm{T}(x^*) \cap S)$ using Proposition 3.

First, let $\operatorname{spp} x^*$ denote the support of x^*, that is the indexes of its nonzero components. Then by Propositions A.13 and A.19 we have

$$\mathrm{N}(x^*) = \{s \in \mathbb{R}^n : s_i = t \operatorname{sign}(x_i^*) \, \forall i \in \operatorname{spp}(x^*),\ s_i \in [-t, t] \, \forall i \in (\operatorname{spp} x^*)^c\}. \quad (92)$$

For fixed $z \in X$, we therefore have

$$\begin{aligned}
&(w(\mathrm{T}(x^*) \cap S_2))^2 && (93)\\
&\leq \inf_{u \in \mathrm{N}(x^*)} \|u - z\|^2 && (94)\\
&= \inf_{t>0} \left(\sum_{i \in \operatorname{spp} x^*} (z_i - t \operatorname{sign} x_i^*)^2 + \sum_{i \in (\operatorname{spp} x^*)^c} \inf_{u_i \in [-t,t]} (z_i - u_i)^2 \right) && (95)\\
&= \inf_{t>0} \left(\sum_{i \in \operatorname{spp} x^*} (z_i - t \operatorname{sign} x_i^*)^2 + \sum_{i \in (\operatorname{spp} x^*)^c} \operatorname{shrink}(z_i, t)^2 \right) && (96)\\
& && (97)
\end{aligned}$$

where

$$\operatorname{shrink}(z, t) := (z + t)\mathbf{1}_{\mathbb{R}_-}(z + t) + (z - t)\mathbf{1}_{\mathbb{R}_+}(z - t), \quad (98)$$

[7] Note that, like in Section 2, the specific structure of γ as a gauge function was not used here.

and for any set C, $\mathbf{1}_C$ is the characteristic function of C. Picking a $t > 0$, substituting the random vector g for z, and taking expectations yields

$$\mathbf{E} \inf_{u \in N_{\gamma(x^*)}} \|u - g\|^2 \tag{99}$$

$$\leq \mathbf{E} \sum_{i \in \operatorname{spp} x^*} (g_i - t \operatorname{sign} x_i^*)^2 + \mathbf{E} \sum_{i \in (\operatorname{spp} x^*)^c} \operatorname{shrink}(g_i, t)^2 \tag{100}$$

$$= \mathbf{E} \sum_{i \in \operatorname{spp} x^*} (g_i^2 + t^2 - 2 g_i t \operatorname{sign} x_i) + \mathbf{E} \sum_{i \in (\operatorname{spp} x^*)^c} \operatorname{shrink}(g_i, t)^2 \tag{101}$$

$$= \underbrace{k(1+t^2)}_{\substack{\text{Since } g_i \text{ is } N(0,1) \\ \text{and } \operatorname{spp} x^* \text{ has } k \text{ elements}}} + \sum_{i \in (\operatorname{spp} x^*)^c} \mathbf{E} \operatorname{shrink}(g_i, t)^2 \tag{102}$$

To proceed, we must bound the second term. We make use of the symmetry of the $\operatorname{shrink}(\cdot, t)$ function.

$$E \operatorname{shrink}(g_i, t)^2 \tag{103}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \operatorname{shrink}(q, t)^2 \exp(\frac{-q^2}{2}) \, \mathrm{d}q \tag{104}$$

$$= \frac{2}{\sqrt{2\pi}} \int_{t}^{\infty} (q - t)^2 \exp(\frac{-q^2}{2}) \, \mathrm{d}q \qquad \text{by symmetry} \tag{105}$$

$$= \underbrace{\frac{2}{\sqrt{2\pi}} \int_{t}^{\infty} q^2 \exp(\frac{-q^2}{2}) \, \mathrm{d}q}_{\text{call this } A} - \underbrace{\frac{4t}{\sqrt{2\pi}} \int_{t}^{\infty} q \exp(\frac{-q^2}{2}) \, \mathrm{d}q}_{\text{call this } B} + \frac{2t^2}{\sqrt{2\pi}} \int_{t}^{\infty} \exp(\frac{-q^2}{2}) \, \mathrm{d}q \tag{106}$$

Working on via integration by parts on term A,

$$A = \frac{2}{\sqrt{2\pi}} \int_{t}^{\infty} (-q)(-q \exp(\frac{-q^2}{2})) \, \mathrm{d}q \tag{107}$$

$$= \frac{2}{\sqrt{2\pi}} \int_{t}^{\infty} (-q)(\frac{d}{dq} \exp(\frac{-q^2}{2})) \, \mathrm{d}q \tag{108}$$

$$= \frac{2}{\sqrt{2\pi}} \left(-q \exp(-\frac{q^2}{2}) \Big|_{t}^{\infty} + \int_{t}^{\infty} \exp(\frac{-q^2}{2}) \, \mathrm{d}q \right) \tag{109}$$

$$= \frac{2t}{\sqrt{2\pi}} \exp(\frac{-t^2}{2}) + \frac{2}{\sqrt{2\pi}} \int_{t}^{\infty} \exp(\frac{-q^2}{2}) \, \mathrm{d}q. \tag{110}$$

In the case of term B, we have

$$B = \frac{4t}{\sqrt{2\pi}} \int_{t}^{\infty} \frac{d}{dq} (-\exp(-\frac{t^2}{2})) \, \mathrm{d}q \tag{111}$$

$$= \frac{4t}{\sqrt{2\pi}} \exp(\frac{-t^2}{2}). \tag{112}$$

Combining the above yields

$$\int_{-\infty}^{\infty} \operatorname{shrink}(q,t)^2 \,\mathrm{d}q \tag{113}$$

$$= \frac{2(1+t^2)}{\sqrt{2\pi}} \int_t^{\infty} \exp(\frac{-q^2}{2}) \,\mathrm{d}q - \frac{2t}{\sqrt{2\pi}} \exp(\frac{-t^2}{2}) \tag{114}$$

$$\leq \frac{2(1+t^2)}{\sqrt{2\pi}} \frac{1}{t} \exp(\frac{-t^2}{2}) - \frac{2t}{\sqrt{2\pi}} \exp(\frac{-t^2}{2}) \qquad \text{by Proposition C.8} \tag{115}$$

$$= \frac{2}{\sqrt{2\pi}} \frac{1}{t} \exp(\frac{-t^2}{2}). \tag{116}$$

Using this bound in (102) yields

$$\mathbf{E} \inf_{u \in \mathrm{N}(x^*)} \|u - g\|^2 \tag{117}$$

$$\leq k(1+t^2) + (n-k)\frac{2}{\sqrt{2\pi}} \frac{1}{t} \exp(\frac{-t^2}{2}). \tag{118}$$

for any $t > 0$. Substitute $t = \sqrt{2\log\left(\frac{n}{k}\right)}$ to get

$$k(1 + 2\log\left(\frac{n}{k}\right)) + (n-k)\frac{2}{\sqrt{2\pi}} \frac{1}{\sqrt{2\log\left(\frac{n}{k}\right)}} \frac{k}{n} \tag{119}$$

$$= k(1 + 2\log\left(\frac{n}{k}\right)) + \frac{k\left(1-\frac{k}{n}\right)}{\sqrt{\pi\log\left(\frac{n}{k}\right)}}. \tag{120}$$

By Taylor series expansion with remainder, for $x > 1$ we have

$$\log x > 0 + \frac{1}{x}(x-1) = 1 - \frac{1}{x}. \tag{121}$$

Thus $\frac{1-\frac{1}{x}}{\sqrt{\log x}} \leq \sqrt{1-\frac{1}{x}} \leq 1$. Thus the bound in (120) becomes[8]

$$\left(w(\mathrm{T}(x^*) \cap S_2)\right)^2 \leq 2k\log\left(\frac{n}{k}\right) + (1 + \frac{1}{\sqrt{\pi}})k \tag{122}$$

□

With this result, we have carried out the plan set out in Remark 1 for the case of sparse vector recovery. Using Proposition 8, we can bound the Gaussian width of a tangent cone $\mathrm{T}(x^*)$, and using Proposition 7, we can use this width to determine when we can recover x^* with high probability. In the $\epsilon = 0$ case, for example, we have.

[8] Actually $\frac{1-\frac{1}{x}}{\sqrt{\log x}} < 0.65$, so a somewhat better bound can be achieved with more work.

Proposition 9. Let $x^* \in \mathbb{R}^n$ be k-sparse. Let Φ be a random $m \times n$ matrix whose entries have zero mean and variance $\frac{1}{m}$. Let $b = \Phi x^*$. Assume

$$m \geq 4\left(2k \log\left(\frac{n}{k}\right) + (1 + \frac{1}{\sqrt{\pi}})k + 1\right). \tag{123}$$

Then x^* is the unique solution to

$$\text{minimize } \|x\|_1 \text{ subject to } \Phi x = b \tag{124}$$

with probability exceeding

$$1 - \exp\left(-C\frac{1}{4}\frac{m^2}{m+1}\right) \tag{125}$$

where C is as in Proposition 4.

Proof. Proposition 8 and (123) show

$$4((w(\mathrm{T}(x^*) \cap S_2))^2 + 1) \leq m. \tag{126}$$

Repeating the steps performed in (70)-(75), we get

$$\lambda_m \geq \frac{\sqrt{4(w(\Omega)+1)}}{1 + \frac{1}{m}} > 2w(\Omega). \tag{127}$$

The hypothesis of the first part of Proposition 7 are satisfied. So solving (170) will find x^* with probability exceeding

$$1 - \exp\left(-C\left(\lambda_m - w(\Omega)\right)^2\right) \tag{128}$$

$$\geq 1 - \exp\left(-C(\frac{1}{2}\lambda_m)^2\right) \qquad \text{by (127)} \tag{129}$$

$$\geq 1 - \exp\left(-C\frac{1}{4}\frac{m^2}{m+1}\right) \qquad \text{by Proposition C.10} \tag{130}$$

□

Similarly, In the $\epsilon > 0$ case, we have

Proposition 10. Let $x^* \in \mathbb{R}^n$ be k-sparse. Let Φ be a random $m \times n$ matrix whose entries have zero mean and variance $\frac{1}{m}$. Let $b = \Phi x^* + \nu$ where $\|\nu\| \leq \epsilon$. Assume

$$m \geq 16\left(2k \log\left(\frac{n}{k}\right) + (1 + \frac{1}{\sqrt{\pi}})k + 1\right) + 2. \tag{131}$$

Let $\hat{x}$ be any solution of

$$\text{minimize } \|x\|_1 \text{ subject to } \|\Phi x - b\| \leq \epsilon. \tag{132}$$

Then $\|x^* - \hat{x}\| \leq 4\epsilon$ with probability exceeding

$$1 - \exp\left(-\frac{C}{16}\frac{(m-1)^2}{m+1}\right) \tag{133}$$

Proof. Set $\delta = \frac{1}{2}$. Proposition 8 and (131) show

$$(1-\delta)^2 m - \frac{1}{2} \geq 4(w(\Omega)+1). \tag{134}$$

Repeating the steps performed in (76)-(86) we get

$$\lambda_m - \sqrt{m}\delta \geq 2w(\Omega). \tag{135}$$

The hypothesis of the second part of Proposition 7 are satisfied. So (173) will find a solution $\hat{x}$ such that $\|\hat{x} - x^*\| \leq \frac{2\epsilon}{\delta} = 4\epsilon$ with probability exceeding

$$1 - \exp\left(-C\left(\lambda_m - w(\Omega) - \sqrt{m}\delta\right)^2\right) \tag{136}$$

$$\geq 1 - \exp\left(-C\left(\frac{1}{2}(\lambda_m - \sqrt{m}\delta)\right)^2\right) \qquad \text{Using (135)} \tag{137}$$

$$\geq 1 - \exp\left(-\frac{C}{4}\left(\frac{m}{\sqrt{m+1}} - \sqrt{m}\delta\right)^2\right) \qquad \text{using Proposition C.10} \tag{138}$$

$$\geq 1 - \exp\left(-\frac{C}{4}\left(\frac{(1-\delta)m - \delta}{\sqrt{m+1}}\right)^2\right) \tag{139}$$

$$= 1 - \exp\left(-\frac{C}{16}\frac{(m-1)^2}{m+1}\right) \qquad \text{Using } \delta = \frac{1}{2} \tag{140}$$

as desired. □

7.2 Nuclear Norm Minimization

Having bounded the Gaussian width for l_1 minimization, we move to our second main example and bound the Gaussian width for nuclear norm minimization.

Proposition 11 (Proposition 3.11 in Chandrasekaran et al.[3])**.** Let $x^* \in \mathbb{R}^{n_1 \times n_2}$ be a rank r matrix and let $\gamma = \|\cdot\|_*$. Let $\mathrm{T}(x^*)$ be as in Section 2. Let S_2 denote the unit sphere in $\mathbb{R}^{n_1 \times n_2}$. Then

$$\left(w(\mathrm{T}(x^*) \cap S_2)\right)^2 \leq 3r(n_1 + n_2 - r) + 2r\sqrt{\pi}\sqrt{n_1 + n_2 - 2r} \tag{141}$$

Proof. Let x^* have singular value decomposition

$$x^* = UDV^T. \tag{142}$$

Where $U \in \mathbb{R}^{n_1 \times r}$, $D \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{n_2 \times r}$.

By Proposition B.1 and A.19 we have

$$\mathrm{N}\|x^*\|_* = \{tUV^T + W,\, W \in \mathbb{R}^{n_1 \times n_2},\, \|W\| \leq t,\, WV = 0,\, U^T W = 0,\, t \geq 0\}. \tag{143}$$

Let $u_1, \ldots, u_r$ denote the columns of U and let $v_1, \ldots, v_r$ denote the columns of V. We define the subspace

$$\Delta = \operatorname{span}\{u_i v_j^T,\, 1 \le i \le r,\, 1 \le j \le r\} \oplus \operatorname{span}\{a v_j^T,\, 1 \le j \le r,\, a \in \{u_1, \ldots, u_r\}^\perp\} \tag{144}$$

$$\oplus \operatorname{span}\{u_i b^T,\, 1 \le i \le r,\, b \in \{v_1, \ldots, v_r\}^\perp\} \tag{145}$$

Since Δ is the direct sum of three subspaces that are mutually perpendicular in the Frobenius inner-product, its dimension is the sum of their dimensions:

$$\dim \Delta = r^2 + (n_1 - r)r + r(n_2 - r) \tag{146}$$

$$= r(n_1 + n_2 - r). \tag{147}$$

Let g be a Gaussian matrix in $\mathbb{R}^{n_1 \times n_2}$. We will upper bound the distance from g to the normal cone $\mathrm{N}\|x^*\|_*$ by computing the distance from g to

$$Z(g) := \left(\|\Pi_{\Delta^\perp}(g)\| UV^T + \Pi_{\Delta^\perp}(g)\right) \in \mathrm{N}\|x^*\|_* \tag{148}$$

where for any closed convex set C, we let Π_C be the projection operator onto C. Let $\|\cdot\|_F$ denote the Frobenius norm. We have

$$\mathbf{E}\|g - Z(g)\|_F^2 \tag{149}$$

$$= \mathbf{E}\|\Pi_\Delta(g) + \Pi_{\Delta^\perp}(g) - \|\Pi_{\Delta^\perp}(g)\| UV^T - \Pi_{\Delta^\perp}(g)\|_F^2 \tag{150}$$

$$= \mathbf{E}\|\Pi_\Delta(g) - \|\Pi_{\Delta^\perp}(g)\| UV^T\|_F^2 \tag{151}$$

$$= \mathbf{E}\|\Pi_\Delta(g)\|_F^2 - 2\|\Pi_{\Delta^\perp}(g)\|\langle \Pi_\Delta(g), UV^T\rangle + \|\Pi_{\Delta^\perp}(g)\|^2 \|UV^T\|_F^2 \tag{152}$$

$$= \mathbf{E}\|\Pi_\Delta(g)\|_F^2 + \|\Pi_{\Delta^\perp}(g)\|^2 \|UV^T\|_F^2 \quad \text{Since } \Pi_\Delta(g) \text{ and } \Pi_{\Delta^\perp}(g) \text{ are independent} \tag{153}$$

By Proposition C.10, we know that

$$\mathbf{E}\|\Pi_\Delta(g)\|_F^2 = \dim \Delta = r(n_1 + n_2 - r). \tag{154}$$

Also, because the singular values of UV^T consist of r copies of 1, we have

$$\|UV^T\|_F^2 = r. \tag{155}$$

It remains to bound $\mathbf{E}\|\Pi_{\Delta^\perp}(g)\|$. Applying Proposition C.11 (Theorem II.13 in Davidson and Szarek[5]) we have that for $s \ge 0$

$$P(\|\Pi_{\Delta^\perp}(g)\| \ge \sqrt{n_1 - r} + \sqrt{n_2 - r} + s) \le \exp(-\frac{s^2}{2}). \tag{156}$$

Define $\mu := \sqrt{n_1 - r} + \sqrt{n_2 - r}$. Now compute

$$\begin{aligned}
&\mathbf{E}\|\Pi_{\Delta^\perp}(g)\|^2 && (157)\\
&= \int_0^\infty P(\|\Pi_{\Delta^\perp}(g)\|^2 > h)\,\mathrm{d}h && (158)\\
&\leq \mu^2 + \int_{\mu^2}^\infty P(\|\Pi_{\Delta^\perp}(g)\|^2 > h)\,\mathrm{d}h && (159)\\
&= \mu^2 + \int_0^\infty 2(t+\mu)P(\|\Pi_{\Delta^\perp}(g)\|^2 > (t+\mu)^2)\,\mathrm{d}t \qquad \text{letting } h = (t+\mu)^2 && (160)\\
&= \mu^2 + \int_0^\infty 2(t+\mu)P(\|\Pi_{\Delta^\perp}(g)\| > (t+\mu))\,\mathrm{d}t && (161)\\
&\leq \mu^2 + \int_0^\infty 2(t+\mu)\exp(\frac{-t^2}{2})\,\mathrm{d}t \qquad \text{By (156)} && (162)\\
&= \mu^2 + 2 + 2\mu\sqrt{\frac{\pi}{2}}. && (163)
\end{aligned}$$

Combining (153), (154), (155), and (163) we get

$$\begin{aligned}
&\mathbf{E}\, d(g, \mathrm{N}\|x^*\|_*) && (164)\\
&\leq r(n_1+n_2-r) + r(\sqrt{n_1-r}+\sqrt{n_2-r})^2 + 2r + 2r(\sqrt{n_1-r}+\sqrt{n_2-r})\sqrt{\frac{\pi}{2}} && (165)\\
&\leq r(n_1+n_2-r) + 2r(n_1+n_2-2r) + 2r + 2r\sqrt{2(n_1+n_2-2r)}\sqrt{\frac{\pi}{2}} && (166)\\
&\leq 3r(n_1+n_2-r) + 2r\sqrt{\pi}\sqrt{n_1+n_2-2r} && (167)\\
& && (168)
\end{aligned}$$

where we have used that for real a, b we have $(a+b)^2 \leq 2a^2 + 2b^2$ which follows from $(a-b)^2 \geq 0$.

An application of Proposition 3 gives the desired result[9]. □

We have the following analogue to Proposition 9

Proposition 12. Let $x^* \in \mathbb{R}^{n_1 \times n_2}$ be a rank r matrix. Let Φ be a random linear transformation from $\mathbb{R}^{n_1 \times n_2}$ to $\mathbb{R}^m$, which when represented as a matrix has entries with zero mean and variance $\frac{1}{m}$. Let $b = \Phi x^*$. Assume

$$m \geq 4\left(3r(n_1+n_2-r) + 2r\sqrt{\pi}\sqrt{n_1+n_2-2r} + 1\right). \tag{169}$$

Then x^* is the unique solution to

$$\text{minimize } \|x\|_* \text{ subject to } \Phi(x) = b \tag{170}$$

[9]There seems to be an error in the proof of Proposition 3.11 in Chandrasekaran et al.[3]: equation 85 in that paper does not appear to follow from equation 84. I modified their proof to correct this apparent mistake, and so got a slightly worse bound on the Gaussian width.

with probability exceeding

$$1 - \exp\left(-C\frac{1}{4}\frac{m^2}{m+1}\right) \tag{171}$$

where C is as in Proposition 4.

Proof. Essentially identical to that of Proposition 9. □

Similarly, In the $\epsilon > 0$ case, we have

Proposition 13. Let $x^* \in \mathbb{R}^{n_1 \times n_2}$ be a rank r matrix. Let Φ be a random linear transformation from $\mathbb{R}^{n_1 \times n_2}$ to $\mathbb{R}^m$, which when represented as a matrix has entries with zero mean and variance $\frac{1}{m}$. Let $b = \Phi x^* + \nu$ where $\|\nu\| \leq \epsilon$. Assume

$$m \geq 16\left(3r(n_1 + n_2 - r) + 2r\sqrt{\pi}\sqrt{n_1 + n_2 - 2r} + 1\right) + 2. \tag{172}$$

Let $\hat{x}$ be any solution of

$$\text{minimize } \|x\|_* \text{ subject to } \|\Phi(x) - b\| \leq \epsilon. \tag{173}$$

Then $\|x^* - \hat{x}\| \leq 4\epsilon$ with probability exceeding

$$1 - \exp\left(-\frac{C}{16}\frac{(m-1)^2}{m+1}\right) \tag{174}$$

where C is as in Proposition 4.

Proof. Basically identical to Proposition 10. □

8 Conclusion

Frequently, convexity serves as a nexus, fusing discrete or discontinuous problems with continuous ones. This role is apparent in some of the simplest convex objects. For example, a polyhedron has both a discrete vertex/edge/face structure a continuous interior structure. We have seen numerous other instances in CS798, including the solution of minimum cut problems by gradient descent, the use of a randomized projection algorithm to find a partly integral point in a large convex body, and the proof of an approximate caratheodory theorem via mirror descent.

The class of problems we presented in this essay provide yet another example. One solves these problems by replacing a highly nonsmooth set with a much larger, much smoother set, and then minimizing a nonsmooth function over the smoothed set. The structure of the original nonsmooth set is partly preserved in the nonsmooth function. Moreover, the nonsmooth function is tractable to minimize because it is convex.

The strategy of transferring nonsmoothness from a set to a function in such a way that the function remains tractable has been a fruitful one. I predict that cunning mathematicians will continue to find new ways to use it the future.

9 Extensions

9.1 Random Convex Programs

Amelunxen et al.[1] observe that the question considered in this essay– whether a given distinguished point is likely to be the a solution to a linear inverse problem with random data– is a special case of the problem of determining the probability that a randomly rotated convex cone intersects a fixed convex cone. Employing results from the theory of spherical integral geometry (covered in chapter 6.5 of Schneider and Weil[16]), they characterize this probability in terms of a quantity called the statistical dimension of a cone. The statistical dimension is related to but distinct from the Gaussian width we used in this essay. In addition to allowing them to treat a broader class of problems, Amelunxen et al.'s approach also enables them to produce more detailed results. For example, in the case of sparse signal recovery, they characterize not only the region in which l_1 minimization is likely to succeed, but also the region in which it is likely to fail, and the transition between the two regions.

9.2 Total Variation

If an image or signal is modeled as a real function of one or more real variables, its total variation is the L_1 norm of the pointwise l_2 norm of its gradient. For practical images, which are of course discrete, a discretized total variation is used. The great utility of the total variation regularizer in signal and image processing lends considerable importance to the task of understanding its theoretical properties. Cai and Xu[2] undertake this task, using strategies mirroring those of Chandrasekaran et al.[3] to relate the sparsity of an signal's gradient to the odds it can be recovered by total variation minimization.

A Appendix: Results from Convex Analysis

Here are some useful results from Convex Analysis. This material is mostly drawn from Hiriart-Urruty and Lemaréchal ([12], [13]). In what follows, let X be a Euclidean space.

A.1 Convex Functions

Proposition A.1 (The Criterion of Increasing Slopes)**.** A function $f : X \to \mathbb{R}$ is convex if and only if for all $t_2 > t_1 > 0$ and for all vectors $x, d \in X$

$$\frac{f(x + t_2 d) - f(x)}{t_2} \geq \frac{f(x + t_1 d) - f(x)}{t_1}, \tag{175}$$

which holds if an only if

$$\frac{f(x + t_2 d) - f(x + t_1 d)}{t_2 - t_1} \geq \frac{f(x + t_1 d) - f(x)}{t_1}. \tag{176}$$

Proof. (175) is equivalent to

$$t_1(f(x + t_2 d) - f(x)) \geq t_2(f(x + t_1 d) - f(x_1)) \tag{177}$$

$$\frac{t_1}{t_2} f(x + t_2 d) + (1 - \frac{t_1}{t_2}) f(x) \geq f(x + t_1 d) \tag{178}$$

but this is equivalent to the definition of convexity. To get (176), define $x_1 := 0$, $y_1 := t_1$, $z_1 := t_2$, $x_2 := f(x)$, $y_2 := f(x + t_1 d)$, and $z_2 := f(x + t_2 d)$. Then

$$\frac{z_2 - y_2}{z_1 - y_1} - \frac{y_2 - x_2}{y_1 - x_1} \tag{179}$$

$$= \frac{(y_1 - x_1)(z_2 - y_2) - (z_1 - y_1)(y_2 - x_2)}{(z_1 - y_1)(y_1 - x_1)} \tag{180}$$

$$= \frac{y_1(z_2 - y_2 + y_2 - x_2) - x_1(z_2 - y_2) - z_1(y_2 - x_2)}{(z_1 - y_1)(y_1 - x_1)} \tag{181}$$

$$= \frac{y_1(z_2 - x_2) - x_1(z_2 - y_2) - z_1(y_2 - x_2)}{(z_1 - y_1)(y_1 - x_1)} \tag{182}$$

$$= \frac{y_1(z_2 - x_2) - x_1(z_2 - x_2) - x_1(x_2 - y_2) - z_1(y_2 - x_2)}{(z_1 - y_1)(y_1 - x_1)} \tag{183}$$

$$= \frac{(y_1 - x_1)(z_2 - x_2) - (z_1 - x_1)(y_2 - x_2)}{(z_1 - y_1)(y_1 - x_1)} \tag{184}$$

$$= \frac{\frac{z_2 - x_2}{z_1 - x_1} - \frac{y_2 - x_2}{y_1 - x_1}}{\frac{z_1 - y_1}{z_1 - x_1}}. \tag{185}$$

The positivity of (179) is equivalent to (176), while the positivity of (185) is equivalent to (175) (since $t_2 > t_1 > 0$). But we have shown the two quantities in question are equal, so we are done. □

Proposition A.2 (Theorem B.3.1.2 in Hiriart-Urruty and Lemaréchal [13])**.** Let $f : X \to \mathbb{R}$ be convex. Let $S \subset X$ be convex and compact. Then there exists a positive real number $L(S)$ such that on S, f is $L(S)$-lipshitz

Proof. See Hiriart-Urruty and Lemaréchal[13]. □

Proposition A.3 (Proposition IV.2.5.1/IV.2.5.2/IV.2.5.4 in Hiriart-Urruty and Lemaréchal [12])**.** Let $g : X \to \mathbb{R}$ and suppose there exists an affine function $b : X \to \mathbb{R}$ such that $g \geq b$. Then the following functions (called the closed convex hull of g and denoted $\operatorname{cl}\operatorname{conv} g$) are convex and equal:

$$f_1(x) = \inf\{r : (x, r) \in \operatorname{cl}\operatorname{conv}\operatorname{epi} g\} \tag{186}$$

$$f_2(x) = \sup\{a(x) : \ a \text{ affine and } a(x) \leq g(x)\, \forall x \in X\}. \tag{187}$$

Furthermore, if $g = \min\{g_1, \ldots, g_m\}$, where the g_i are finite convex functions then

$$\operatorname{cl}\operatorname{conv} g(x) = \inf\{\sum_{j=1}^{m} \alpha_j g_j(x_j) : \alpha_i \geq 0\, \forall i, \sum_{i=1}^{m} \alpha_i = 1, \sum_{i=1}^{m} \alpha_j x_j = x\} \tag{188}$$

Proof. See Proposition IV.2.5.1/IV.2.5.2/IV.2.5.4 in Hiriart-Urruty and Lemaréchal[12]. □

A.2 Convex Cones

Given a convex set $C \subset X$ and a point $x \in C$, define the normal cone

$$\mathrm{N}_C(x) = \{s \in X : \langle y - x, s\rangle \leq 0 \, \forall y \in C\}. \tag{189}$$

and the tangent cone

$$\mathrm{T}_C(x) = \mathrm{cl}\, \mathbb{R}_+ \left(C - x\right). \tag{190}$$

Proposition A.4.

$$N_C(x) = (T_C(x))^\circ \tag{191}$$

Proof. First, let $s \in (T_C(x))^\circ$. Since $C - x \subset T_C(x)$ we have for all $y \in C$

$$\langle y - x, s\rangle \leq 0 \tag{192}$$

so that $s \in N_C(x)$. Conversely suppose $s \in N_C(x)$. Then by definition

$$\langle y - x, s\rangle \leq 0 \, \forall y \in C \tag{193}$$
$$\langle \alpha(y - x), s\rangle \leq 0 \, \forall y \in C \, \forall \alpha \geq 0. \tag{194}$$

Taking limits and using the continuity of the inner product, we have

$$\langle z, s\rangle \leq 0 \, \forall z \in T_C(x) \tag{195}$$

as desired. □

Proposition A.5 (Moreau Decomposition. III.3.2.5 in Hiriart-Urruty and Lemaréchal[12])**.** Let $K \subset X$ be a closed convex cone. Let $x \in X$. Then the following are equivalent.

- y is the projection of x onto K and z is the projection of x onto K°.
- $x = y + z$, $y \in K$, $z \in K^\circ$, and $\langle y, z\rangle = 0$.

Proof. Assume the first statement holds. Then by the characterization of the projection onto a convex set, we have for all $w \in K$

$$\langle x - y, w - y\rangle \leq 0. \tag{196}$$

Since $\alpha y \in K$ for $\alpha \geq 0$ we have

$$(\alpha - 1)\langle x - y, y\rangle \leq 0. \tag{197}$$

Since this must hold for all $\alpha > 0$, we must have

$$\langle x - y, y\rangle = 0. \tag{198}$$

Now use this to compute for $u \in K^\circ$

$$\langle x - (x - y), u - (x - y)\rangle = \langle y, u - (x - y)\rangle = \langle y, u\rangle \leq 0 \tag{199}$$

so that $x - y$ is the projection of x onto K° and thus $x - y = z$. Combined with (198) this gives the second statement.

Now assume the second statement holds. Then for $w \in K$

$$\langle x - y, w - y \rangle = \langle z, w - y \rangle = \langle z, w \rangle - \langle z, y \rangle = \langle z, w \rangle \leq 0 \tag{200}$$

so that y is the projection of x onto K. Symmetrically z is the projection of x onto K°. □

Proposition A.6 (Proposition A.1.4.7 in Hiriart-Urruty and Lemaréchal [13])**.** Let S be a nonempty compact set whose convex hull does not contain the origin. Then the conical hull of S (i.e. $\operatorname{cone} S$) is closed.

Proof. See Proposition A.1.4.7 in Hiriart-Urruty and Lemaréchal [13]. □

A.3 Sublinear Functions

Definition A.1. A function $\sigma : X \to \mathbb{R}$ is said to be *finite sublinear* if it is convex and satisfies for all $x \in X$ and $t > 0$

$$\sigma(tx) = t\sigma(x). \tag{201}$$

The condition (201) is called positive homogeneity.

Definition A.2. Given a convex compact set $C \subset \mathbb{R}^n$ define its support function to be $\sigma_C(x) := \sup\{\langle x, s \rangle : s \in C\}$.

Proposition A.7. The support function σ_C of a convex compact set C is finite sublinear.

Proof. Convexity follows since σ is the supremum of linear functions. The compactness of C makes σ_C finite. Positive homogeneity is a consequence of the interaction of the supremum operator with multiplication by a positive constant. □

Proposition A.8 (Theorem B.3.1.1 in Hiriart-Urruty and Lemaréchal [13])**.** The σ be a finite sublinear function. Then σ is the support function of the set

$$S = \{s \in X : \langle x, s \rangle \leq \sigma(x) \, \forall x \in X\}. \tag{202}$$

Proof. See Theorem B.3.1.1 in [13]. □

Proposition A.9 (Theorem V.3.3.3 in Hiriart-Urruty and Lemaréchal [13])**.** Let J be a finite index set. Let $\{\sigma_j\}_{j \in J}$ be the support functions of the compact convex sets $\{S_j\}_{j \in J}$. Let $S = \bigcap_{j \in J} S_j \neq \emptyset$. Let σ_S be the support function of S. Then

$$\sigma_S = \operatorname{cl} \operatorname{conv} \inf_{j \in J} \sigma_j \tag{203}$$

Proof.

- $s \in S$ is equivalent to $s \in s_j \, \forall j \in J$.
- By Proposition A.8, this is equivalent to $\langle \cdot, s \rangle \leq \sigma_j \, \forall j \in J$.

- This is equivalent to $\langle \cdot, s \rangle \leq \inf_{j \in J} \sigma_j$.
- This is equivalent to $\langle \cdot, s \rangle \leq \operatorname{cl}\operatorname{conv} \inf_{j \in J} \sigma_j$ by Proposition A.3.
- Using Proposition A.8, this last statement implies that the finite sublinear function $\operatorname{cl}\operatorname{conv} \inf_{j \in J} \sigma_j$ is the support function of S (Proposition A.3 implies that $\operatorname{cl}\operatorname{conv} \inf_{j \in J} \sigma_j$ is finite convex. Positive homogeneity follows from the representation (188). So it is finite sublinear.).

□

A.4 Subgradients

Definition A.3. If $f : X \to \mathbb{R}$ is convex, define its subgradient at x X

$$\partial f(x) = \{s \in X : f(y) \geq f(x) + \langle y - x, s \rangle \, \forall y \in X\}. \tag{204}$$

Proposition A.10 (pg 168 in Hiriart-Urruty and Lemaréchal [13])**.** If $f : X \to \mathbb{R}$ is convex then $\partial f(x)$ is convex and compact for all $x \in X$.

Proof. The closeness and convexity of $\partial f(x)$ follows from the continuity and convexity of the inner product, and the preservation of convexity and closedness under intersection of sets. To see the boundness of $\partial f(x)$, let L be the Lipshitz constant provided by Proposition A.2 for the set $B_2(x, 2)$, the Euclidean ball of radius 2 centered on x. For nonzero $s \in \partial f(x)$ let $y = x + \frac{s}{\|s\|}$. Then

$$f(x) + L \geq f(y) \geq f(x) + \langle y - x, s \rangle = f(x) + \|s\| \tag{205}$$

so that

$$L \geq \|s\|. \tag{206}$$

□

Definition A.4. If $f : X \to \mathbb{R}$ is convex and $x, d \in X$ define its directional derivative at x in direction D by

$$f'(x, d) = \inf_{t > 0} \frac{f(x + td) - f(x)}{t}. \tag{207}$$

Note that by Proposition A.1, $\frac{f(x+td)-f(x)}{t}$ is decreases monotonically as $t \downarrow 0$. Thus

$$f'(x, d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}. \tag{208}$$

Proposition A.11 (Proposition D.1.1.2 in Hiriart-Urruty and Lemaréchal [13])**.** Let $f : X \to \mathbb{R}$ be convex. For fixed $x \in X$ the function $f'(x, \cdot)$ is finite sublinear.

Proof. First we show convexity. Pick $d_1, d_2 \in X$ and $\alpha_1, \alpha_2 \geq 0$ such that $\alpha_1 + \alpha_2 = 1$. Let $t > 0$. We have

$$f(x + t(\alpha_1 d_1 + \alpha_2 t_2)) - f(x) = f(\alpha_1(x + td_1) + \alpha_2(x + td_2)) - f(x) \tag{209}$$

$$\leq \alpha_1(f(x + td_1) - f(x)) + \alpha_2(f(x + td_2) - f(x)) \quad \text{by convexity of } f \tag{210}$$

Divide both sides by t and let $t \downarrow 0$ to show convexity.

To show positive homogeneity let $s > 0$

$$f'(x, sd) = \lim_{t \downarrow 0} \frac{f(x + tsd) - f(x)}{t} = s \lim_{u \downarrow 0} \frac{f(x + ud) - f(x)}{u} = sf'(x, d) \tag{211}$$

where we use a change of variables. To show finiteness, note that by Proposition A.2 for small enough t we have that $|f(x + td) - f(x)| \leq Lt\|d\|$. □

Proposition A.12. Let $f : X \to \mathbb{R}$ be convex and let $x \in X$. The function $f'(x, \cdot)$ is the support function of the set $\partial f(x)$.

Proof. To say that $s \in \partial f(x)$ is to say that for all $t > 0$ all $d \in X$ we have

$$f(x + td) \geq f(x) + t\langle d, s\rangle \tag{212}$$

$$\frac{f(x + td) - f(x)}{t} \geq \langle d, s\rangle \tag{213}$$

Since Proposition A.1 tells us that $t \mapsto \frac{f(x+td)-f(x)}{t}$ is monotonic, (213) holding for all $t > 0$ is equivalent to

$$\lim_{\downarrow 0} \frac{f(x + td) - f(x)}{t} \geq \langle d, s\rangle \tag{214}$$

$$f'(x, d) \geq \langle d, s\rangle. \tag{215}$$

By Proposition A.8, this is what we want. □

Proposition A.13 (pg 180 of Hiriart-Urruty and Lemaréchal [13]. Proof is mine.)**.** If $C \subset X$ is a convex compact set, and σ_C is its support function then for all $x \in \mathbb{R}^n$

$$\partial\sigma_C(x) = \{s \in C : \langle x, s\rangle \geq \langle x, z\rangle \, \forall z \in C\} =: F_C(x) \tag{216}$$

Proof. We aim to show the statement $s \in F_C(x)$ is equivalent to the statement that for all y

$$\sup_{z_1 \in C} \langle y, z_1\rangle \geq \sup_{z_2 \in C} \langle x, z_2\rangle + \langle y - x, s\rangle. \tag{217}$$

First let $s \in \partial f(x)$. To generate a contradiction assume $s \notin F_C(x)$. Then there exists a $z \in C$ such that $\langle x, z\rangle > \langle x, s\rangle$. But then for all y

$$\sup_{z_1 \in C} \langle y, z_1\rangle \geq \sup_{z_2 \in C} \langle x, z_2\rangle + \langle y - x, s\rangle \geq \langle x, z\rangle + \langle y - x, s\rangle > \langle y, s\rangle. \tag{218}$$

Take $y = 0$ to get a contradiction. Conversely let $s \in F_C(x)$ Then $\langle x, z\rangle \leq \langle x, s\rangle$ for all $z \in C$. So

$$\sup_{z_2 \in C} \langle x, z_2\rangle + \langle y - x, s\rangle \leq \langle y, s\rangle \leq \sup_{z_1 \in C} \langle y, z_1\rangle. \tag{219}$$

as desired. □

Proposition A.14.

Proposition A.15 ([12] Theorem VI.4.4.2)**.** Let J be a compact subset of a metric space. Let $\{f_j : j \in J\}$ be a family of convex functions on $\mathbb{R}^n$. Let $f = \sup_{j\in J} f_j$ and assume f is finite everywhere. Assume that the maps $j \to f_j(x)$ are upper semicontinuous for each $x \in \mathbb{R}^n$. Let $J(x) := \{j \in J : f_j(x) = f(x)\}$ Then for all x

$$\partial f(x) = \operatorname{conv} \bigcup_{j \in J(x)} \partial f_j(x). \tag{220}$$

A.5 Descent

For any convex function f, and point x denote by $Sf(x) = \{y \in \mathbb{R}^n : f(y) \leq f(x)\}$, the sublevel set of f at x.

Proposition A.16 (VI.1.3.2 in Hiriart-Urrutty and Lemaréchal[12])**.** For any convex function $f : X \to \mathbb{R}$ we have that

$$\mathrm{T}_{Sf(x)}(x) \subset \{d : f'(x, d) \leq 0\} \tag{221}$$

Proof. We have

$$\mathrm{T}_{Sf(x)}(x) = \mathrm{cl}\{\alpha d : f(x + d) - f(x) \leq 0, \alpha \geq 0\}. \tag{222}$$

By Proposition A.1, we know that the slope of a convex function is monotone, so $f(x+d)-f(x) \leq 0$ implies $f'(x, d) \leq 0$. Thus

$$\begin{aligned}
\mathrm{T}_{Sf(x)}(x) &\subset \mathrm{cl}\{\alpha d : f'(x, d) \leq 0, \alpha \geq 0\} && (223)\\
&= \mathrm{cl}\{d : f'(x, d) \leq 0\} && f'(x, \cdot) \text{ is positively homogeneous} \quad (224)\\
&= \{d : f'(x, d) \leq 0\} && (225)
\end{aligned}$$

where the last line follows because, as a finite convex function, $f'(x, \cdot)$ is continuous by proposition A.2, so has closed sublevel sets. □

Proposition A.17 (IV.1.3.4 in Hiriart-Urrutty and Lemaréchal[12])**.** Let $f : X \to \mathbb{R}$ be convex. Suppose there exists $x_0 \in X$ such that

$$f(x_0) < 0. \tag{226}$$

Then

$$\mathrm{cl}\{x \in X : f(x) < 0\} = \{x \in X : f(x) \leq 0\} \tag{227}$$

Proof. By Proposition A.2 we know f is continuous, so $\{x \in X : f(x) \leq 0\}$ is closed and hence

$$\mathrm{cl}\{x \in X : f(x) < 0\} \subset \{x \in X : f(x) \leq 0\}. \tag{228}$$

To show the other inclusion let $y \in \{x \in X : f(x) \leq 0\}$. Let $x_\alpha := (1-\alpha)x_0 + \alpha y$. By convexity we have for $0 < \alpha < 1$ that

$$f(x_\alpha) \leq (1-\alpha)f(x_0) + \alpha f(y) \tag{229}$$

$$< 0 \qquad (226) \text{ and def of } y \tag{230}$$

Since $x_\alpha \to y$ as $\alpha \to 1$ we have $y \in \mathrm{cl}\{x \in X : f(x) < 0\}$ as desired. □

Proposition A.18 (Theorem VI.1.3.4 in Hiriart-Urrutty and Lemaréchal[12])**.** Let $f : X \to \mathbb{R}$ be convex. Assume x is such that $0 \notin \partial f(x)$. Then

$$\mathrm{T}_{Sf(x)}(x) = \{d \in X : f'(x,d) \leq 0\}. \tag{231}$$

Proof. By Propositions A.5 we already have the $\subset$ inclusion. To prove the other inclusion, note that if $d \in X$ satisfies the strict inequality $f'(x,d) < 0$ then by Definition A.4 there is a $t > 0$ such that $f(x+td) - f(x) < 0$. We have $d = \frac{x+td-x}{t}$ so that $d \in \mathrm{T}_{Sf(x)}(x)$. So we have shown

$$\mathrm{T}_{Sf(x)}(x) \supset \{d : f'(x,d) < 0\}. \tag{232}$$

By the assumption $0 \notin \partial f(x)$ we know that x does not minimize f and thus by A.4, the set $\{d : f'(x,d) < 0\}$ is nonemtpy. This nonemptiess allows us to apply Proposition A.17 to the convex function $f'(x,\cdot)$ to get

$$\mathrm{T}_{Sf(x)}(x) \supset \{d : f'(x,d) \leq 0\} \tag{233}$$

as desired. □

Proposition A.19 (Theorem VI.1.3.5 in Hiriart-Urruty and Lemaréchal[12].)**.** Let $f : X \to \mathbb{R}$ be convex, and let $x \in \mathbb{R}^n$ be such that $0 \notin \partial f(x)$. Then

$$\mathrm{N}_{Sf(x)}(x) = \mathbb{R}_+ \partial f(x). \tag{234}$$

Proof. We have:

$$\mathrm{T}_{Sf(x)}(x) = \{d : f'(x,d) \leq 0\} \qquad \text{by A.18} \tag{235}$$

$$= \{d : \langle s, d\rangle \leq 0 \, \forall s \in \partial f(x)\} \qquad \text{by A.12} \tag{236}$$

$$= \{d : \langle ts, d\rangle \leq 0 \, \forall s \in \partial f(x), \, \forall t \geq 0\} \tag{237}$$

$$= (\mathbb{R}_+ \partial f(x))^\circ. \tag{238}$$

Take polars of both sides to get

$$\mathrm{N}_{Sf(x)}(c) = \mathrm{cl}\, \mathbb{R}_+ \partial f(x). \tag{239}$$

We know that $0 \notin \partial f(x)$, so that by Proposition A.6

$$\mathrm{cl}(\mathbb{R}_+ \partial f(x)) = \mathbb{R}_+ \partial f(x) \tag{240}$$

and we are done. □

B The Nuclear Norm

Let $A \in \mathbb{R}^{m\times n}$. The nuclear norm of A, denoted $\|A\|_*$, is the sum of its singular values. The pertinent facts about the nuclear norm are

1. The nuclear norm is a norm on $\mathbb{R}^{m\times n}$.

2. **Proposition B.1.** Let $A \in \mathbb{R}^{m\times n}$. Let A have singular value decomposition $A = FDG^T$ where $F \in \mathbb{R}^{m\times r}$ $D \in \mathbb{R}^{r\times r}$ and $G \in \mathbb{R}^{n\times r}$. Then

$$\partial\|A\|_* = \begin{cases} B_{\|\cdot\|} \text{ if } A = 0 \\ \{C \in \mathbb{R}^{m\times n} : C = FG^T + W, F^TW = WG = 0, \|W\| \leq 1\} \text{ else} \end{cases} \tag{241}$$

Because this essay is already too long, I will not go into details. However, I believe the key attributes of the nuclear norm are consequences of this proposition.

Proposition B.2. Let $A, B \in \mathbb{R}^{m\times n}$ then

$$|\langle A, B\rangle| \leq \|A\|\|B\|_* \tag{242}$$

where we are using the Frobenius inner product.

Proof. Let B have singular value decomposition $B = \sum_{i=1}^r \sigma_i f_i g_i^T$.

$$\begin{aligned}
|\langle A, B\rangle| &= |\operatorname{tr}(A^TB)| && (243)\\
&= |\sum_{i=1}^r \sigma_i \operatorname{tr}(A^T f_i g_i^T)| && (244)\\
&= |\sum_{i=1}^r \sigma_i g_i^T A^T f_i| && \text{Property of Trace} \quad (245)\\
&\leq \sum_{i=1}^r \sigma_i |g_i^T A^T f_i| && (246)\\
&\leq \sum_{i=1}^r \sigma_i \|A^T f_i\| && \text{Cauchy-Schwarz} \quad (247)\\
&\leq \sum_{i=1}^r \sigma_i \|A\| && \text{definition of operator norm} \quad (248)\\
&= \|A\|\|B\|_* && \text{definition of nuclear norm} \quad (249)
\end{aligned}$$

□

C Probability

C.1 General Facts

Proposition C.1 (Markov's inequality)**.** Let Λ be a nonnegative random variable. Let $a > 0$. Then

$$P(\Lambda \geq a) \leq \frac{\mathbf{E}\,\Lambda}{a} \tag{250}$$

Proof. We have

$$\Lambda = \Lambda \mathbf{1}_{\Lambda \geq a} + \Lambda \mathbf{1}_{\Lambda < a} \geq a \mathbf{1}_{\Lambda \geq a}. \tag{251}$$

Take the expectation of both sides to get the result. □

Proposition C.2 (Jensen's Inequality for Integrals. From Exercise 3.42 in Folland[7])**.** Let μ be a probability measure on on $\mathbb{R}$ such that $\int x \,\mathrm{d}\mu(x)$ is finite. Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$f\left(\int x \,\mathrm{d}\mu(x)\right) \leq \int f(x) \,\mathrm{d}\mu(x). \tag{252}$$

Proof. Let $x_0 := \int x \,\mathrm{d}\mu(x)$. Since f is a finite convex function, we can find a non-vertical hyperplane supporting its epigraph at $(x_0, f(x_0))$. In other words, we can find an affine function A such that

$$A(x) \leq f(x) \,\forall x \tag{253}$$
$$A(x_0) = f(x_0). \tag{254}$$

Using this and the fact that affine functions can be passed through integrals with respect to probability measures, we have

$$f\left(\int x \,\mathrm{d}\mu(x)\right) = f(x_0) = A(x_0) = A\left(\int x \,\mathrm{d}\mu(x)\right) = \int A(x) \,\mathrm{d}\mu(x) \leq \int f(x) \,\mathrm{d}\mu(x). \tag{255}$$

□

Proposition C.3. Let Λ_1, Λ_2 be independent random variables on the Euclidean space X. Let $g : X \times X \to \mathbb{R}$ be measurable. Assume $\mathbf{E}\, g(\Lambda_1, \Lambda_2)$ is finite. Define the function $h(x) : X \to \mathbb{R}$ by $h(x) = \mathbf{E}\, g(x, \Lambda_2)$. Then we have

$$\mathbf{E}\, g(\Lambda_1, \Lambda_2) = \mathbf{E}\, h(\Lambda_1) \tag{256}$$

Proof. Let $(\Omega, \mathscr{F}, P)$ Be the underlying measure space. r Tao[17], to say that Λ_1, Λ_2 are independent is to say that this probability space can be factored as $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \times \mathscr{F}_2, P_1 \times P_2)$ such that for any $(s_1, s_2) \in \Omega_1 \times \Omega_2$, $\Lambda_1(s_1, s_2)$ depends only on s_1 and $\Lambda_2(s_1, s_2)$ depends only on s_2.

Then applying Fubini-Tonei, we have

$$\begin{aligned}
&\mathbf{E}\, g(\Lambda_1, \Lambda_2) && (257)\\
&= \int g(\Lambda_1(s_1), \Lambda_2(s_2))\, \mathrm{d}P(s_1, s_2) && (258)\\
&= \iint g(\Lambda_1(s_1), \Lambda_2(s_2))\, \mathrm{d}P_2(s_2)\, \mathrm{d}P_1(s_1) && (259)\\
&= \int h(\Lambda_1(s_1))\, \mathrm{d}P_1(s_1) && (260)\\
&= \mathbf{E}\, h(\Lambda_1). && (261)
\end{aligned}$$

□

C.2 Gaussian Vectors

The following result is useful when working with Gaussian.

Proposition C.4 (Integral of the exponential of a quadratic)**.** Let $a \in \mathbb{R}$, $b > 0$. Then

$$\int \exp(ax - bx^2) dx = \sqrt{\frac{\pi}{b}} \exp \frac{a^2}{4b}. \tag{262}$$

Proof.

$$\begin{aligned}
\int \exp(ax - bx^2)\, \mathrm{d}x &= \exp\left(\frac{a^2}{4b}\right) \int \exp\left(-(bx^2 - ax + \frac{a^2}{4b})\right)\, \mathrm{d}x && (263)\\
&= \exp\left(\frac{a^2}{4b}\right) \int \exp\left(-\left(\sqrt{b}x - \frac{a}{2\sqrt{b}}\right)^2\right)\, \mathrm{d}x. && (264)
\end{aligned}$$

Now change variables to $z = \sqrt{b}x - \frac{a}{2\sqrt{b}}$ so that $dx = \frac{1}{\sqrt{b}} dz$. Then we have

$$\frac{\exp\left(\frac{a^2}{4b}\right)}{\sqrt{b}} \int \exp(-z^2)\, \mathrm{d}z = \sqrt{\frac{\pi}{b}} \exp \frac{a^2}{4b}. \tag{265}$$

as desired. □

Proposition C.5.

1. If Λ has distribution Normal$(0, \sigma^2)$ and $a \in \mathbb{R}$ is nonzero then $a\Lambda$ has distribution Normal$(0, a^2\sigma^2)$.
2. If Λ_1 and Λ_2 have distributions Normal$(0, \sigma_1^2)$ and Normal$(0, \sigma_2^2)$ and are independent, then $\Lambda_1 + \Lambda_2$ has distribution Normal$(0, \sigma_1^2 + \sigma_2^2)$

Proof.

1. If $a > 0$ we have

$$P(a\Lambda \leq s) = P(\Lambda \leq \frac{s}{a}) \tag{266}$$

$$= \int_{-\infty}^{\frac{s}{a}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-x^2}{2\sigma^2})\,\mathrm{d}x \tag{267}$$

$$= \int_{-\infty}^{s} \frac{1}{\sqrt{2\pi a^2\sigma^2}} \exp(\frac{-x^2}{2a^2\sigma^2})\,\mathrm{d}x. \tag{268}$$

If $a < 0$, note that by symmetry $-\Lambda$ has the same distribution as Λ and apply the above.

2. Using the independence of Λ_1 and Λ_2, the density of their sum is the convolution of their densities.

$$f(z) = \int \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}} \exp(\frac{-x^2}{2\sigma_1^2}) \exp(-\frac{(z-x)^2}{2\sigma_2^2})\,\mathrm{d}x \tag{269}$$

$$= \int \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}} \exp\left(\frac{-1}{2\sigma_1^2\sigma_2^2}\left((\sigma_1^2+\sigma_2^2)x^2 - 2\sigma_1^2 zx + \sigma_1^2 z^2\right)\right)\,\mathrm{d}x. \tag{270}$$

Apply Proposition C.4 with $b = \frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}$ $a = \frac{z}{\sigma_2^2}$ to get

$$f(z) = \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}} \sqrt{\frac{2\pi\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} \exp\left(\frac{z^2}{\sigma_2^4}\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\frac{1}{4} - \frac{\sigma_1^2 z^2}{2\sigma_1^2\sigma_2^2}\right) \tag{271}$$

$$= \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}} \sqrt{\frac{2\pi\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}} \exp\left(\frac{z^2}{2\sigma_2^2}\left(\frac{\sigma_1^2}{\sigma_1+\sigma_2^2} - 1\right)\right) \tag{272}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left(-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{273}$$

as desired.

□

The next result is sometimes useful in relation to standard Gaussian vectors, who distributions are rotationally invariant.

Proposition C.6. Fix a positive integer n and let I_n denote the $n \times n$ identity matrix. For any $\theta \in [0, 2\pi]$ the $2n \times 2n$ matrix

$$U_\theta := \begin{bmatrix} \cos\theta I_n & \sin\theta I_n \\ -\sin\theta I_n & \cos\theta I_n \end{bmatrix} \tag{274}$$

defines an isometry.

Proof. It suffices to check that $U^*U = I$. This follows from the Pythagorean identity. □

Proposition C.7 (Generating Function of a Gaussian Random Variable)**.** Let $\Lambda \sim \text{Normal}(\mu, \sigma^2)$. Let $t \in \mathbb{R}$. Then

$$\mathbf{E}\exp(t\Lambda) = \exp(t\mu + \frac{1}{2}t^2\sigma^2) \tag{275}$$

Proof.

$$\mathbf{E}\exp(t\Lambda) = \frac{1}{\sqrt{2\sigma^2\pi}}\int \exp(tx)\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx \tag{276}$$

$$= \frac{\exp(\frac{-\mu^2}{2\sigma^2})}{\sqrt{2\sigma^2\pi}}\int \exp(-\frac{x^2}{2\sigma^2} + (t+\frac{\mu}{\sigma^2})x)\,\mathrm{d}x. \tag{277}$$

Define $a := (t + \frac{\mu}{\sigma^2})$ and $b := \frac{1}{2\sigma^2}$ and apply Proposition C.4. This gives us

$$\sqrt{2\sigma^2\pi}\frac{\exp(\frac{-\mu^2}{2\sigma^2})}{\sqrt{2\sigma^2\pi}}\exp\left(\frac{1}{4}\left(t+\frac{\mu}{\sigma^2}\right)^2 2\sigma^2\right) = \exp(\frac{-\mu^2}{2\sigma^2})\exp\left(\frac{1}{2}\sigma^2\left(t^2 + \frac{2t\mu}{\sigma^2} + \frac{\mu^2}{\sigma^4}\right)\right) \tag{278}$$

$$= \exp(t\mu + \frac{\sigma^2 t^2}{2}) \tag{279}$$

as desired. □

The following result bounds the tail probability of Gaussian random variables.

Proposition C.8. Let $x > 0$

$$\int_x^\infty \exp\frac{-t^2}{2}\,\mathrm{d}t \leq \frac{1}{x}\exp(\frac{-x^2}{2}) \tag{280}$$

Proof.

$$\int_x^\infty \exp\frac{-t^2}{2}\,\mathrm{d} \leq \int_x^\infty \frac{t}{x}\exp\frac{-t^2}{2}\,\mathrm{d}t \tag{281}$$

$$= \frac{1}{x}\int_x^\infty \frac{d}{dt}\left(-\exp\frac{-t^2}{2}\right)\,\mathrm{d}t \tag{282}$$

$$= \frac{1}{x}\exp\frac{-x^2}{2} \tag{283}$$

□

Proposition C.9. [The χ^2 distribution. Lemma 7.9 in Foucart and Rauhut[6]] Let g be a standard Gaussian vector in $\mathbb{R}^n$. Then $Z := \|g\|^2$ has density function

$$\phi(u) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}u^{\frac{n}{2}-1}\exp(-\frac{u}{2}) \text{ if } u > 0 \\ 0 \text{ else} \end{cases} \tag{284}$$

Proof. See Lemma 7.9 in Foucart and Rauhut[6]. □

Proposition C.10 (Norms of Gaussian vectors. Proposition 8.1 in Foucart and Rauhut[6])**.** Let g be a standard Gaussian vector in an n-dimensional Euclidean space. Then

1.

$$\mathbf{E}\|g\|^2 = n \tag{285}$$

2.

$$\frac{n}{\sqrt{n+1}} \leq \mathbf{E}\|g\| = \sqrt{2}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \leq \sqrt{n} \tag{286}$$

Proof. To prove (285) note that the quantity in question is the sum of the variances of n standard normal random variables.

To show, (286), first apply the convexity of $-\sqrt{\cdot}$ and Proposition C.2 (Jensen's Inequality):

$$-\mathbf{E}\|g\| = \mathbf{E} - \sqrt{\|g\|^2} \geq -\sqrt{\mathbf{E}\|g\|^2} = -\sqrt{n} \tag{287}$$

so that

$$\mathbf{E}\|g\| \leq \sqrt{n}. \tag{288}$$

Next use Proposition C.9 to compute

$$\begin{aligned}
\mathbf{E}\|g\| &= \int_0^\infty u^{\frac{1}{2}}\phi(u)\,\mathrm{d}u && (289)\\
&= \int_0^\infty \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} u^{\frac{n}{2}-\frac{1}{2}} \exp(\frac{-u}{2})\,\mathrm{d}u && (290)\\
&= \frac{2^{\frac{n}{2}+\frac{1}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_0^\infty t^{\frac{n}{2}-\frac{1}{2}} \exp(-t)\,\mathrm{d}t \qquad \text{Let } t = \frac{u}{2} && (291)\\
&= \frac{\sqrt{2}}{\Gamma(\frac{n}{2})}\Gamma(\frac{n}{2}+\frac{1}{2}) && (292)
\end{aligned}$$

where the last line follows from the definition of the gamma function (see Folland[7] page 58). Finally, to get the lower bound, note that if we let g_{n+1} denote a standard Gaussian vector in an $n+1$ dimensional vector space, then

$$(\mathbf{E}\|g\|)(\mathbf{E}\|g_{n+1}\|) = 2\frac{\Gamma(\frac{n}{2}+\frac{1}{2})}{\Gamma(\frac{n}{2})}\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n}{2}+\frac{1}{2})} = 2\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n}{2})} = 2(\frac{n}{2}) = n \tag{293}$$

where we have used the functional equation for the gamma function (see Folland[7] page 58). Then using (288)

$$\mathbf{E}\|g\| = \frac{n}{\mathbf{E}\|g_{n+1}\|} \geq \frac{n}{\sqrt{n+1}} \tag{294}$$

as desired. □

The following result is used by Chandrasekaran et al.[3] in their bound of the Gaussian width of the tangent cone of the nuclear norm. It appears as part of a very dense paper, and so I have not investigated it further.

Proposition C.11 (Theorem II.13 Davidson and Szarek[5])**.** Let Λ be a random linear transformation between Euclid an spaces whose matrix with respect to a pair of orthonomal bases is $m \times n$ with all entries independent Normal$(0,1)$. Let $t \geq 0$. Then

$$P(\|\Lambda\| \geq \sqrt{m} + \sqrt{n} + t) \leq \exp(\frac{-t^2}{2}) \tag{295}$$

References

[1] Amelunxen, Lotz, Mccoy, and Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: a Journal of the IMA* (2014) 3, 224-294.

[2] Cai and Xu. Guarantees of total variation minimization for signal recovery. *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on. IEEE* (2013).

[3] Chandrasekaran, Recht, Parrilo, and Willsky. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, (2012) 12: 805-849.

[4] Chen, Donoho, and Saunders, . Atomic decomposition by basis pursuit. *SIAM review* (2001), 43(1), 129-159.

[5] Davidson and Szarek. *Local Operator Theory, Random Matricies, and Banach Spaces* in Handbook of Geometry of Banach Spaces Volume 1 (2001).

[6] Foucart and Rauhut. *A Mathematical Introduction to Compressive Sensing.* Birkhauser (2013).

[7] Folland. *Real Analysis.* Wiley and Sons (1999).

[8] Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric aspects of Functional Analysis* (1988), 1317: 84-106.

[9] Grimmett and Stirzaker. *Probability and Random Processes.* Oxford University press (1992).

[10] Gordon. Some inequalities For Gaussian Processes and Applications. *Israel Journal of Mathematics* (1985). Vol 50 no 4:265-289.

[11] Halmos. *Finite-Dimensional Vector Spaces.* Litton (1958).

[12] Hiriart-Urruty and Lemaréchal. *Convex Analysis and Minimization Algorithms.* Springer-Verlag (1993).

[13] Hiriart-Urruty and Lemaréchal. *Fundamentals of Convex Analysis.* Springer-Verlag (2004).

[14] Keller. Inverse Problems. *The American Mathematical Monthly.* 83:107-118.

[15] Recht, Fazel, amd Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* (2010), 52(3), 471-501.

[16] Schneider and Weil. *Stochastic and Integral Geometry.* Springer-Verlag (2008).

[17] Tao, Terrance. *Topics in Random Matrix Theory.* American Mathematical Society (2012).

[18] Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996). Vol. 58, No. 1 , 267-288

[19] Wright, Ganesh, Rao, Peng, and Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems* (2009). 2080-2088.