Convex Relaxation for Low-Dimensional Representation: Phase Transitions and Limitations

Thesis by

Samet Oymak

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2015

(Defended June 12, 2014)

Dedicated to my family

Acknowledgments

To begin with, it was a great pleasure to work with my advisor Babak Hassibi. Babak has always been a fatherly figure to me and my labmates. Most of our group are thousands of miles away from their home, and we are fortunate to have Babak as advisor, who always helps us with our troubles, whether they are personal or professional. As a graduate student, Babak's intuition on identifying new problems and motivation for math was a great inspiration for me. He always encouraged me to do independent research and to be persistent with the toughest (mostly mathematical) challenges.

I thank my thesis committee, Professor Maryam Fazel, Professor Joel Tropp, Professor Venkat Chandrasekaran, and Professor PP Vaidyanathan. Maryam has literally been a second advisor to me, and I am very thankful for her professional guidance. I would like to thank Joel for the material he taught me in class as well as for answering my research questions with great patience and for his helpful feedback for this thesis. It was always great to interact with Venkat and PP. Thanks to them, Caltech has been a better and more stimulating place for me.

I would like to thank fellow graduate students and lab-mates for their help and support. My labmates Amin Khajehnejad, Ramya Korlakai Vinayak, Kishore Jaganathan, and Chris Thrampoulidis were also my research collaborators with whom I spent countless hours discussing research and meeting deadlines. Kishore, Hyoung Jun Ahn, and Eyal En Gad were my partners in the fiercest racquetball games. Matthew Thill, Wei Mao, and Ravi Teja Sukhavasi were a crucial part of all the fun in the lab.

During my time at Caltech, Shirley Slattery and Tanya Owen were amazingly helpful with providing information and daily issues. I cannot possibly count how many times I asked Shirley "Is Babak around?".

I can't imagine how I could enjoy my time here without my dear friends Necmiye Ozay, Murat Acar, Aycan Yurtsever, Selim Hanay, and Jiasi Chen. I guess I could meet these great people only at Caltech. They are now literally all over the world; but I do hope our paths will cross again.

Finally, I am most grateful to my family for always being there for me. Their continuous support has been a source of my ambition and motivation since primary school. It has been a challenging five years away from them, and I am very happy to know that, we are still as connected as ever.

Abstract

There is a growing interest in taking advantage of possible patterns and structures in data so as to extract the desired information and overcome the *curse of dimensionality*. In a wide range of applications, including computer vision, machine learning, medical imaging, and social networks, the signal that gives rise to the observations can be modeled to be approximately sparse and exploiting this fact can be very beneficial. This has led to an immense interest in the problem of efficiently reconstructing a sparse signal from limited linear observations. More recently, low-rank approximation techniques have become prominent tools to approach problems arising in machine learning, system identification and quantum tomography.

In sparse and low-rank estimation problems, the challenge is the inherent intractability of the objective function, and one needs efficient methods to capture the low-dimensionality of these models. Convex optimization is often a promising tool to attack such problems. An intractable problem with a combinatorial objective can often be "relaxed" to obtain a tractable but almost as powerful convex optimization problem. This dissertation studies convex optimization techniques that can take advantage of low-dimensional representations of the underlying high-dimensional data. We provide *provable* guarantees that ensure that the proposed algorithms will succeed under reasonable conditions, and answer questions of the following flavor:

- For a given number of measurements, can we reliably estimate the true signal?
- If so, how good is the reconstruction as a function of the model parameters?

More specifically, i) Focusing on linear inverse problems, we generalize the classical error bounds known for the least-squares technique to the lasso formulation, which incorporates the signal model. ii) We show that intuitive convex approaches do not perform as well as expected when it comes to signals that have multiple low-dimensional structures simultaneously. iii) Finally, we propose convex relaxations for the graph clustering problem and give sharp performance guarantees for a family of graphs arising from the so-called stochastic block model. We pay particular attention to the following aspects. For i) and ii), we aim to provide a general geometric framework, in which the results on sparse and low-rank estimation can be obtained as special cases. For i) and iii), we investigate the precise performance characterization, which yields the right constants in our bounds and the true dependence between the problem parameters.

Contents

Chapter 1

Introduction

The amount of data that is being generated, measured, and stored has been increasing exponentially in recent years. As a result, there is a growing interest in taking advantage of possible patterns and structures in the data so as to extract the desired information and overcome the *curse of dimensionality*. In a wide range of applications, including computer vision, machine learning, medical imaging, and social networks, the signal that gives rise to the observations can be modeled to be approximately sparse. This has led to an immense interest in the problem of efficiently reconstructing a sparse signal from limited linear measurements, which is known as the compressed sensing (CS) problem [42, 43, 45, 73]. Exploiting sparsity can be extremely beneficial. For instance, MRI acquisition can be done faster with better spatial resolution with CS algorithms [140]. For image acquisition, the benefits of sparsity go beyond MRI thanks to applications such as the "single pixel camera" [11].

Sparse approximation can be viewed as a specific, albeit major, example of a low-dimensional representation (LDR). The typical problem we consider is one for which the ambient dimension of the signal is very large (think of a high resolution image, gene expression data from a DNA microarray, social network data, etc.), yet is such that its desired properties lie in some low-dimensional structure (sparsity, low-rankness, clusters, etc.). More recently, for instance, low-rank approximation has become a powerful tool, finding use in applications varying from face recognition to recommendation systems [40, 94, 178]. The revolutionary results that started CS are now a decade old; however, CS and LDR are still active research topics opening doors to new applications as well as new challenges.

1.1 Sparse signal estimation

Sparse approximation aims to represent a signal $\mathbf{x}$ as a linear combination of a few elements from a given dictionary $\Psi \in \mathbb{R}^{n \times d}$. In particular, we can write $\mathbf{x} = \Psi\alpha$, where α has few nonzero entries. Ψ depends on the application, for instance; one can use wavelets for natural images. The aim is to parsimoniously represent $\mathbf{x}$ and take advantage of this representation when the time comes. The typical problem in compressed sensing assumes the linear observations of $\mathbf{x}$ of the form

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

Here $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the measurement matrix and $\mathbf{z}$ is the additive noise (statisticians would use the notation $\mathbf{y} = \mathbf{X}\beta + \mathbf{z}$). Depending on the application, $\mathbf{A}$ can be enforced by the problem or it can be up to us to design. To simplify the discussion, we will discard Ψ (or assume it to be identity) with the change of variable $\mathbf{A}\Psi \to \mathbf{A}$ and $\alpha \to \mathbf{x}$. Assuming $\mathbf{A}$ is full-rank, the problem is rather trivial when $m \geq n$, as we can estimate $\mathbf{x}$ with the pseudo-inverse of $\mathbf{A}$. However,

- In a growing list of applications, the signal $\mathbf{x}$ is high-dimensional and the amount of observations m may be significantly smaller than n.
- If we do know that the true signal is approximately sparse, we need a way of encouraging sparsity in our solution even in the overdetermined regime $m \geq n$.

In the noiseless setup ($\mathbf{z} = 0$), to find $\mathbf{x}$, we can enforce $\mathbf{y} = \mathbf{A}\mathbf{x}'$ while trying to minimize the number of nonzero entries of the candidate solution $\mathbf{x}'$

$$\min \|\mathbf{x}'\|_0 \quad \text{subject to} \quad \mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x}.$$

The challenge in this formulation, especially in the $m < n$ regime, is the fact that the sparse structure we would like to enforce is combinatorial and often requires exponential search. Assuming $\mathbf{x}$ has k nonzero entries, $\mathbf{x}$ lies on one of the $\binom{n}{k}$ k-dimensional subspaces induced by the locations of its nonzero entries. $\mathbf{x}$ can be possibly found by trying out each of the $m \times k$ submatrices of $\mathbf{A}$; however, this method becomes exponentially difficult with increasing n and k.

Sparse approximation has been of significant interest at least since the 1990's. Various algorithms have been proposed to tackle this problem. Initial examples include the matching pursuit algorithms by Mallat and Zhang in 1993 [143], the lasso estimator of Tibshirani in 1996 [201] and Basis Pursuit by Donoho and

Chen [58, 59]. Perhaps the most well-known technique is to replace the cardinality function with the ℓ_1 norm of the signal, i.e., the sum of the absolute values of the entries. This takes us from a combinatorially challenging problem to a tractable one which can be solved in polynomial time. Moving from ℓ_0 quasi-norm to ℓ_1 norm as the objective is known as "convex relaxation". The new problem is known as the Basis Pursuit (BP) and is given as [58, 59]

$$\min \|\mathbf{x}'\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x} \tag{1.1}$$

As it will be discussed further in Section 1.2.1, convex relaxation techniques are not limited to sparse recovery. Important signal classes that admit low-dimensional representation allow for convex relaxation.

In general, compressed sensing and sparse approximation techniques aim to provide efficient ways to deal with the original combinatorial problem and reliably estimate $\mathbf{x}$ in the underdetermined regime $m < n$. From a theoretical point of view, our aim is to understand the extent to which the data ($\mathbf{x}$) can be undersampled while allowing for efficient reconstruction. We will often try to answer the following question:

Question 1 *How many observations m do we need to reliably and efficiently estimate $\mathbf{x}$?*

Preferably, the answer should depend on $\mathbf{x}$, only through the problem parameters k and n. The answer is also highly dependent on the specific algorithm we are using. It should be noted that the computational efficiency is another important concern and there is often a tradeoff between the computational efficiency and the estimation performance of the associated algorithm [49].

To address Question 1, since the late 1990's there has been significant efforts in understanding the performance of sparse approximation algorithms. In 2001, Donoho and Huo provided the initial results on the recovery of a sparse signal from linear observations via BP [79]. Tropp jointly studied orthogonal matching pursuit and BP and found conditions on $\mathbf{A}$ for which both both approaches can recover a sparse signal [203].

In practice, BP often performs much better than the theoretical guarantees of these initial works. Later on, it was revealed that with the help of randomness (over $\mathbf{A}$), one can show significantly stronger results for BP. For instance, when $\mathbf{A}$ is obtained by picking m rows of the Discrete Fourier Transform matrix uniformly at random, it has been shown that signals up to $\mathcal{O}\left(\frac{m}{\log n}\right)$ sparsity can be recovered via BP[1]. This is in fact the celebrated result of Candes, Tao, and Romberg that started CS as a field [43]. To see why this is

[1]We remark that Gilbert et al. also considered reconstruction of signals that are sparse in frequency domain using sublinear time combinatorial algorithms [107].

remarkable, observe that m grows almost linearly in the sparsity k and it can be significantly smaller than the ambient dimension n. We should remark that no algorithm can require less than k measurements. We will expand more on the critical role of randomness in CS. The measurement ensembles for which we have strong theoretical guarantees (i.e. linear scaling in sparsity) are mostly random. For instance, an important class of measurement ensembles for which BP provably works are the matrices with independent and identically distributed entries (under certain tail/moment conditions) [45, 73].

1.1.1 Overview of recovery conditions

In general, conditions for sparse recovery ask for $\mathbf{A}$ to be well-behaved and are often related with each other. In linear regression, well-behaved often means that $\mathbf{A}$ is well-conditioned. In other words, denoting maximum singular value by $\sigma_{\max}(\mathbf{A})$ and minimum singular value by $\sigma_{\min}(\mathbf{A})$, we require $\frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$ to be small. For sparse approximation, in the more interesting regime $m \ll n$, we have $\sigma_{\min}(\mathbf{A}) = 0$, hence, one needs to look for other conditions. Some of the conditions that guarantee success of BP are as follows.

• Restricted Isometry Property (RIP) [32,45]: This condition asks for submatrices of $\mathbf{A}$ to be well-conditioned. Let $1 \leq s \leq n$ be an integer. Then, $\mathbf{A}$ satisfies the RIP with restricted isometry constant δ_s, if for all $m \times s$ submatrices $\mathbf{A}_s$ of $\mathbf{A}$, one has,

$$(1-\delta_s)\|\mathbf{v}\|_2^2 \leq \|\mathbf{A}_s\mathbf{v}\|_2^2 \leq (1+\delta_s)\|\mathbf{v}\|_2^2$$

Observe that this is a natural generalization of conditioning of a matrix in the standard linear regression setup. When the system is overdetermined, setting $n = s$, δ_s characterizes the relation between minimum and maximum singular values of $\mathbf{A}$. When RIP holds with $\delta_{2k} \approx 0.453$, it is known that BP will successfully get back to k-sparse $\mathbf{x}$. RIP is alternatively known as the uniform uncertainty principle [46].

• Incoherence [79, 203]: This asks for columns of $\mathbf{A}$ to have low correlation with each other. In particular, the coherence of the matrix $\mathbf{A}$ captures the maximum correlation between any two columns of $\mathbf{A}$ and is defined as follows:

$$\mu_{\mathbf{A}} = \max_{i \neq j} \frac{\left\langle \mathbf{A}_{\{i\}}, \mathbf{A}_{\{j\}} \right\rangle}{\|\mathbf{A}_{\{i\}}\|_2 \|\mathbf{A}_{\{j\}}\|_2}.$$

Unlike the restricted isometry constant, $\mu_{\mathbf{A}}$ can be easily calculated. However, the type of guarantees are not as strong. We should remark that earlier results (before CS was introduced) on BP were based on coherence. However, the number of observations m grew quadratically in sparsity rather than linearly. The

later works [41, 205] show that, almost linear scaling can be achieved by introducing randomness to the sparsity pattern of the signal.

• Null-Space Property (NSP) [74, 85]: NSP is a condition on the null space of $\mathbf{A}$. A typical version is the following.

Definition 1.1 $\mathbf{A}$ *satisfies the* ℓ_1*-NSP of order* k*; if all nonzero* $\mathbf{w}$ *that satisfies* $\mathbf{A}\mathbf{w} = 0$*, also satisfies* $\|\mathbf{w}\|_1 > 2\|\mathbf{w}^k\|_1$*. Here,* $\mathbf{w}^k$ *is the* k *sparse approximation of* $\mathbf{w}$ *obtained by setting all entries* 0 *except the largest* k *(in absolute value).*

RIP and incoherence based conditions are often sufficient but not necessary for BP. Unlike these, NSP is "if and only if" (see Proposition 1.1). If $\mathbf{A}$ satisfies ℓ_1-NSP, BP can recover any k-sparse $\mathbf{x}$; conversely, if NSP does not hold, there exists a k sparse signal for which BP fails. Consequently, careful analysis of NSP can lead us to understand the exact characteristics of BP. The first such analysis is due to Donoho and Tanner who developed precise undersampling theorems when $\mathbf{A}$ has independent standard normal entries [83].

Proposition 1.1 ([96]) *Suppose* $\mathbf{A}$ *satisfies* ℓ_1*-NSP of order* k*. Then,* (1.1) *can recover any* k *sparse* $\mathbf{x}$*. Conversely, if* $\mathbf{A}$ *does not satisfy* ℓ_1*-NSP, there exists a* k*-sparse* $\mathbf{x}$*, for which* (1.1) *fails.*

Proof: Let $\mathbf{x}$ be a k-sparse vector and suppose $\mathbf{x}^*$ is the minimizer of (1.1). Then, $\mathbf{w} = \mathbf{x}^* - \mathbf{x} \in \text{Null}(\mathbf{A})$. Let S be a subset of $\{1, 2, \ldots, n\}$ be the set of nonzero locations (support) of $\mathbf{x}$. We will use the fact that, for $i \in S$, $|\mathbf{x}_i + \mathbf{w}_i| \geq |\mathbf{x}_i| - |\mathbf{w}_i|$. It follows that

$$0 \geq \|\mathbf{x}^*\|_1 - \|\mathbf{x}\|_1 \geq \sum_{i \in S} (|\mathbf{x}_i + \mathbf{w}_i| - |\mathbf{x}_i|) - \sum_{i \notin S} |\mathbf{w}_i| \geq \sum_{i \notin S} |\mathbf{w}_i| - \sum_{i \in S} |\mathbf{w}_i|.$$

Observe that $\sum_{i \notin S} |\mathbf{w}_i| - \sum_{i \in S} |\mathbf{w}_i| = \|\mathbf{w}\|_1 - 2\sum_{i \in S} |\mathbf{w}_i| \geq \|\mathbf{w}\|_1 - 2\|\mathbf{w}^k\|_1 > 0$ for all nonzero $\mathbf{w}$. This implies $\mathbf{w} = 0$. Conversely, if a nonzero $\mathbf{w} \in \text{Null}(\mathbf{A})$ satisfies, $\|\mathbf{w}\|_1 \leq 2\|\mathbf{w}^k\|_1$. Then, choose $\mathbf{x}$ to be $-\mathbf{w}^k$ and observe that $\|\mathbf{x} + \mathbf{w}\|_1 \leq \|\mathbf{x}\|_1$ and $\mathbf{x}$ is not the unique minimizer. ■

1.2 Low-dimensional representation via convex optimization

The abundance of results on sparsity naturally motivates us to extend CS theory beyond sparse recovery. The idea is to apply the powerful techniques developed for CS to new applications. It turns out that this is indeed possible to do, both theoretically and algorithmically. Focusing on the convex optimization techniques, we will exemplify how convex relaxation can be applied to other problems.

1.2.1 Examples of Structured Signals

• **Block sparse signals and $\ell_{1,2}$ minimization:** Block sparsity [91, 147, 174, 175, 195, 204, 208] is a generalization of sparsity in which nonzero entries appear in blocks. One of the first work on such signals is by Rao and Kreutz-Delgado in 1999 [174] (also see [56, 204, 208] for earlier works). Assume $n = bt$ for some positive integers b and t. Given $\mathbf{x} \in \mathbb{R}^n$, partition its entries into t vectors $\{\mathbf{x}^i\}_{i=1}^t \in \mathbb{R}^b$ such that $\mathbf{x} = [\mathbf{x}^1\ \mathbf{x}^2\ \ldots\ \mathbf{x}^t]^T$. $\mathbf{x}$ is called a block sparse signal if only a few of its blocks are nonzero. The "structure exploiting" function is the $\ell_{1,2}$ norm, which is given as

$$\|\mathbf{x}\|_{1,2} = \sum_{i=1}^{t} \|\mathbf{x}^i\|_2.$$

Observe that, the ℓ_1 norm is a special case of the $\ell_{1,2}$ norm where the block length d is equal to 1.

• **Sparse representation over a dictionary:** As we have mentioned previously, often the signal $\mathbf{x}$ is not sparse but it has a sparse representation α over a known dictionary Ψ [34, 75, 89]. In this case, to estimate the signal from compressed observations one can use

$$\hat{\alpha} = \arg\min_{\alpha'} \|\alpha'\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{A}\Psi\alpha', \tag{1.2}$$

and let $\hat{\mathbf{x}} = \Psi\hat{\alpha}$. There are several alternatives to (1.2) (see [34, 211]). We remark that, often, instead of recovery from linear observations $\mathbf{A}\mathbf{x}$, we are simply interested in finding a sparse representation given the signal $\mathbf{x}$. Properties of Ψ plays a critical role in the recoverability of α and $\mathbf{x}$. A related topic is learning a dictionary to sparsely represent a group of signals, which is an active research area by itself [88, 141].

• **Low rank matrices and nuclear norm minimization:** In this case, our signal is a low-rank matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. In order to exploit the low rank structure, one can use the nuclear norm heuristic [39,40,94,128,178]. This is convex relaxation of the rank function. Denoting the i'th largest singular value of a matrix $\mathbf{X}$ by $\sigma_i(\mathbf{X})$, its nuclear norm is denoted by $\|\mathbf{X}\|_\star$ and is given as follows

$$\|\mathbf{X}\|_\star = \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(\mathbf{X}).$$

• **Discrete total variation:** In many imaging applications [26,160,231] the signal of interest $\mathbf{x}$ rarely changes as a function of the coordinates. Consequently, letting $\mathbf{d}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ for $1 \leq i \leq n-1$, the difference vector $\mathbf{d} \in \mathbb{R}^{n-1}$ becomes a sparse vector. To induce this structure, one may minimize the total variation of $\mathbf{x}$,

namely,

$$\|\mathbf{x}\|_{TV} = \|\mathbf{d}\|_1. \tag{1.3}$$

• **Nonuniformly sparse signals and weighted ℓ_1 minimization:** Sometimes, we might have prior information regarding the sparsity pattern of the signal [48, 127, 165, 212]. In particular, the signal $\mathbf{x}$ might be relatively sparser over a certain region and denser over another. To exploit this additional information, we can use a modified ℓ_1 minimization where different weights are assigned to different regions. More rigorously, assume that the set of entries $\{1,2,\ldots,n\}$ is divided into t disjoint sets $S_1,\ldots,S_t$ that correspond to regions with different sparsity levels. Then, given a nonnegative weight vector $\mathbf{w} = [w_1\ w_2 \ldots\ w_t]$, the weighted ℓ_1 norm is given as

$$\|\mathbf{x}\|_w = \sum_{i=1}^{t} w_i \sum_{j\in S_i} |x_j|.$$

• **Other examples:** Low-rank plus sparse matrices (see Section 1.2.3); simultaneously sparse and low-rank matrices, low-rank tensors (see Chapter 5); sparse inverse covariance in graphical models [102]; incorporating convex constraints (e.g., nonnegativity, ℓ_∞-norm, positive semidefiniteness [84, 126]).

1.2.2 Generalized Basis Pursuit

These examples suggest that the success of the ℓ_1 minimization is not an isolated case, and the power of convex relaxation is a commonly accepted phenomenon. A natural question is whether one needs to study these problems individually, or there is a general line of attack to such problems. If we focus our attention to the linear inverse problems, we can consider the generalized basis pursuit (GBP),

$$\min_{\mathbf{x}'} f(\mathbf{x}') \quad \text{subject to} \quad \mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x}. \tag{1.4}$$

Here, $f(\cdot)$ is a convex function that tries to capture the low-dimensionality of $\mathbf{x}$. $f(\cdot)$ is often obtained by relaxing a combinatorial objective function. The core arguments underlying GBP are due to the paper of Rudelson and Vershynin in 2006 [184], and subsequent work by Mendelson et al. [151, 153]. These ideas are later on further generalized and polished by Chandrasekaran et al. [50]. Similar to (1.1), we can consider a null-space condition that ensures success of (1.4). To introduce this, we shall first define the descent set.

Definition 1.2 (Descent set) *Given a convex function $f(\cdot):\mathbb{R}^n \to \mathbb{R}$, the descent set of $f(\cdot)$ at $\mathbf{x}$ is denoted by $\mathscr{D}_f(\mathbf{x})$ and is equal to $\{\mathbf{w}\in\mathbb{R}^n \big| f(\mathbf{x}+\mathbf{w}) \le f(\mathbf{x})\}$.*

The following lemma gives an equivalent of Proposition 1.1 for (1.4).

Proposition 1.2 $\mathbf{x}$ *is the unique minimizer of* (1.4) *if and only if* $\text{Null}(\mathbf{A}) \cap \mathscr{D}_f(\mathbf{x}) = \{0\}$.

Proof: $\text{Null}(\mathbf{A}) \cap \mathscr{D}_f(\mathbf{x}) = \{0\}$ and $\mathbf{x}^*$ is the minimizer. Let $\mathbf{x}^* - \mathbf{x} = \mathbf{w}$. If $\mathbf{w} \neq 0$, $\mathbf{w} \notin \mathscr{D}_f(\mathbf{x})$; which implies $f(\mathbf{x}^*) = f(\mathbf{x}+\mathbf{w}) > f(\mathbf{x})$ and contradicts with the optimality of $\mathbf{x}^*$. Conversely, if there exists a nonzero $\mathbf{w} \in \text{Null}(\mathbf{A}) \cap \mathscr{D}_f(\mathbf{x})$, we have $f(\mathbf{x}+\mathbf{w}) \leq f(\mathbf{x})$. ■

Proposition 1.2 shows that the descent set $\mathscr{D}_f(\mathbf{x})$, and the null space $\text{Null}(\mathbf{A})$ determine the fate of (1.4). In Chapter 2, we will see that Karush-Kuhn-Tucker optimality conditions will provide an alternative condition which is dual to NSP. It will be more convenient to study NSP in terms of the tangent cone of $f(\cdot)$ at $\mathbf{x}$.

Definition 1.3 (Tangent cone) *Tangent cone is denoted by* $\mathscr{T}_f(\mathbf{x})$ *and is obtained by taking the closure of conic hull of* $\mathscr{D}_f(\mathbf{x})$ *(see Chapter 2).*

Clearly, $\text{Null}(\mathbf{A}) \cap \mathscr{T}_f(\mathbf{x}) = \{0\} \implies \text{Null}(\mathbf{A}) \cap \mathscr{D}_f(\mathbf{x}) = \{0\}$. We next introduce the restricted singular value of a matrix.

Definition 1.4 (Restricted singular value (RSV)) *Let* S *be a cone in* $\mathbb{R}^n$ *and let* $\mathbf{A} \in \mathbb{R}^{m \times n}$. *The minimum and maximum restricted singular values of* $\mathbf{A}$ *at* S *are respectively defined as*

$$\sigma_S(\mathbf{A}) = \min_{\mathbf{v} \in S, \|\mathbf{v}\|_2 = 1} \|\mathbf{A}\mathbf{v}\|_2, \; \Sigma_S(\mathbf{A}) = \max_{\mathbf{v} \in S, \|\mathbf{v}\|_2 = 1} \|\mathbf{A}\mathbf{v}\|_2.$$

It should be noted that similar definitions exist in the literature [50, 129, 206]. For instance, observe that the restricted isometry constant defined in Section 1.1.1 can be connected to RSV by choosing the cone S to be the set of at most s sparse vectors. The restricted singular value provides an alternative point of view on GBP. Observe that, $\sigma_{\mathscr{T}_f(\mathbf{x})}(\mathbf{A}) > 0$ is equivalent to $\text{Null}(\mathbf{A}) \cap \mathscr{T}_f(\mathbf{x}) = \{0\}$. Hence, we have

$$\sigma_{\mathscr{T}_f(\mathbf{x})}(\mathbf{A}) > 0 \implies \mathbf{x} \text{ is the unique minimizer of (1.4).} \tag{1.5}$$

On the other hand, a larger restricted singular value will imply that (1.4) is better conditioned and is more robust to noise [50].

When we have noisy observations $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$, we can consider the Basis Pursuit Denoising or lasso variation of this problem.

$$\min_{\mathbf{x}'} \lambda f(\mathbf{x}') + \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}'\|_2^2 \tag{1.6}$$

This version tries to induce a structured signal with the help of $f(\cdot)$ and also tries to fit the observations $\mathbf{y}$ to the estimate $\mathbf{x}'$ with the second term $\|\mathbf{y}-\mathbf{A}\mathbf{x}'\|_2^2$. There are several questions we wish to answer regarding these approaches.

1. How many measurements do we need to recover $\mathbf{x}$ in the noiseless case?
2. What are the bounds on estimation error in the noisy setup?
3. Are there simple and intuitive quantities capturing the behavior of these problems?
4. Is there a systematic way to construct convex $f(\cdot)$ given the signal structure?
5. Is there a gap between what can be done in theory and the performance of the relaxed approaches?

These questions have recently been subject of considerable interest. One of our main contributions will be a comprehensive answer to the second and third questions. In general, it is difficult to find answers that work for all measurement ensembles. As we seek better guarantees, we need to sacrifice the generality of the results. For instance, most results in CS require $\mathbf{A}$ to be randomly generated. The sharp guarantees we obtain in Chapter 3 will require $\mathbf{A}$ to be i.i.d. Gaussian.

1.2.3 Demixing problems

Most of our attention will focus on the linear inverse problem (1.4). However, we shall now introduce the closely related demixing problem, which will be important for Chapter 6. In demixing, we often get to observe the true signal $\mathbf{x}$; however, the signal originates from a linear combination of several structured signals. The task is to identify these components.

Example: Robust principal component analysis. Suppose the matrix of interest $\mathbf{X}$ can be decomposed into a low rank piece $\mathbf{L}$ and a sparse piece $\mathbf{S}$, and hence it is a "mixture" of the low rank and sparse structures. This model is useful in applications such as video surveillance and face recognition [36, 171, 235]. The task is to split $\mathbf{X}$ into its sparse and low-rank components. Often $\mathbf{S}$ is a dense sparse corruption on the desirable data $\mathbf{L}$, hence the name robust PCA. To decompose $\mathbf{X}$, the natural optimization we wish to carry out has the form

$$\{\hat{\mathbf{L}},\hat{\mathbf{S}}\} = \arg \inf_{\mathbf{L}'+\mathbf{S}'=\mathbf{X}} \operatorname{rank}(\mathbf{L}') + \gamma\|\mathbf{S}'\|_0.$$

Candès et al. and Chandrasekaran et al. independently proposed relaxing both objectives to end up with the infimal convolution of the ℓ_1 norm and the nuclear norm, [36, 51]

$$\{\hat{\mathbf{L}}, \hat{\mathbf{S}}\} = \arg \inf_{\mathbf{L}'+\mathbf{S}'=\mathbf{X}} \|\mathbf{L}'\|_\star + \gamma \|\mathbf{S}'\|_1. \tag{1.7}$$

The optimization on the right hand side simultaneously emphasizes sparse and low rank pieces in the given matrix. In Chapter 6, we will cast the graph clustering problem as a low-rank and sparse decomposition problem and propose convex approaches based on (1.7). The performance analysis of (1.7) focuses on the allowable levels of $\text{rank}(\mathbf{L}_0)$ and $\|\mathbf{S}_0\|_0$ for which (1.7) succeeds in identifying the individual components.

The other major example is the morphological component analysis (see Elad et al. [90]). In this case, the signal is a linear combination of signals that are sparse in different basis. $\mathbf{x} = \Psi_1 \alpha_1 + \Psi_2 \alpha_2$. In this case, we can minimize

$$\{\hat{\alpha}_1, \hat{\alpha}_2\} = \arg \min_{\Psi_1 \alpha_1' + \Psi_2 \alpha_2' = \mathbf{x}} \|\alpha_1'\|_1 + \gamma \|\alpha_2'\|_1. \tag{1.8}$$

In a separate line of work, McCoy and Tropp proposed a general formulation for demixing problems in a similar manner to (1.4) [144, 145]. In particular, they studied the case where one of the components is multiplied by a random unitary matrix. This formulation allowed them to obtain sharper bounds compared to the related literature that deals with more stringent conditions [36, 90, 235].

The tools to study the convexified demixing problems (1.7) and (1.8) often parallel those of linear inverse problems. For instance, subgradient calculus plays an important role in both problems. Consequently, joint analysis of both problems appear in several works [4, 36, 133, 221].

1.3 Phase Transitions

Returning to the basis pursuit problem (1.4), we can ask a stronger and more specific version of Question 1.

Question 2 *What is the exact tradeoff between sparsity and measurements to recover* $\mathbf{x}$ *via* (1.1)*?*

Focusing on the sparse recovery setup, in Section 1.1, we have mentioned that $m \sim \mathcal{O}(k \log n)$ samples are sufficient for recovery. However, for practical applications it is crucial to know the true tradeoff between problem parameters. Question 2 has first been studied by Donoho and Tanner for sparse signals and for Gaussian measurement matrices. They obtain upper bounds on the required number of Gaussian measurements which are tight in practice. In short, they show that $m \geq 2k \log \frac{2n}{k}$ samples are sufficient for successful

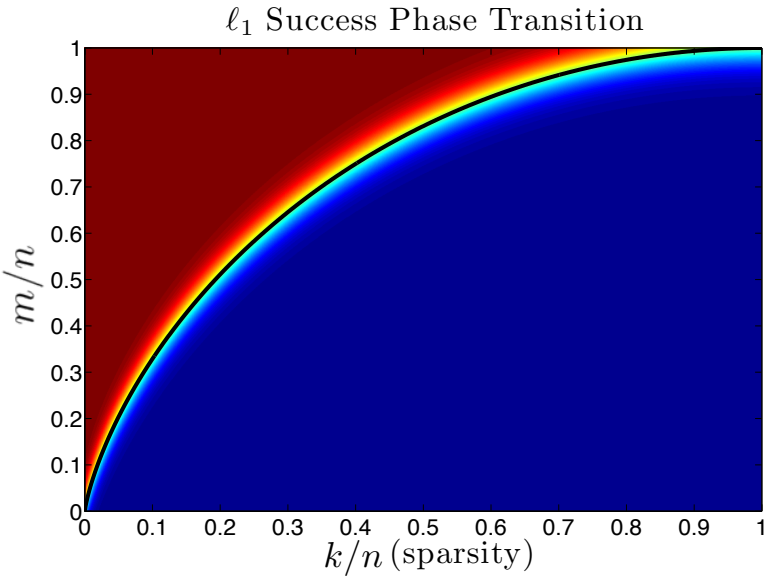


Figure 1.1: The y-axis is the normalized number of measurements. x-axis is the normalized sparsity. The gradient illustrates the gradual increase in the success of BP. As there are more measurements per sparsity (towards the red region), the likelihood of success increases. The black line is the Donoho-Tanner phase transition curve.

reconstruction with high probability. However, they derive the precise behavior as well, which is known as the Donoho-Tanner bound (illustrated in Figure 1.1). Due to the universality phenomenon that we discuss in Section 1.4.5, studying properties of Gaussian matrices gives a good idea about other measurement ensembles as well.

In general, phase transition tries to capture the exact tradeoff between problem parameters and is not limited to linear inverse problems. In Chapter 3, we will investigate a stronger version of Question 2, namely, the precise error bounds for basis pursuit denoising. On the other hand, in Chapter 6, we will investigate the phase transitions of the graph clustering problem where we make use of convex relaxation.

1.4 Literature Survey

With the introduction of compressed sensing, recent years saw an explosion of interest to high dimensional estimation problems. There have been several theoretical, algorithmic, and applied breakthroughs in a relatively short period of time, and signal processing, statistics, and machine learning have been converging to a unified setting. Our literature review will mostly focus on contributions in the theory of high dimensional statistics and convex optimization techniques. Let us start with the results on sparse recovery, in particular,

ℓ_1 minimization.

1.4.1 Sparse signal estimation and ℓ_1 minimization

Sparse signals show up in a variety of applications, and their properties have drawn attention since the 1990's. Donoho and Johnstone used ℓ_1 regularization on wavelet coefficients in the context of function estimation [80]. Closer to our interests, the lasso was introduced by Tibshirani [201] in 1994 as a noise robust version of (1.1). The original formulation was

$$\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \tau.$$

where $\tau \geq 0$ is a tuning parameter. At the same time, Donoho and Chen studied BP from a sparse signal representation perspective, where $\mathbf{A}$ is a dictionary and the aim is to represent $\mathbf{y}$ as a linear combination of few elements (columns of $\mathbf{A}$) [58, 59, 79]. We should remark that "orthogonal matching pursuit" is a greedy algorithm and has been extensively studied as an alternative technique [203, 207].

With the introduction of compressed sensing, randomness started playing a critical role in theoretical guarantees. While guarantees for deterministic matrices require $m \geq \mathscr{O}\left(k^2\right)$, for reasonable random measurement ensembles one requires $m \geq \mathscr{O}\left(k \log n\right)$. First such result is due to Candes, Tao, and Romberg [43]. They considered randomly subsampling the rows of the Discrete Fourier Transform matrix, and have shown that sparse signals can be recovered from incomplete frequencies. Later results included guarantees for matrices with i.i.d entries [42, 45, 73]. Today, state of the art results make even weaker assumptions which cover a wide range of measurement ensembles. As a generalization of sparsity, block-sparsity and non-uniform sparsity have been studied in detail [48, 91, 127, 195]. We should emphasize that the literature on sparse recovery is vast and we only attempt to cover a small but relevant portion of it here.

1.4.2 Phase transitions of sparse estimation

As we have discussed previously, it is desirable to understand the precise behavior of sparse estimation problems, hence Question 2. For the Gaussian measurement ensemble it is known that the recovery of a sparse vector $\mathbf{x}$ depends only on its sparsity level, and is independent of the locations or values of the nonzero entries. Hence, we are interested in the relation between the sparsity k, ambient dimension n, and the number of samples m that guarantees success of BP.

Recovery types: Question 2 is interested in so-called weak-recovery, as we are interested in a recovery of a particular vector $\mathbf{x}$. There is also the notion of strong recovery, which asks for the same (random) measurement matrix to recover all k-sparse vectors via BP. Observe that, Proposition 1.1 gives the condition for the strong recovery. Clearly, strong recovery will require more measurements for the same sparsity level compared to the weak recovery.

Neighborly polytopes: As we discussed in the previous section, Donoho and Tanner found upper bounds on the required number of samples by studying neighborly polytopes and grassman angle computations. They show that their bounds are tight in the regime where the relative sparsity $\frac{k}{n}$ tends to zero. Remarkably, these upper bounds were observed to be tight in simulation for all sparsity regimes [74, 83, 85]. The fact that the Donoho-Tanner bound is apparently tight gained considerable attention. As follow-up works Xu, Khajehnejad, and Hassibi studied several variations of the phase transition problem by extending the Grassman Angle approach. In [225, 226], they showed that when m is above the bound, BP also enjoys robustness when recovering approximately sparse signals. In [127, 224, 227] they have analyzed weighted and reweighted ℓ_1 minimization algorithms, and have provably shown that such algorithms could allow one to go beyond the Donoho-Tanner bound.

Gaussian comparison inequalities: In 2008, Versynin and Rudelson used Gaussian comparison inequalities to study BP [184]. They found $\approx 8k \log \frac{n}{k}$ measurements to be sufficient with a short argument. This is strikingly close to the optimal closed form bound $2k \log \frac{n}{k}$. Later on, Stojnic carried out a more careful analysis of comparison inequalities to obtain better bounds for ℓ_1 minimization [192]. Remarkably, Stojnic's analysis was able to recover the Donoho-Tanner bound with a more generalizable technique (compared to Grassman angle), which can be extended to other structured signal classes. The tightness of Donoho-Tanner bound has been proven by Amelunxen et al. [4], as well as Stojnic [194]. Namely, they show that, below the Donoho-Tanner bound, BP will fail with high probability (also see Section 1.4.4).

Approximate Message Passing (AMP) algorithms: Message-passing algorithms have been introduced by Donoho, Maleki, and Montanari as an alternative to BP thanks to their computational efficiency [81]. Remarkably, it has been observed that the sparsity–measurement trade-off of AMP matches the Donoho-Tanner bound; which makes it appealing as one can have as good performance as ℓ_1 minimization with a low-complexity algorithm. This was later proven by Bayati and Montanari [13, 14]. It should be noted that, AMP can be extended to accommodate several other penalties in a similar manner to GBP [70]

1.4.3 Low-rank estimation

The abundance of strong theoretical results on sparse estimation led researchers to consider other low-dimensional representation problems. Low-rank matrices show up in a variety of applications and they resemble sparse signals to a good degree. The rank minimization problem has first been considered in low-order control system design problems [95, 178]. Convex relaxation for the rank minimization (RM) problem is due to Fazel, who replaced rank function with nuclear norm and also cast it as a semidefinite program [94]. Initial results on RM were limited to applications and implementations.

The significant theoretical developments in this problem came relatively later. This was made possible by advances in sparse estimation theory and similarities between sparse and low-rank recovery problems. Recht, Fazel, and Parrilo studied the RM problem from a compressed sensing point of view, where they studied the number of measurements required to recover a low-rank matrix [178]. They introduced matrix RIP and showed that it is sufficient to recover low-rank matrices via NNM. For i.i.d subgaussian measurements, they also show that matrix RIP holds with $\mathcal{O}(rn\log n)$ measurements. This is a significant result, as one needs at least $\mathcal{O}(rn)$ measurements to accomplish this task (information theoretic lower bound). Candes and Plan improved their bound to $\mathcal{O}(rn)$, which is information theoretically optimal [39].

Another major theoretical development on RM came with results on matrix completion (MC). Low-rank matrix completion is a special case of the rank minimization problem. In MC, we get to observe the entries of the true matrix, which we believe to be low-rank. Hence, measurement model is simpler compared to i.i.d measurements obtained by linear combinations of the entries weighted by independent random variables. This measurement model also has more applications, particularly in image processing and recommendation systems. Results on MC requires certain incoherence conditions on the underlying matrix. Basically, the matrix should not be spiky and the energy should be distributed smoothly over the entries. The first result on MC is due to Candes and Recht [40], who show that $\mathcal{O}\left(r^2 n\cdot \text{polylog}(n)\right)$ measurements are sufficient for MC via nuclear norm minimization. There is a significant amount of theory dedicated to improving this result [37, 60, 125, 128]. In particular, the current best known results require $\mathcal{O}\left(rn\log^2 n\right)$, where the theoretical lower bound is $\mathcal{O}(rn\log n)$ [47, 177].

Phase transitions: When the measurement map is i.i.d Gaussian, Recht et al. obtained the initial Donoho-Tanner type bounds for the nuclear norm minimization [179]. Their rather loose bounds were significantly improved by Oymak and Hassibi who found the exact rank-measurement tradeoff by careful "Gaussian width" calculations following Stojnic's approach [163, 165]. In particular, as small as $6rn$

measurements are sufficient to guarantee low-rank recovery via NNM. Chandrasekaran et al. have similar results, which will be discussed next [50].

1.4.4 General approaches

The abundance of results on the estimation of structured signals naturally led to the development of a unified understanding of these problems. Focusing on the linear inverse problems, we can tackle the generalized basis pursuit,

$$\min f(\mathbf{x}') \quad \text{subject to} \quad \mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x}. \tag{1.9}$$

This problem has recently been a popular topic. The aim is to develop a theory for the general problem and recover specific results as an application of the general framework. A notable idea in this direction is the atomic norms, which are functions that aim to parsimoniously represent the signal as a sum of a few core signals called "atoms". For instance, if $\mathbf{x}$ is sparse over the dictionary Ψ, the columns of Ψ will be the corresponding atoms. This idea goes back to mid 1990's where it appears in the approximation theory literature [68]. Closer to our discussion, importance of minimizing these functions, is first recognized by Donoho in the context of Basis Pursuit [58]. More recently, in connection to atomic norms, Chandrasekaran et al. [50] analyze the generalized basis pursuit with i.i.d Gaussian measurements using Gaussian comparison results. They generalize the earlier results of Vershynin, Rudelson and Stojnic [184, 192] (who used similar techniques to study BP) and find upper bounds to the phase transitions of GBP, which are seemingly tight. Compared to the neighborly polytopes analysis of [74, 83, 224], Gaussian comparison inequalities result in more geometric and intuitive results; in particular, "Gaussian width" naturally comes up as a way to capture the behavior of the problem (1.9). While the Gaussian width was introduced in 1980's [111, 112], its importance in compressed sensing is discovered more recently by Vershynin and Rudelson [184]. Gaussian width is also important for our exposition in Section 1.5, hence we will describe how it shows up in (1.9).

Definition 1.5 (Gaussian width) *Let $\mathbf{g} \in \mathbb{R}^n$ be a vector with independent standard normal entries. Given a set $S \subset \mathbb{R}^n$, its Gaussian width is denoted by $\omega(S)$ and is defined as*

$$\omega(S) = \mathbb{E}[\sup_{\mathbf{v} \in S} \mathbf{v}^T \mathbf{g}].$$

Gaussian width is a way of measuring size of the set, and it captures how well a set is aligned with a random vector. The following result is due to Gordon [112] and it finds a bound on the minimum restricted singular

values of Gaussian matrices.

Proposition 1.3 *Denote the unit ℓ_2-ball by $\mathscr{B}^{n-1}$. Suppose S is a cone in $\mathbb{R}^n$. Let $\mathbf{G} \in \mathbb{R}^{m\times n}$ have independent standard normal entries. Then, with probability $1-\exp(-\frac{t^2}{2})$*

$$\sigma_S(\mathbf{G}) \geq \sqrt{m-1} - \omega(S\cap\mathscr{B}^{n-1}) - t.$$

This result is first used by Rudelson and Vershynin for basis pursuit [184]. Further developments in this direction are due to Mendelson et al. [151, 153]. Chandrasekaran et al. have observed that this result can be used to establish success of the more general problem GBP. In particular, combining (1.5) and Proposition 1.3 ensures that, if $m \geq (\omega(\mathscr{T}_f(\mathbf{x})\cap\mathscr{B}^{n-1})+t)^2+1$, GBP will succeed with high probability. Stojnic was the first person to do a careful and tight analysis of this for basis pursuit and to show that Donoho-Tanner phase transition bound is in fact equal to $\omega(\mathscr{T}_{\ell_1}(\mathbf{x})\cap\mathscr{B}^{n-1})^2$ [192].

The Gaussian width has been subject of several follow-up works, as it provides an easy way to analyze the linear inverse problems [4, 18, 49, 163, 165, 175]. In particular $\omega(\mathscr{T}_f(\mathbf{x})\cap\mathscr{B}^{n-1})^2$ has been established as an upper bound on the minimum number of measurements to ensure success of GBP. Remarkably, this bound was observed to be tight in simulation, in other words, when $m < \omega(\mathscr{T}_f(\mathbf{x})\cap\mathscr{B}^{n-1})^2$, GBP would fail with high probability.

It was proven to be tight more recently by Amelunxen et al. [4]. Amelunxen et al.'s study is based on the intrinsic volumes of convex cones, and their results are applicable in demixing problems as well as linear inverse problems. Stojnic's lower bound is based on a duality argument; however, he still makes use of comparison inequalities as his main tool [194].

1.4.5 Universality of the phase transitions

Donoho-Tanner type bounds are initially proven only for Gaussian measurements. However, in simulation, it is widely observed that the phase transition points of different measurement ensembles match [71]. Examples include,

- Matrices with i.i.d subgaussian entries
- Randomly subsampling rows of certain deterministic matrices such as Discrete Fourier Transform and Hadamard matrices

While this is a widely accepted phenomenon, theoretical results are rather weak. Bayati et al. recently showed the universality of phase transitions for BP via connection to the message passing algorithms for i.i.d subgaussian ensembles [12][2]. Universality phenomenon is also observed in the general problem (1.4); however, we are not aware of a significant result on this. For GBP, the recent work by Tropp shows that, subgaussian measurements have similar behavior to Gaussian ensemble up to an unknown constant oversampling factor [206] (also see works by Mendelson et al. [129, 152] and Ai et al. [2]). In Chapter 4.1, we investigate the special case of Bernoulli measurement ensemble, where we provide a short argument that states that $7\omega(\mathscr{T}_f(\mathbf{x})\cap\mathscr{B}^{n-1})^2$ Bernoulli measurements are sufficient for the success of GBP.

1.4.6 Noise Analysis

One of the desirable properties of an optimization algorithm is stability to perturbations. The noisy estimation of structured signals has been studied extensively in the recent years. In this case, we get to observe corrupted linear observations of the form $\mathbf{y}=\mathbf{A}\mathbf{x}+\mathbf{z}$. Let us call the recovery stable if the estimate $\hat{\mathbf{x}}$ satisfies $\|\hat{\mathbf{x}}-\mathbf{x}\|_2\leq C\|\mathbf{z}\|_2$ for a positive constant C.

Results on sparse estimation: In case of ℓ_1 minimization, most conditions that ensure recovery of a sparse signal from noiseless observations also ensure stable recovery from noisy observations (e.g., RIP, NSP). Both the lasso estimator (1.6) and the Dantzig selector do a good job at emphasizing a sparse solution while suppressing the noise [19, 31, 201]. For related literature see [19, 31, 45]. There is often a distinction between worst-case noise analysis and the average-case noise analysis.

Assume $\mathbf{A}$ has i.i.d standard normal entries with variance $\frac{1}{m}$ and $m\geq\mathscr{O}\left(k\log\frac{2n}{k}\right)$. It can be shown that, with high probability, $\|\hat{\mathbf{x}}-\mathbf{x}\|_2\leq C\|\mathbf{z}\|_2$ for a positive constant for all vectors $\mathbf{z}$ [32, 43, 45, 50, 225]. This includes the adversarial noise scenario where $\mathbf{z}$ can be arranged (as a function of $\mathbf{A},\mathbf{x}$) to degrade the performance. On the other hand, the average-case reconstruction error $\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2$ ($\mathbf{z}$ is independent of $\mathbf{A}$) is known to scale as $\mathscr{O}\left(\frac{k\log n}{m}\|\mathbf{z}\|_2^2\right)$ [19, 31, 33, 82, 172]. In this case, $\mathbf{z}$ does not have the knowledge of $\mathbf{A}$ and is not allowed to be adversarial.

Average-case noise is what we face in practice and it is an attractive problem to study. The exact error behavior has been studied by Donoho, Maleki, and Montanari [13, 82] in connection to the Approximate Message Passing framework. They are able to find the true error formulas (so-called the noise-sensitivity bounds) which hold in high dimensions, as a function of the sparsity k, m, and the signal-to-noise ratio.

[2] They have several other constraints. For instance, the results are true asymptotically in m,n,k. They also require that the subgaussian distribution has a small Gaussian component.

Results on the generalized basis pursuit: For the general problem, Chandrasekaran et al. have worst-case error bounds, which follows from arguments very similar to the noiseless case [50]. Negahban et al. provide order-optimal convergence rates under a "decomposability" assumption on f [161]. The average-case analysis for arbitrary convex functions and sharp error bounds with correct constants are studied by Oymak et al. in [170, 200] and they are among the contributions of this dissertation.

1.4.7 Low-rank plus sparse decomposition

In several applications, the signal $\mathbf{X}$ can be modeled as the superposition of a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$. From both theory and application motivated reasons, it is important to understand under what conditions we can split $\mathbf{X}$ into $\mathbf{L}$ and $\mathbf{S}$, and whether it can be done in an efficient manner. The initial work on this problem is due to Chandrasekaran et al. where authors derived deterministic conditions under which (1.7) works [51]. These conditions are based on the incoherence between the low-rank and the sparse signal domains. In particular we require the low-rank column-row spaces to not be spiky (diffused entries), and the support of the sparse component to have a diffused singular value spectrum. Independently, Candès et al. studied the problem in a randomized setup where the nonzero support of the sparse component is chosen uniformly at random. Similar to the discussion in Section 1.1, randomness results in better guarantees in terms of the sparsity-rank tradeoff [36]. The problem is also related to the low-rank matrix completion, where we observe few entries of a low-rank matrix corrupted by additive sparse noise [1,36,235]. In Chapter 6, we propose a similar problem formulation for the graph clustering problem.

1.5 Contributions

We will list our contributions in four topics. These are:

- A general theory for noisy linear inverse problems
- Elementary equivalences in compressed sensing
- Recovery of simultaneously structured models
- Graph clustering via convex optimization

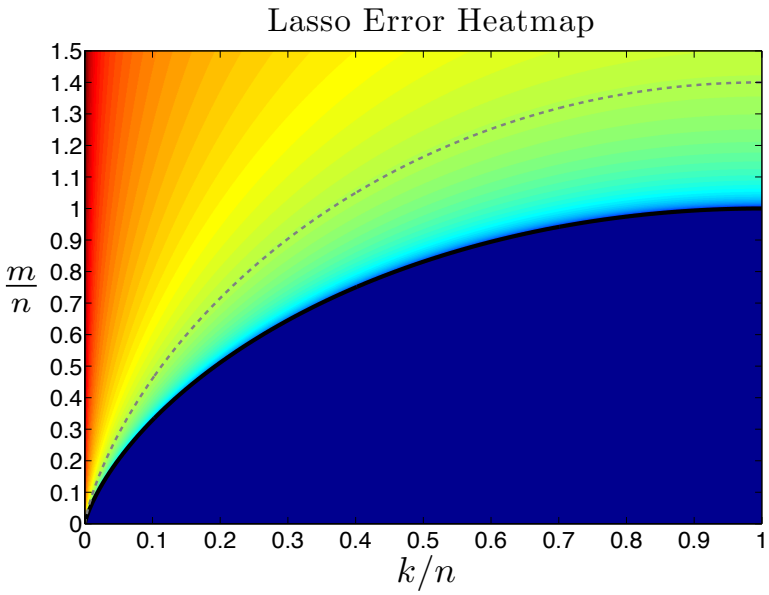


Figure 1.2: We plot (1.11) for sparse recovery. The black line is the Donoho-Tanner bound; above which the recovery is robust. The warmer colors correspond to better reconstruction guarantees; hence, this figure adds an extra dimension to Figure 1.1; which only reflects "success" and "failure". Dashed line corresponds to the fixed reconstruction error. Remark: The heatmap is clipped to enhance the view (due to singularities of (1.11)).

1.5.1 A General Theory of Noisy Linear Inverse Problems

In Chapter 3, we consider the noisy system $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$. We are interested in estimating $\mathbf{x}$ and the normalized estimation error $NSE = \frac{\mathbb{E}\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2}{\mathbb{E}\|\mathbf{z}\|_2^2}$ when $\mathbf{A} \in \mathbb{R}^{m\times n}$ has independent standard normal entries. When $m > n$, the standard approach to solve this overdetermined system of equations is the least squares method. This method is credited to Legendre and Gauss and is approximately 200 years old. The estimate is given by the pseudo-inverse $\hat{\mathbf{x}} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}^T\mathbf{y}$ and it can be shown that the error is approximately $\frac{n}{m-n}$, where $m-n$ corresponds to the statistical degrees of freedom (here the difference between the number of equations and the number of unknowns). When the system is underdetermined ($m < n$), as is the case in many applications, the problem is ill-posed and unique reconstruction is not possible in general. Assuming that the signal has a structure, the standard way to overcome this challenge is to use the lasso formulation (1.6). In Chapter 3, we are able to give precise error formulas as a function of,

- Penalty parameter λ
- Number of measurements m
- Convex structure inducing function $f(\cdot)$
- Noise level z

To give an exposition to our result, let us first consider the following variation of the problem.

$$\min_{\mathbf{x}'} \|\mathbf{y} - \mathbf{A}\mathbf{x}'\|_2 \quad \text{subject to} \quad f(\mathbf{x}') \leq f(\mathbf{x}) \tag{1.10}$$

We prove that the NSE satisfies (with high probability),

$$\frac{\mathbb{E}\,\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2}{\mathbb{E}\,\|\mathbf{z}\|_2^2} \lesssim \frac{\omega(\mathscr{T}_f(\mathbf{x}) \cap \mathscr{B}^{n-1})^2}{m - \omega(\mathscr{T}_f(\mathbf{x}) \cap \mathscr{B}^{n-1})^2}. \tag{1.11}$$

Furthermore, the equality is achieved when signal-to-noise ratio $\frac{\|\mathbf{x}\|_2^2}{\mathbb{E}[\|\mathbf{z}\|_2^2]}$ approaches ∞. We also show that, this is the best possible bound that is based on the knowledge of the first order statistics of the function. Here, first order statistics means knowledge of the subgradients at $\mathbf{x}$, and will become clear in Chapter 3.

From Section 1.4.4, recall that $\omega(\mathscr{T}_f(\mathbf{x}) \cap \mathscr{B}^{n-1})^2$ is the quantity that characterizes the fate of the noiseless problem (1.4). When $\mathbf{x}$ is a k-sparse vector and we use ℓ_1 optimization, (1.11) reduces to the best known error bounds for sparse recovery, namely,

$$\frac{\mathbb{E}\,\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2}{\mathbb{E}\,\|\mathbf{z}\|_2^2} \lesssim \frac{2k \log \frac{n}{k}}{m - 2k \log \frac{n}{k}}.$$

This particular bound was previously studied by Donoho, Maleki, and Montanari under the name "noise sensitivity" [13, 82]. (1.11) has several unifying aspects:

- We generalize the results known for noiseless linear inverse problems. Stable reconstruction is possible if and only if $m > \omega(\mathscr{T}_f(\mathbf{x}) \cap \mathscr{B}^{n-1})^2$.
- Setting $f(\cdot) = 0$ reduces our results to standard least-squares technique where $\omega(\mathscr{T}_f(\mathbf{x}) \cap \mathscr{B}^{n-1})^2 = n$. Hence, we recover the classic result $\frac{n}{m-n}$.
- We provide a generic guarantee for structured recovery problems. Instead of dealing with specific cases such as sparse signals, low-rank matrices, block-sparsity, etc, we are able to handle any abstract norm, and hence treat these specific cases systematically.

Penalized problems: (1.10) requires information about the true signal, namely, $f(\mathbf{x})$. The more useful formulation (1.6) uses penalization instead. In Chapter 3, we also study (1.6) as well as its variation,

$$\min_{\mathbf{x}'} \|\mathbf{y} - \mathbf{A}\mathbf{x}'\|_2 + \lambda f(\mathbf{x}'). \tag{1.12}$$

For penalized problems, we come up with the quantity "Gaussian distance", which is a natural generalization

of Gaussian width. We show that the error bounds for (1.12) are captured by this new term, which can reflect the precise dependence on λ. In addition to error bounds, our results yield the optimal penalty parameters for (1.6) and (1.12), which can achieve the bound (1.11). We defer the detailed analysis and rigorous statements of our results to Chapter 3.

1.5.2 Elementary equivalences in compressed sensing

In Chapter 4, we present two results which are relatively short and are based on short and elementary arguments. These are

- investigating properties of the Bernoulli measurement ensemble via connection to Gaussian ensemble,
- investigating RIP conditions for low-rank recovery via connection to RIP of sparse recovery.

1.5.2.1 Relating the Bernoulli and Gaussian ensembles

So far, we have discussed the importance of the Gaussian ensemble. In particular, one can find the exact performance when the sensing matrix has independent $\mathcal{N}(0,1)$ entries. We have also discussed the universality phenomenon which is partially solved for ℓ_1-minimization. In Chapter 4.1, we consider (1.4) when $\mathbf{A}$ has symmetric Bernoulli entries which are equally likely to be ± 1. To analyze this, we write a Gaussian matrix $\mathbf{G}$ as,

$$\mathbf{G} = \sqrt{\frac{2}{\pi}}\text{sign}(\mathbf{G}) + \mathbf{R}$$

where $\text{sign}(\cdot)$ returns the element-wise signs of the matrix, i.e. $+1$ if $\mathbf{G}_{i,j} \geq 0$ and -1 else. Observe that, this decomposition ensures $\text{sign}(\mathbf{G})$ is identical to a symmetric Bernoulli matrix. Furthermore, $\mathbf{R}$ is a zero-mean matrix conditioned on $\text{sign}(\mathbf{G})$. Based on this decomposition, we show that,

$$m \approx 7\omega(\mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{B}^{n-1})^2$$

Bernoulli samples are sufficient for successful recovery. Compared to the related works [2, 152, 206], our argument is concise and yields reasonably small constants. We also find a deterministic relation between the restricted isometry constants and restricted singular values of Gaussian and Bernoulli matrices. In particular, restricted isometry constant corresponding to a Bernoulli matrix is at most $\frac{\pi}{2}$ times that of Gaussian.

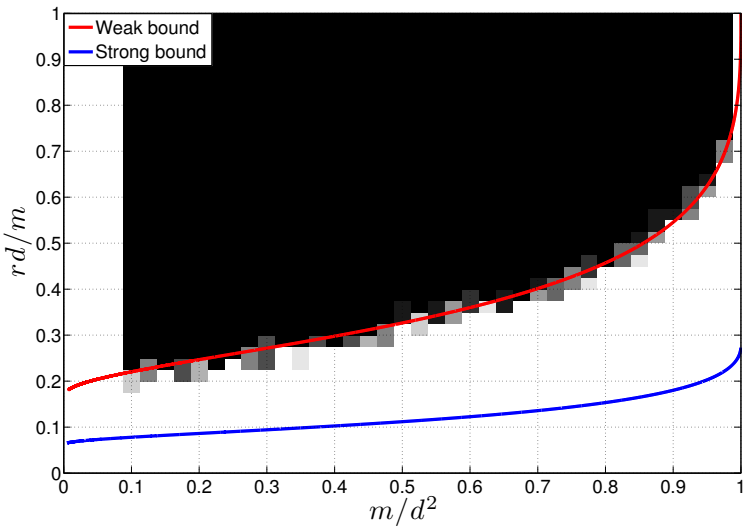


Figure 1.3: To illustrate the linear growth of the phase transition point in rd, we chose the y-axis to be $\frac{m}{rd}$ and the x-axis to be the normalized measurements $\frac{m}{d^2}$. Observe that only weak bound can be simulated and it shows a good match with simulations. The numerical simulations are done for a 40×40 matrix and when $m \geq 0.1d^2$. The dark region implies that (1.13) failed to recover $\mathbf{X}$.

1.5.2.2 Relating the recovery conditions for low-rank and sparse recovery

The tools that are used for low-rank approximation often originates from the sparse approximation techniques. For instance, the initial guarantee of Recht, Fazel and Parrilo [178] is based on a restricted isometry property specialized for low-rank estimation. Similarly, the null space property proposed by [179] also parallels that of sparse estimation. Despite the efforts of [39, 156, 178], the restricted isometry constants for low-rank estimation was weaker than that of the sparse recovery and required a more complicated analysis. Furthermore, the low-rank null space property given in [179] was rather loose and not "if and only if". In Chapter 4.2, we first find a tight null space condition for the success of nuclear norm minimization that compares well with Proposition 1.1. With the help of this result, we establish a framework to "translate" RIP conditions that guarantee sparse recovery (which we call "vector RIP") to RIP conditions that guarantee low-rank recovery (which we call "matrix RIP"). Our eventual result states that if a set of "vector RIP" conditions guarantee sparse recovery, than the equivalent "matrix RIP" conditions can guarantee low-rank recovery. Our results yield immediate improvement over those of [39, 87, 156, 178].

1.5.2.3 Phase transitions for nuclear norm minimization

Related to Section 1.5.2.2 and as mentioned in Section 1.4.3, in [163, 165] we study the sample complexity of the low-rank matrix estimation via nuclear norm minimization[3]. In particular, given a rank r matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$, we are interested in recovering it from $\mathscr{A}(\mathbf{X})$ where $\mathscr{A}(\cdot) : \mathbb{R}^{d \times d} \to \mathbb{R}^m$ is a linear i.i.d Gaussian map, i.e. $\mathscr{A}(\mathbf{X})$ is equivalent to casting $\mathbf{X}$ into a $d^2 \times 1$ vector and multiplying with an $m \times d^2$ matrix with independent standard normal entries. We consider

$$\arg\min_{\mathbf{X}'} \|\mathbf{X}'\|_\star \quad \text{subject to} \quad \mathscr{A}(\mathbf{X}') = \mathscr{A}(\mathbf{X}) \tag{1.13}$$

as the estimator. In [163], we find Donoho-Tanner type precise undersampling bounds for this problem (illustrated in Figure 1.3). As mentioned in Section 1.4.2, the weak bound asks for the recovery of a particular rank r matrix, while the strong bound asks for the recovery of all rank r matrices simultaneously by the same realization of $\mathscr{A}$. We showed that when m is above the bounds (1.13) will succeed. However, the recent progress on phase transitions strongly indicates that, our "weak threshold" is indeed tight [4, 77, 194]. In [165], we provide a robustness analysis of (1.13) and also find closed-form bounds:

- To ensure weak recovery one needs $m \gtrsim 6rd$ samples.
- To ensure strong recovery one needs $m \gtrsim 16rd$ samples.

Considering that a rank r matrix has $2dr - r^2$ degrees of freedom [40], the weak bound suggests that, one needs to oversample $\mathbf{X}$ by only a factor of 3 to efficiently recover it from underdetermined observations.

1.5.3 Simultaneously structured signals

Most of the attention in the research community has focused on signals that exhibit a single structure, such as variations of sparsity and low-rank matrices. However, signals exhibiting multiple low-dimensional structures do show up in certain applications. For instance, in certain applications, we wish to encourage a solution which is not only sparse but whose entries also vary slowly, i.e., the gradient of the signal is approximately sparse as well (recall (1.3)). Tibshirani proposed fused lasso optimization for this task [202],

$$\arg\min \lambda_{\ell_1} \|\mathbf{x}'\|_1 + \lambda_{TV} \|\mathbf{x}'\|_{TV} + \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}'\|_2^2.$$

Fused lasso aims to encourage a solution with two structures: sparsity and sparsity of the gradient.

[3] Due to time and space limitations, the technical details are not included in the dissertation.

As a second example, we will talk about matrices that are simultaneously sparse and low-rank. Such matrices typically arise from sparse vectors, with the mapping $\mathbf{a} \to \mathbf{A} = \mathbf{a}\mathbf{a}^T$, where $\mathbf{a}$ is a sparse vector itself. This mapping is also known as lifting as we move from lower dimensional $\mathbb{R}^n$ to a higher dimensional space $\mathbb{R}^{n \times n}$. The new variable $\mathbf{A}$ has rank 1 and is also sparse. This model shows up in several applications, including sparse principal component analysis [236] and sparse phase retrieval [44, 92, 122].

Example: Sparse Phase Retrieval. Phase retrieval is the problem of estimating a signal from its phaseless observations. For instance, assume we get to observe power spectral density (PSD) of a signal. PSD is obtained by taking the square of the Fourier Transform and is phaseless. In certain applications, most notably X-Ray crystallography, one only has access to PSD information and the task is to find a solution to the given PSD. There are multiple signals that can yield the same power spectral density, and hence we often need an additional criteria to optimize over. This criteria is often the sparsity of the underlying signal, which yields the problem "sparse phase retrieval". In finite dimensions, the noiseless phase retrieval problem assumes phaseless measurements $\mathbf{y}_i = \{|\mathbf{a}_i^T \mathbf{x}|^2\}_{i=1}^m$ of the true vector $\mathbf{x}$, where $\{\mathbf{a}_i\}_{i=1}^m$ are the measurement vectors. Hence, sparse PR problem is given as

$$\|\mathbf{x}'\|_0 \quad \text{subject to} \quad |\mathbf{a}_i^T \mathbf{x}'|^2 = \mathbf{y}_i.$$

This problem is nontrivial due to the quadratic equality constraints as well as the combinatorial objective. Applying the lifting $\mathbf{x} \to \mathbf{X} = \mathbf{x}\mathbf{x}^T$ proposed by Balan et al. [9], we end up with linear measurements in the new variable $\mathbf{X}$,

$$\|\mathbf{X}'\|_0 \quad \text{subject to} \quad \left\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X}' \right\rangle = \mathbf{y}_i, \ \text{rank}(\mathbf{X}') = 1, \ \mathbf{X}' \succeq 0.$$

Relaxing the sparsity and rank constraints by using ℓ_1 and nuclear norm, we find a convex formulation that encourages a low-rank and sparse solution.

$$\|\mathbf{X}'\|_1 + \lambda \|\mathbf{X}'\|_\star \quad \text{subject to} \quad \mathbf{X}' \succeq 0, \ \left\langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X}' \right\rangle = \mathbf{y}_i, \ 1 \le i \le m. \tag{1.14}$$

Traditional convex formulations for the sparse PCA problem have a striking similarity to (1.14) and also make use of nuclear norm and ℓ_1 norm [64].

Finally, we remark that low-rank tensors are yet another example of simultaneously structured signals [103, 158] and they will be discussed in more detail in Chapter 5.

Our contributions: Convex relaxation is a powerful tool because it often yields almost-optimal perfor-

mance guarantees, i.e., we don't lose much by solving the relaxed objective compared to the true objective. For example, countless papers in literature show that ℓ_1 norm does a great job in encouraging sparsity. In (1.14), the signal has two structures, and hence it has far fewer degrees of freedom compared to an only low-rank or only sparse matrix. We investigate whether it is possible to do better (i.e., use fewer measurements) by making use of this fact. We show that the answer is negative for any cost function that combines the ℓ_1 and nuclear norms. To be more precise, by combining convex penalties, one cannot reduce the number of measurements much beyond what is needed for the best performing individual penalty (ℓ_1 or nuclear norm). In Chapter 5, we will study the problem for abstract signals that have multiple structures and arrive at a more general theory that can be applied to the specific signal types.

Our results are easy to interpret and apply to a wide range of measurement ensembles. In particular, we show the limitations of standard convex relaxations for the sparse phase retrieval and the low-rank tensor completion problems. For the latter one, each observation is a randomly chosen entry of the tensor. To give a flavor of our results, let us return to the sparse and low-rank matrices, where we investigate,

$$\|\mathbf{X}'\|_1 + \lambda \|\mathbf{X}'\|_\star \quad \text{subject to} \quad \mathbf{X}' \succeq 0, \ \left\langle \mathbf{G}_i, \mathbf{X}' \right\rangle = \mathbf{y}_i, \ 1 \leq i \leq m. \tag{1.15}$$

For simplicity, let $\{\mathbf{G}_i\}_{i=1}^m$ be matrices with independent standard normal entries and $\mathscr{A}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is the measurement operator. Suppose the true signal is $\mathbf{X} = \mathbf{x}\mathbf{x}^T$, where $\mathbf{x} \in \mathbb{R}^n$ is a k-sparse vector. To simplify the consequent notation, let us assume that the nonzero entries of $\mathbf{x}$ are comparable, i.e., $\frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} \approx \mathscr{O}\left(\sqrt{k}\right)$. Then, we have the following,

- If only ℓ_1 norm is used ($\lambda = 0$), CS theory requires $\mathscr{O}\left(k^2 \log \frac{n}{k}\right)$ measurements to retrieve $\mathbf{X}$ via (1.15).
- If only nuclear norm is used, results on rank minimization requires $\mathscr{O}(n)$ measurements to retrieve $\mathbf{X}$.
- Our results in Chapter 5 ensure that one needs at least $\Omega(\min\{k^2, n\})$ measurements for any choice of λ.
- Replace the objective function in (1.15) by $\|\mathbf{X}'\|_0 + \lambda \operatorname{rank}(\mathbf{X}')$. There is a suitable λ, for which $\mathscr{O}\left(k \log \frac{n}{k}\right)$ measurements are sufficient.

There is a significant gap between what can be done by the optimally tuned convex approach ($\min\{k^2, n\}$) and the non-convex objective ($k \log \frac{n}{k}$). On the other hand, the optimally tuned convex approach is not better (in order) than using only the best of the ℓ_1 norm and nuclear norm.

1.5.4 Convex Optimization for Graph Clustering

Our results so far focused on the study of linear inverse problems. In Chapter 6, we consider a more application-oriented problem, namely, graph clustering. Graphs are important tools to represent data efficiently. An important task regarding graphs is to partition the nodes into groups that are densely connected. In the simpler case, we may wish to find the largest clique in the graph, i.e., the largest group of nodes that are fully connected to each other. In a similar flavor to the sparse estimation problem, the clustering problem is challenging and highly combinatorial. Can this problem be cast in the convex optimization framework? It turns out the answer is positive. We pose the clustering problem as a demixing problem where we wish to decompose the adjacency matrix of the graph into a sparse and low-rank component. We then formulate two optimizations, one of which is tailored for sparse graphs. Remarkably, performance of convex optimization is on par with alternative state-of-the-art algorithms [6, 7, 146, 164] (also see Chapter 6). We carefully analyze these formulations and obtain intuitive quantities, dubbed "effective density", that sharply capture the performance of the proposed algorithms in terms of cluster sizes and densities.

Let the graph have n nodes, and $\mathbf{A}$ be the adjacency matrix, where 1 corresponds to an edge between two nodes and 0 corresponds to no edge. $\mathbf{A}$ is a symmetric matrix, and, without loss of generality, assume diagonal entries are 1. Observe that a clique corresponds to a submatrix of all 1's. The rank of this submatrix is simply 1. With this observation, Ames and Vavasis [6] proposed to find cliques via rank minimization, and used nuclear norm as a convex surrogate of rank. We cast the clustering problem in a similar manner to the clique finding problem, where cliques can have imperfections in the form of missing edges. Assuming there are few missing edges, each cluster corresponds to the sum of a rank 1 matrix and a sparse matrix.

Let us assume the clusters are disjoint. Our first method (dubbed simple method) solves the following,

$$\begin{aligned} &\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \quad \|\mathbf{L}\|_\star + \lambda \|\mathbf{S}\|_1 \\ &\text{subject to} \\ &\qquad 1 \geq \mathbf{L}_{i,j} \geq 0 \quad \text{for all } i,j \in \{1,2,\dots n\} \\ &\qquad \mathbf{L} + \mathbf{S} = \mathbf{A}. \end{aligned} \tag{1.16}$$

The nuclear norm and the ℓ_1 norm are used to induce a low-rank $\mathbf{L}$ and sparse $\mathbf{S}$, respectively. The hope is that $\mathbf{L}_{i,j}$ will be 1 whenever nodes i and j will lie in the same cluster, and 0 otherwise. This way, rank$(\mathbf{L})$ will be equal to the number of clusters.

We investigate this problem for the well-known stochastic block model [117], which is essentially a nonuniform Erdös-Renyi graph. While Chapter 6 considers a general model, let us introduce the following setup for the exposition.

Definition 1.6 (Simple Block Model) *Assume that $\mathscr{G}$ is a random graph with n nodes with t clusters, where each cluster has size d. Let $\mathbf{A}$ be the corresponding adjacency matrix. Further assume that the existence of each edge is independent of each other, and*

$$\mathbb{P}(\mathbf{A}_{i,j} = 1) = \begin{cases} p & \text{if } i,j \text{ is in the same cluster} \\ q & \text{else} \end{cases}$$

for some constants $p > q > 0$.

Assuming d, n are large, $d = o(n)$ and λ is well-tuned, for the simple block model, we find that,

- (1.16) correctly identifies the planted clusters if $q < \frac{1}{2}$ and $d(2p-1) > 4\sqrt{q(1-q)n}$.
- (1.16) fails if $d(2p-1) < \sqrt{qn}$ or $q > \frac{1}{2}$.

Here, $d(2p-1)$ jointly captures the size and density of the planted clusters; hence, we call it effective density. Assuming $q < \frac{1}{2}$, $d(2p-1)$ is tightly sandwiched between $\sqrt{qn}$ and $4\sqrt{qn}$. This result indicates that cluster sizes should grow with $\sqrt{n}$, which is consistent with state of the art results on clique finding [6, 7, 61, 67, 185].

The simple method has critical drawbacks, as we require $p > \frac{1}{2} > q$. The algorithm tailored for sparse graphs (dubbed "improved method") is able to overcome this bottleneck. This new algorithm only requires $p > q$ and will succeed when

$$d(p-q) > 2\sqrt{q(1-q)n}.$$

This comes at the cost of requiring additional information about the cluster sizes, however, this new bound is highly consistent with the existing results on very sparse graphs [146]. Unlike the comparable results in the literature [3, 5–7, 61, 62], our bounds are not only order optimal, but also have small constants that show a good match with numerical simulations. We defer the extensive discussion and further literature survey to Chapter 6.

1.5.5 Organization

In Chapter 2, we go over the mathematical notation and tools that will be crucial to our discussion throughout this dissertation.

In Chapter 3, we study the error bounds for the noisy linear inverse problems. We formulate three versions of the lasso optimization and find formulas that accurately capture the behavior based on the summary parameters Gaussian width and Gaussian distance.

Chapter 4 will involve two topics. We first analyze Bernoulli measurement ensemble via connection to the Gaussian measurement ensemble. Next, we study the low-rank approximation problem by establishing a relation between the sparse and low-rank recovery conditions.

Chapter 5 is dedicated to the study of simultaneously structured signals. We formulate intuitive convex relaxations for recovery of these signals and show that there is a significant gap between the performance of convex approaches and what is possible information theoretically.

In Chapter 6, we develop convex relaxations for the graph clustering problem. We formulate two approaches, where one is particularly tailored for sparse graphs. We find intuitive parameters based on the density and size of the clusters that sharply characterize the performance of these formulations. We also numerically show that performance is on par with more traditional algorithms.

In Chapter 7, we discuss the related open problems and possible extensions to our results that are left to be researched.

Chapter 2

Preliminaries

We will now introduce some notation and definitions that will be used throughout the dissertation.

2.1 Notation

Vectors and matrices: Vectors will be denoted by bold lower case letters. Given a vector $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^T$ will be used to denote its transpose. For $p \geq 1$, ℓ_p norm of $\mathbf{a}$ will be denoted by $\|\mathbf{a}\|_p$ and is equal to $(\sum_{i=1}^n |\mathbf{a}_i|^p)^{1/p}$. The ℓ_0 quasi-norm returns the number of nonzero entries of the vector and will be denoted by $\|\mathbf{a}\|_0$. For a scalar a, $\operatorname{sgn}(a)$ returns its sign i.e. $a \cdot \operatorname{sgn}(a) = |a|$ and $\operatorname{sgn}(0) = 0$. $\operatorname{sgn}(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ returns a vector consisting of the signs of the entries of the input.

Matrices will be denoted by bold upper case letters. $n \times n$ identity matrix will be denoted by $\mathbf{I}_n$. For a given matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, its null space and range space will be denoted by $\text{Null}(\mathbf{A})$ and $\text{Range}(\mathbf{A})$, respectively. To vectorize a matrix, we can stack its columns on top of each other to obtain the vector $\text{vec}(\mathbf{A}) \in \mathbb{R}^{n_1 n_2}$. $\text{rank}(\mathbf{A})$ will denote the rank of the matrix. The minimum and maximum singular values of a matrix $\mathbf{A}$ are denoted by $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$. $\sigma_{\max}(\mathbf{A})$ is equal to the spectral norm $\|\mathbf{A}\|$. $\|\mathbf{A}\|_F$ will denote the Frobenius norm. This is essentially equivalent to the ℓ_2 norm of the vectorization of a matrix: $\|\mathbf{A}\|_F^2 = (\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbf{A}_{i,j}^2)^{1/2}$. $\operatorname{tr}(\cdot)$ will return the trace of a matrix.

Probability: We will use $\mathbb{P}(\cdot)$ to denote the probability of an event. $\mathbb{E}[\cdot]$ and $\text{Var}[\cdot]$ will be the expectation and variance operators, respectively. Gaussian random variables will play a critical role in our results. A multivariate (or scalar) normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ will be denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. "Independent and identically distributed" and "with high probability" will be abbreviated as i.i.d and w.h.p respectively. The symbol "$\sim$" should be read "is distributed as". There may be additional definitions specific to the individual chapters and they will be introduced accordingly.

Convex geometry: Denote the unit ℓ_2 sphere and the unit ℓ_2 ball in $\mathbb{R}^n$ by $\mathscr{S}^{n-1}$ and $\mathscr{B}^{n-1}$, respectively. $\dim(\cdot)$ will return the dimension of a linear subspace. For convex functions, the subgradient will play a critical role. $\mathbf{s}$ is a subgradient of $f(\cdot)$ at the point $\mathbf{v}$, if for all vectors $\mathbf{w}$, we have that,

$$f(\mathbf{v}+\mathbf{w}) \geq f(\mathbf{v}) + \mathbf{s}^T\mathbf{w}.$$

The set of all subgradients $\mathbf{s}$ is called the subdifferential and is denoted by $\partial f(\mathbf{v})$. If $f:\mathbb{R}^n \to \mathbb{R}$ is continuous at $\mathbf{v}$, then the subdifferential $\partial f(\mathbf{v})$ is a convex and compact set [23]. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|$ denotes a general norm and $\|\mathbf{x}\|^* = \sup_{\|z\|\leq 1} \langle \mathbf{x}, \mathbf{z}\rangle$ is the corresponding dual norm. Finally, the Lipschitz constant of the function $f(\cdot)$ is a number L so that, for all $\mathbf{v}, \mathbf{u}$, we have that, $|f(\mathbf{v}) - f(\mathbf{u})| \leq L\|\mathbf{v}-\mathbf{u}\|_2$.

Sets: Given sets $S_1, S_2 \in \mathbb{R}^n$, $S_1 + S_2$ will be the Minkowski sum of these sets, i.e., $\{\mathbf{v}_1 + \mathbf{v}_2 \big| \mathbf{v}_1 \in S_1,\ \mathbf{v}_2 \in S_2\}$. Closure of a set is denoted by $\mathrm{Cl}(\cdot)$. For a scalar $\lambda \in \mathbb{R}$ and a nonempty set $S \subset \mathbb{R}^n$, λS will be the dilated set, and is equal to $\{\lambda \mathbf{v} \in \mathbb{R}^n \ \big|\ \mathbf{v} \in S\}$. The cone induced by the set S will be denoted by $\mathrm{cone}(S)$ and is equal to $\{\lambda \mathbf{v} \big| \lambda \geq 0,\ \mathbf{v} \in S\}$. It can also be written as a union of dilated sets as follows,

$$\mathrm{cone}(S) = \bigcup_{\lambda \geq 0} \lambda S.$$

The polar cone of S is defined as $S^\circ = \{\mathbf{v} \in \mathbb{R}^n \ \big|\ \mathbf{v}^T\mathbf{u} \leq 0\ \forall\ \mathbf{u} \in S\}$. The dual cone is $S^* = -S^\circ$.

2.2 Projection and Distance

Given a point $\mathbf{v}$ and a closed and convex set $\mathscr{C}$, there is a unique point $\mathbf{a}$ in $\mathscr{C}$ satisfying $\mathbf{a} = \arg\min_{\mathbf{a}' \in \mathscr{C}} \|\mathbf{v} - \mathbf{a}'\|_2$. This point is the projection of $\mathbf{v}$ onto $\mathscr{C}$ and will be denoted as $\mathrm{Proj}_{\mathscr{C}}(\mathbf{v})$ or $\mathrm{Proj}(\mathbf{v}, \mathscr{C})$. The distance vector will be denoted by $\Pi_{\mathscr{C}}(\mathbf{v}) = \mathbf{v} - \mathrm{Proj}_{\mathscr{C}}(\mathbf{v})$. The distance to a set is naturally induced by the definition of the projection. We will let $\mathrm{dist}_{\mathscr{C}}(\mathbf{v}) = \|\mathbf{v} - \mathrm{Proj}_{\mathscr{C}}(\mathbf{v})\|_2$.

When $\mathscr{C}$ is a closed and convex cone, we have the following useful identity due to Moreau [157].

Fact 2.1 (Moreau's decomposition theorem) *Let $\mathscr{C}$ be a closed and convex cone in $\mathbb{R}^n$. For any $\mathbf{v} \in \mathbb{R}^n$, the following two are equivalent:*

1. $\mathbf{v} = \mathbf{a} + \mathbf{b}$, $\mathbf{a} \in \mathscr{C}, \mathbf{b} \in \mathscr{C}^\circ$ *and* $\mathbf{a}^T\mathbf{b} = 0$.
2. $\mathbf{a} = \mathit{Proj}(\mathbf{v}, \mathscr{C})$ *and* $\mathbf{b} = \mathit{Proj}(\mathbf{v}, \mathscr{C}^\circ)$.

Fact 2.2 (Properties of the projection, [17, 23]) *Assume $\mathscr{C} \subseteq \mathbb{R}^n$ is a nonempty, closed, and convex set and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are arbitrary points. Then,*

- *The projection $Proj(\mathbf{a}, \mathscr{C})$ is the unique vector satisfying, $Proj(\mathbf{a}, \mathscr{C}) = \arg\min_{\mathbf{v} \in \mathscr{C}} \|\mathbf{a} - \mathbf{v}\|_2$.*
- *$\langle Proj(\mathbf{a}, \mathscr{C}), \mathbf{a} - Proj(\mathbf{a}, \mathscr{C}) \rangle = \sup_{\mathbf{s} \in \mathscr{C}} \langle \mathbf{s}, \mathbf{a} - Proj(\mathbf{a}, \mathscr{C}) \rangle$.*
- *$\|Proj(\mathbf{a}, \mathscr{C}) - Proj(\mathbf{b}, \mathscr{C})\|_2 \leq \|\mathbf{a} - \mathbf{b}\|_2$.*

Descent set and tangent cone: Given a function $f(\cdot)$ and a point $\mathbf{x}$, descent set is denoted by $\mathscr{D}_f(\mathbf{x})$, and is defined as

$$\mathscr{D}_f(\mathbf{x}) = \{\mathbf{w} \in \mathbb{R}^n \mid f(\mathbf{x} + \mathbf{w}) \leq f(\mathbf{x})\}.$$

We also define the tangent cone of f at $\mathbf{x}$ as $\mathscr{T}_f(\mathbf{x}) := \mathrm{Cl}(\mathrm{cone}(\mathscr{D}_f(\mathbf{x})))$. In words, tangent cone is the closure of the conic hull of the descent set. These concepts will be quite important in our analysis. Recall that Proposition 1.2 is based on the descent cone and is essentially the null-space property for the GBP. The tangent cone is related to the subdifferential $\partial f(\mathbf{x})$ as follows [182].

Proposition 2.1 *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and continuous function. The tangent cone has the following properties.*

- *$\mathscr{D}_f(\mathbf{x})$ is a convex set and $\mathscr{T}_f(\mathbf{x})$ is a closed and convex cone.*
- *Suppose $\mathbf{x}$ is not a minimizer of $f(\cdot)$. Then, $\mathscr{T}_f(\mathbf{x})^\circ = \mathrm{cone}(\partial f(\mathbf{x}))$.*

2.2.1 Subdifferential of structure inducing functions

Let us introduce the subdifferential of the ℓ_1 norm.

Proposition 2.2 ([72]) *Given $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{s} \in \partial \|\mathbf{x}\|_1$ if and only if, $\mathbf{s}_i = \mathrm{sgn}(\mathbf{x}_i)$ for all $i \in \mathrm{supp}(\mathbf{x})$ and $|\mathbf{s}_i| \leq 1$ else.*

Related to this, we define the soft-thresholding (a.k.a. shrinkage) operator.

Definition 2.1 (Shrinkage) *Given $\lambda \geq 0$, $shrink_\lambda(\cdot) : \mathbb{R} \to \mathbb{R}$ is defined as,*

$$shrink_\lambda(x) = \begin{cases} x - \lambda & \text{if } x \geq \lambda \\ x + \lambda & \text{if } x \leq -\lambda \\ 0 & \text{if } |x| < \lambda \end{cases}$$

While the importance of shrinkage will become clear later on in the simplest setup, it shows up when one considers denoising by ℓ_1 minimization [72].

Proposition 2.3 *Suppose we see the noisy observations* $\mathbf{y} = \mathbf{x} + \mathbf{z}$ *and we wish to estimate* $\mathbf{x}$ *via* $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}'} \lambda \|\mathbf{x}'\|_1 + \frac{1}{2}\|\mathbf{y} - \mathbf{x}'\|_2^2$. *The solution* $\hat{\mathbf{x}}$ *is given by*

$$\hat{\mathbf{x}}_i = shrink_\lambda(\mathbf{y}_i) \quad for \quad 1 \leq i \leq n.$$

This in turn is related to the distance of a vector to the scaled subdifferential, which will play an important role in Chapter 3. It follows from Proposition 2.2 that $\Pi(\mathbf{v}, \lambda\partial\|\mathbf{x}\|_1)_i = \mathbf{v}_i - \lambda \operatorname{sgn}(\mathbf{x}_i)$ when $i \in \operatorname{supp}(\mathbf{x}_i)$ and $\operatorname{shrink}(\mathbf{v}_i)$ when $i \notin \operatorname{supp}(\mathbf{x}_i)$.

We will also discuss that subdifferentials of $\ell_{1,2}$ norm and the nuclear norm have very similar forms to that of ℓ_1 norm, and the distance to the subdifferential can again be characterized by the shrinkage operator. For instance, the nuclear norm is associated with shrinking the singular values of the matrix while the $\ell_{1,2}$ norm is associated with shrinking the ℓ_2 norms of the individual blocks [50, 70, 77].

2.3 Gaussian width, Statistical dimension and Gaussian distance

We have discussed Gaussian width (Def. 1.5) in Chapter 1.4 and its importance in dimension reduction via Gaussian measurements. We now introduce two related quantities that will be useful for characterizing how well one can estimate a signal by using a structure inducing convex function[1]. The first one is the Gaussian squared-distance.

Definition 2.2 (Gaussian squared-distance) *Let* $S \in \mathbb{R}^n$ *be a subset of* $\mathbb{R}^n$ *and* $\mathbf{g} \in \mathbb{R}^n$ *have independent* $\mathcal{N}(0,1)$ *entries. Define the Gaussian squared-distance of* S *to be,*

$$\mathbf{D}(S) = \mathbb{E}[\inf_{\mathbf{v} \in S} \|\mathbf{g} - \mathbf{v}\|_2^2]$$

Definition 2.3 (Statistical dimension) *Let* $\mathscr{C} \in \mathbb{R}^n$ *be a closed and convex cone and* $\mathbf{g} \in \mathbb{R}^n$ *have independent* $\mathcal{N}(0,1)$ *entries. Define the statistical dimension of* $\mathscr{C}$ *to be,*

$$\delta(\mathscr{C}) = \mathbb{E}[\|Proj(\mathbf{g}, \mathscr{C})\|_2^2]$$

[1] Another closely related quantity is the "mean width" of [172]

	k-**sparse,** $\mathbf{x}_0 \in \mathbb{R}^n$	**Rank** r**,** $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$	k-**block sparse,** $\mathbf{x}_0 \in \mathbb{R}^{tb}$
$\delta(\mathscr{T}_f(\mathbf{x}_0))$	$2k(\log \frac{n}{k}+1)$	$6dr$	$4k(\log \frac{t}{k}+b)$
$\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$	$(\lambda^2+3)k$ for $\lambda \geq \sqrt{2\log \frac{n}{k}}$	$\lambda^2 r+2d(r+1)$ for $\lambda \geq 2\sqrt{d}$	$(\lambda^2+b+2)k$ for $\lambda \geq \sqrt{b}+\sqrt{2\log \frac{t}{k}}$

Table 2.1: Closed form upper bounds for $\delta(\mathscr{T}_f(\mathbf{x}_0))$ ([50, 101]) and $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ corresponding to sparse, block-sparse signals and low-rank matrices described in Section 1.2.1. See Section A.7 for the proofs.

The next proposition provides some basic relations between these quantities.

Proposition 2.4 *Let $C \in \mathbb{R}^n$ be a closed and convex set, $\mathscr{C} = \text{cone}(C)$ and $\mathbf{g} \in \mathbb{R}^n$ have independent $\mathcal{N}(0,1)$ entries. Then,*

- $\omega(\mathscr{C}\cap\mathscr{B}^{n-1})^2 \leq \delta(\mathscr{C}) \leq \omega(\mathscr{C}\cap\mathscr{B}^{n-1})^2+1$.
- $\mathbf{D}(C) \geq \mathbf{D}(\mathscr{C}) = \delta(\mathscr{C}^\circ)$.

Proof: Let us first show that for any vector $\mathbf{a}$,

$$\|\text{Proj}(\mathbf{a},\mathscr{C})\|_2 = \sup_{\mathbf{v}\in\mathscr{C}\cap\mathscr{B}^{n-1}} \mathbf{v}^T\mathbf{a}.$$

Using Moreau's decomposition, $\mathbf{a} = \text{Proj}(\mathbf{a},\mathscr{C}) + \text{Proj}(\mathbf{a},\mathscr{C}^\circ)$. For any unit length $\mathbf{v} \in \mathscr{C}$, we have, $\mathbf{a}^T\mathbf{v} \leq \langle \text{Proj}(\mathbf{a},\mathscr{C}),\mathbf{v}\rangle \leq \|\text{Proj}(\mathbf{a},\mathscr{C})\|_2$ hence right hand side is less than or equal to the left hand side. On the other hand, the equality can be achieved by choosing $\mathbf{v}=0$ if $\text{Proj}(\mathbf{a},\mathscr{C})=0$ and $\mathbf{v} = \frac{\text{Proj}(\mathbf{a},\mathscr{C})}{\|\text{Proj}(\mathbf{a},\mathscr{C})\|_2}$ else.

With this observe that, $\omega(\mathscr{C}\cap\mathscr{B}^{n-1}) = \mathbb{E}[\|\text{Proj}(\mathbf{g},\mathscr{C})\|_2]$. From Jensen's inequality $\mathbb{E}[\|\text{Proj}(\mathbf{g},\mathscr{C})\|_2]^2 \leq \mathbb{E}[\|\text{Proj}(\mathbf{g},\mathscr{C})\|_2^2]$, hence $\omega(\mathscr{C}\cap\mathscr{B}^{n-1})^2 \leq \delta(\mathscr{C})$. On the other hand, since $\|\text{Proj}(\mathbf{g},\mathscr{C})\|_2$ is 1-Lipschitz function of $\mathbf{g}$, using Fact 2.3, $\delta(\mathscr{C}) \leq \omega(\mathscr{C}\cap\mathscr{B}^{n-1})^2+1$. $\mathbf{D}(C) \geq \mathbf{D}(\mathscr{C})$ as $C \subseteq \mathscr{C}$ hence, the distance to C is greater than or equal to the distance to $\mathscr{C}$. Finally, using Fact 2.1 again, $\text{dist}(\mathbf{v},\mathscr{C}) = \|\mathbf{v}-\text{Proj}(\mathbf{v},\mathscr{C})\|_2 = \|\text{Proj}(\mathbf{v},\mathscr{C}^\circ)\|_2$, hence $\mathbb{E}[\|\text{Proj}(\mathbf{v},\mathscr{C}^\circ)\|_2^2] = \mathbb{E}[\text{dist}(\mathbf{v},\mathscr{C})^2]$. ■

Gaussian distance is particularly beneficial when one is dealing with weighted combination of functions involving Gaussians. In such setups, Gaussian distance may be helpful in representing the outcome of the problem as a function of penalty parameters. The next section will provide an example of this.

2.4 Denoising via proximal operator

Let us focus on the basic estimation problem, where the aim is to estimate $\mathbf{x}_0$ from $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$. One can use the proximal operator to accomplish this task in a similar manner to (1.4) [157].

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \sigma\lambda f(\mathbf{x}) \tag{2.1}$$

An important question is the estimation error which can be defined as $\mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2]$. When $\mathbf{z}$ has independent $\mathcal{N}(0, \sigma^2)$ entries, the following proposition shows that, the error is related to the Gaussian distance. The reader is referred to [18, 49, 70, 167] for more details.

Proposition 2.5 *Suppose $\lambda > 0$, $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuous and $\mathbf{y} = \mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^n$ where $\mathbf{z}$ has independent $\mathcal{N}(0, \sigma^2)$ entries. Then,*

$$\frac{\mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2^2]}{\sigma^2} \leq \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \tag{2.2}$$

Proof: From the Karush-Kuhn-Tucker optimality conditions, there exists $\mathbf{s} \in \partial f(\hat{\mathbf{x}})$ such that $\mathbf{y} = \hat{\mathbf{x}} + \lambda \mathbf{s}$. Using the change of variable $\hat{\mathbf{w}} = \hat{\mathbf{x}} - \mathbf{x}_0$, equivalently,

$$\mathbf{z} = \hat{\mathbf{w}} + \sigma\lambda \mathbf{s} \tag{2.3}$$

Now, choose $\mathbf{s}_0 \in \partial f(\mathbf{x}_0)$ to be $\mathbf{s}_0 = \frac{1}{\sigma\lambda}\text{Proj}(\mathbf{z}, \sigma\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{w}_0 = \mathbf{z} - \sigma\lambda\mathbf{s}_0$. We will first explore the relation between $\mathbf{w}_0$ and $\hat{\mathbf{w}}$. The following inequality follows from the definition of the subgradient.

$$\hat{\mathbf{w}}^T\mathbf{s} \geq f(\mathbf{x}_0 + \hat{\mathbf{w}}) - f(\mathbf{x}_0) \geq \hat{\mathbf{w}}^T\mathbf{s}_0 \implies \hat{\mathbf{w}}^T(\mathbf{s} - \mathbf{s}_0) \geq 0 \tag{2.4}$$

From (2.3) and (2.4), we will conclude that $\|\hat{\mathbf{w}}\|_2 \leq \|\mathbf{w}_0\|_2$. These are equivalent to:

$$\langle \hat{\mathbf{w}}, (\mathbf{z} - \hat{\mathbf{w}}) - (\mathbf{z} - \mathbf{w}_0)\rangle = \langle \hat{\mathbf{w}}, \mathbf{w}_0 - \hat{\mathbf{w}}\rangle \geq 0 \implies \|\hat{\mathbf{w}}\|_2^2 \leq \langle \hat{\mathbf{w}}, \mathbf{w}_0\rangle \leq \|\hat{\mathbf{w}}\|_2\|\mathbf{w}_0\|_2$$

Hence, we find $\|\mathbf{w}_0\|_2 = \text{dist}(\mathbf{z}, \sigma\lambda\partial f(\mathbf{x}_0)) \geq \|\hat{\mathbf{w}}\|_2$. $\text{dist}(\mathbf{z}, \sigma\lambda\partial f(\mathbf{x}_0)) = \sigma\text{dist}(\frac{\mathbf{z}}{\sigma}, \lambda\partial f(\mathbf{x}_0))$ where $\frac{\mathbf{z}}{\sigma}$ has $\mathcal{N}(0, 1)$ entries. Taking the square and then expectation, we can conclude. ■

It is important to understand what happens when λ is optimally tuned to minimize the upper bound on the error. This has been investigated by [4, 101, 167]. We will state the result from [4] which requires less

assumption.

Proposition 2.6 (Theorem 4.3 **of [4])** *Suppose $f:\mathbb{R}^n\to\mathbb{R}$ is a norm and $\mathbf{x}_0\in\mathbb{R}^n$ is nonzero. Then,*

$$\delta(\mathscr{T}_f(\mathbf{x}_0))\leq \min_{\lambda\geq 0}\mathbf{D}(\lambda\partial f(\mathbf{x}_0))\leq \delta(\mathscr{T}_f(\mathbf{x}_0))+2\frac{\sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)}\|\mathbf{s}\|_2}{f(\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2})} \tag{2.5}$$

Here, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)),\delta(\mathscr{T}_f(\mathbf{x}_0))$ are solely based on the subdifferential, hence, as long as $\partial f(\mathbf{x}_0)$ remains same, one can change $\mathbf{x}_0$ to obtain a tighter bound by making $f(\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2})$ larger.

The left hand side of (2.5) is clear from Proposition 2.4 when we use the fact that $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))\subseteq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))=\mathscr{T}_f(\mathbf{x}_0)^\circ$. Let us consider the right-hand side, for a k-sparse $\mathbf{x}_0\in\mathbb{R}^n$. Letting $\mathbf{x}_0\to \text{sgn}(\mathbf{x}_0)$ do not change the subdifferential and results in $2\frac{\sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)}\|\mathbf{s}\|_2}{f(\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2})}=2\sqrt{\frac{n}{k}}$ whereas $\delta(\mathscr{T}_f(\mathbf{x}_0))\sim 2k\log\frac{en}{k}$ from Table 2.1. Hence, when k and n is proportional and large, $2\sqrt{\frac{n}{k}}$ is not significant, and (2.5) is rather tight.

Proposition 2.6 indicates that, $\min_{\lambda\geq 0}\mathbf{D}(\lambda\partial f(\mathbf{x}_0))\approx\delta(\mathscr{T}_f(\mathbf{x}_0))$ and by optimally tuning λ, one can reduce the upper bound on the normalized error in (2.2) to as small as $\delta(\mathscr{T}_f(\mathbf{x}_0))$. Recall that, this is also the sample complexity of (1.4) as discussed in Chapter 1.4. This relation between the estimation error formula for (2.1) and the sample complexity of (1.4) is first proposed in [70] and rigorously established by the results of [167] and [4]. In Chapter 3, we will find similar error formulas based on the Gaussian width and distance terms for the less trivial lasso problem.

2.5 Inequalities for Gaussian Processes

Gaussian random variables have several nice properties that make their analysis more accessible. The following results are on the Lipschitz functions of Gaussian vectors.

Fact 2.3 (Variance of Lipschitz functions, [131]) *Assume $\mathbf{g}\sim\mathscr{N}(0,\mathbf{I}_p)$ and let $f(\cdot):\mathbb{R}^p\to\mathbb{R}$ be an L-Lipschitz function. Then,*

$$Var[f(\mathbf{g})]\leq L^2.$$

Fact 2.4 (Gaussian concentration Inequality for Lipschitz functions, [131]) *Let $f(\cdot):\mathbb{R}^p\to\mathbb{R}$ be an L-Lipschitz function and $\mathbf{g}\sim\mathscr{N}(0,\mathbf{I}_p)$. Then,*

$$\mathbb{P}\left(|f(\mathbf{g})-\mathbb{E}[f(\mathbf{g})]|\geq t\right)\leq 2\exp(-\frac{t^2}{2L^2}).$$

Our proofs will often involve Lipschitzness to argue strong concentration around mean. For instance, ℓ_2 norm is a 1-Lipschitz function. More generally, given a subset A of $\mathbb{R}^n$, the distance function $\text{dist}(\mathbf{x},A) = \inf_{\mathbf{a}\in A} \|\mathbf{x}-\mathbf{a}\|_2$ is 1-Lipschitz in $\mathbf{x}$. Both of these essentially follow from application of triangle inequality. The following lemma shows that restricted singular value is Lipschitz as well.

Lemma 2.1 *Recall Definition 1.4 of $\sigma_{\mathscr{C}}(\mathbf{A})$ where $\mathscr{C}$ is a closed cone. $\sigma_{\mathscr{C}}(\mathbf{A})$ is 1-Lipschitz function of $\mathbf{A}$.*

Proof: Given $\mathbf{A},\mathbf{B} \in \mathbb{R}^{m\times n}$, define $\mathbf{v_A} = \arg\inf_{\mathbf{v}\in\mathscr{C}\cap\mathscr{S}^{n-1}} \|\mathbf{Av}\|_2$ and $\mathbf{v_B}$ similarly. We have that,

$$\sigma_{\mathscr{C}}(\mathbf{B}) - \sigma_{\mathscr{C}}(\mathbf{A}) = \|\mathbf{Bv_B}\|_2 - \|\mathbf{Av_A}\|_2 \leq \|\mathbf{Bv_A}\|_2 - \|\mathbf{Av_A}\|_2 \leq \|\mathbf{B}-\mathbf{A}\|_F$$

Repeating the same argument for $\sigma_{\mathscr{C}}(\mathbf{A}) - \sigma_{\mathscr{C}}(\mathbf{B})$ gives, $|\sigma_{\mathscr{C}}(\mathbf{B}) - \sigma_{\mathscr{C}}(\mathbf{A})| \leq \|\mathbf{B}-\mathbf{A}\|_F$. ■

2.5.1 Gaussian comparison inequalities

Comparison inequalities will play a major role in our analysis in Chapter 3. Slepian's Lemma is used to compare supremums of two Gaussian processes based on their covariances.

Theorem 2.1 (Slepian's Lemma, [131]) *Let $\{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{y}_i\}_{i=1}^n$ be zero-mean Gaussian random variables. Suppose, for all $1 \leq i \leq n$, $\mathbb{E}[\mathbf{x}_i^2] = \mathbb{E}[\mathbf{y}_i^2]$ and for all $1 \leq i \neq j \leq n$, $\mathbb{E}[\mathbf{x}_i\mathbf{x}_j] \leq \mathbb{E}[\mathbf{y}_i\mathbf{y}_j]$. Then,*

$$\mathbb{E}[\sup_{1\leq i\leq n} \mathbf{y}_i] \leq \mathbb{E}[\sup_{1\leq i\leq n} \mathbf{x}_i].$$

We will use a generalization of this result due to Gordon. Gordon's result will allow us to do minimax type comparisons.

Theorem 2.2 (Gaussian Min-Max Theorem, [111, 112]) *Let $\{X_{ij}\}$ and $\{Y_{ij}\}$, $1 \leq i \leq n$, $1 \leq j \leq m$, be two centered Gaussian processes which satisfy the following inequalities for all choices of indices*

1. $\mathbb{E}\left[|X_{ij} - X_{ik}|^2\right] \leq \mathbb{E}\left[|Y_{ij} - Y_{ik}|^2\right]$,
2. $\mathbb{E}\left[|X_{ij} - X_{\ell k}|^2\right] \geq \mathbb{E}\left[|Y_{ij} - Y_{\ell k}|^2\right], \quad$ *if* $i \neq \ell$.

Then,

$$\mathbb{E}[\min_i \max_j Y_{ij}] \geq \mathbb{E}[\min_i \max_j X_{ij}].$$

Suppose, we additionally have, $\mathbb{E}\left[X_{ij}^2\right] = \mathbb{E}\left[Y_{ij}^2\right]$ *for all* i, j. *Then,*

$$\mathbb{P}\left(\bigcap_{i=1}^{n}\bigcup_{j=1}^{m}[Y_{ij} \geq \lambda_{ij}]\right) \geq \mathbb{P}\left(\bigcap_{i=1}^{n}\bigcup_{j=1}^{m}[X_{ij} \geq \lambda_{ij}]\right),$$

for all scalars $\lambda_{ij} \in \mathbf{R}$.

An immediate application of these lemmas is Proposition 1.3, which can be used to find sharp bounds to minimum and maximum singular values of a Gaussian matrix. Chapter 3 will revisit these results for our own purposes, in particular, we will use the following variations of this result that can obtain lower bounds for certain linear functions of i.i.d Gaussian matrices. The first variation is available in Gordon's own paper [112] as a lemma.

Lemma 2.2 *Let* $\mathbf{G} \in \mathbb{R}^{m\times n}, g \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$ *be independent of each other and have independent standard normal entries. Also, let* $\mathscr{S} \subset \mathbb{R}^n$ *be an arbitrary set and* $\psi : \mathscr{S} \to \mathbb{R}$ *be an arbitrary function. Then, for any* $c \in \mathbb{R}$,

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathscr{S}}\left\{\|\mathbf{G}\mathbf{x}\|_2 + \|\mathbf{x}\|_2 g - \psi(\mathbf{x})\right\} \geq c\right) \geq \mathbb{P}\left(\min_{\mathbf{x}\in\mathscr{S}}\left\{\|\mathbf{x}\|_2\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{x} - \psi(\mathbf{x})\right\} \geq c\right). \tag{2.6}$$

The next variation is a slight generalization of Lemma 2.2 and can be obtained by following techniques that are similar to the proof of Lemma 2.2. The proof is provided in Section A.2.

Proposition 2.7 (Modified Gordon's Lemma) *Let* $\mathbf{G}$, $\mathbf{g}$, $\mathbf{h}$ *be defined as in Lemma 2.2 and let* $\Phi_1 \subset \mathbb{R}^n$ *be arbitrary and* $\Phi_2 \subset \mathbb{R}^m$ *be a compact set. Also, assume* $\psi(\cdot,\cdot) : \Phi_1 \times \Phi_2 \to \mathbb{R}$ *is a continuous function. Then, for any* $c \in \mathbb{R}$:

$$\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\right\} \geq c\right) \geq 2\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\right\} \geq c\right) - 1.$$

Observe that, one can obtain Lemma 2.2 from Proposition 2.7 by setting Φ_2 to be the unit ℓ_2 sphere $\mathscr{S}^{m-1}$ and by allowing ψ to be only a function of $\mathbf{x}$. The remaining difference is the extra term $\|\mathbf{x}\|_2 g$ in Lemma 2.2, which we use a symmetrization argument to get rid of at the expense of a slightly looser estimate on the right hand side of Proposition 2.7.

Chapter 3

A General Theory of Noisy Linear Inverse Problems

Consider the setup where we have noisy observations $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$. Recall from Chapter 1 that, a common approach to estimate $\mathbf{x}_0$ from $\mathbf{y}$ is to use the LASSO algorithm with a proper convex function $f(\cdot)$,

$$\mathbf{x}^*_{LASSO} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}) \right\}. \tag{3.1}$$

LASSO was originally introduced in [201] and has since then been subject of great interest as a natural and powerful approach to do noise robust compressed sensing (CS), [13, 14, 19, 24, 25, 57, 76, 148, 201, 217, 234]. There are also closely related algorithms such as SOCP variations and the Dantzig selector [31, 138]. Of course, applications of (3.1) are not limited to sparse recovery; they extend to various problems including the recovery of block sparse signals [147, 149], the matrix completion problem [37, 128] and the total variation minimization [160]. In each application, $f(\cdot)$ is chosen in accordance to the structure of $\mathbf{x}_0$. In this chapter, we consider arbitrary convex penalty functions $f(\cdot)$ and we will refer to this generic formulation in (3.1) as the "Generalized LASSO" or simply "LASSO" problem.

3.0.2 Motivation

The LASSO problem can be viewed as a "merger" of two closely related problems, which have both recently attracted a lot of attention by the research community; the problems of noiseless CS and that of proximal denoising.

• **Noiseless compressed sensing:** In the noiseless CS problem one wishes to recover $\mathbf{x}_0$ from the random linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0$. The standard approach is solving the generalized basis pursuit (1.4) discussed

in Chapter 1. Recall that a critical performance criteria for the GBP problem (1.4) concerns the minimum number of measurements needed to guarantee successful recovery of $\mathbf{x}_0$ [4, 50, 74, 83, 85, 192, 194]. Here, success means that $\mathbf{x}_0$ is the unique minimizer of (1.4), with high probability, over the realizations of the random matrix $\mathbf{A}$.

• **Proximal denoising:** As mentioned in Section 2.4, the proximal denoising problem tries to estimate $\mathbf{x}_0$ from noisy but uncompressed observations $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ via (2.1). A closely related approach to estimate $\mathbf{x}_0$, which requires prior knowledge $f(\mathbf{x}_0)$ about the signal of interest $\mathbf{x}_0$, is solving the constrained denoising problem:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|_2^2 \quad \text{subject to} \quad f(\mathbf{x}) \leq f(\mathbf{x}_0). \tag{3.2}$$

The natural question to be posed in both cases is how well can one estimate $\mathbf{x}_0$ via (2.1) (or (3.2)) [49, 70, 72, 166, 167]? The minimizer $\mathbf{x}^*$ of (2.1) (or (3.2)) is a function of the noise vector $\mathbf{z}$ and the common measure of performance, is the normalized mean-squared-error which is defined as $\frac{\mathbb{E}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{\sigma^2}$.

3.0.2.1 The "merger" LASSO

The Generalized LASSO problem is naturally merging the problems of noiseless CS and proximal denoising. The compressed nature of measurements, poses the question of finding the minimum number of measurements required to recover $\mathbf{x}_0$ *robustly*, that is with error proportional to the noise level. When recovery is robust, it is of importance to be able to explicitly characterize how good the estimate is. In this direction, when $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$, a common measure of performance for the LASSO estimate $\mathbf{x}^*_{LASSO}$ is defined to be the **normalized squared error** (NSE) :

$$NSE = \frac{1}{\sigma^2} \|\mathbf{x}^*_{LASSO} - \mathbf{x}_0\|_2^2.$$

This is exactly the main topic of this chapter: *proving precise bounds for the NSE of the Generalized LASSO problem.*

In the specific case of ℓ_1-penalization in (3.1), researchers have considered other performance criteria additional to the NSE [76, 217, 234]. As an example, we mention the support recovery criteria [217], which measures how well (3.1) recovers the subset of nonzero indices of $\mathbf{x}_0$. However, under our general setup, where we allow arbitrary structure to the signal $\mathbf{x}_0$, the NSE serves as the most natural measure of perfor-

mance and is, thus, the sole focus in this chapter. In the relevant literature, researchers have dealt with the analysis of the NSE of (3.1) under several settings (see Section 3.0.4). Yet, *we still lack a general theory that would yield precise bounds for the squared-error of* (3.1) *for arbitrary convex regularizer* $f(\cdot)$. *We aim to close this gap.* Our answer involves inherent quantities regarding the geometry of the problem which, in fact, have recently appeared in the related literature, [4, 13, 14, 50, 101, 167].

3.0.3 Three Versions of the LASSO Problem

Throughout the analysis, we assume $\mathbf{A} \in \mathbb{R}^{m\times n}$ has independent standard normal entries and $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$. Our approach tackles various forms of the LASSO all at once, and relates them to each other. In particular, we consider the following three versions:

- ⋆ **C-LASSO**: Assumes a-priori knowledge of $f(\mathbf{x}_0)$ and solves,

$$\mathbf{x}_c^*(\mathbf{A},\mathbf{z}) = \arg\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 \quad \text{subject to} \quad f(\mathbf{x}) \le f(\mathbf{x}_0). \tag{3.3}$$

- ⋆ ℓ_2**-LASSO**[1]: Uses ℓ_2-penalization rather than ℓ_2^2 and solves,

$$\mathbf{x}_{\ell_2}^*(\lambda,\mathbf{A},\mathbf{z}) = \arg\min_{\mathbf{x}} \{\ \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2 + \lambda f(\mathbf{x})\ \}. \tag{3.4}$$

- ⋆ ℓ_2^2**-LASSO**: the original form given in (3.1) :

$$\mathbf{x}_{\ell_2^2}^*(\tau,\mathbf{A},\mathbf{z}) = \arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \sigma\tau f(\mathbf{x}) \right\}. \tag{3.5}$$

C-LASSO in (3.3) stands for "Constrained LASSO". This version of the LASSO problem assumes some a-priori knowledge about $\mathbf{x}_0$, which makes the analysis of the problem arguably simpler than that of the other two versions, in which the role of the penalty parameter (which is meant to compensate for the lack of a-priori knowledge) has to be taken into consideration. To distinguish between the ℓ_2-LASSO and the ℓ_2^2-LASSO, we use λ to denote the penalty parameter of the former and τ for the penalty parameter of the latter. Part of our contribution is establishing useful connections between these three versions of the LASSO problem. We will often drop the arguments $\lambda, \tau, \mathbf{A}, \mathbf{z}$ from the LASSO estimates defined in (3.3)–(3.5), when clear from context.

[1] ℓ_2-lasso is originally introduced by Belloni who dubbed it as "square-root lasso" [16].

3.0.4 Relevant Literature

Precise characterization of the NSE of the LASSO is closely related to the precise performance analysis of noiseless CS and proximal denoising. To keep the discussion short, we defer most of the comments on the connections of our results to these problems to the main body of this chapter. Table 3.1 provides a summary of the relevant literature and highlights the area of our contribution.

	Convex functions	**ℓ_1-minimization**
Noiseless CS	Chandrasekaran et al. [50] Amelunxen et al. [4]	Donoho and Tanner, [83] Stojnic, [192]
Proximal denoising	Donoho et al. [70] Oymak and Hassibi [167]	Donoho [72]
LASSO	Our contribution	Bayati and Montanari, [13], [14] Stojnic, [193]

Table 3.1: Relevant Literature.

The works closest in spirit to our results include [13, 14, 142, 193], which focus on the exact analysis of the LASSO problem, while restricting the attention on sparse recovery where $f(\mathbf{x}) = \|\mathbf{x}\|_1$. In [13, 14], Bayati and Montanari are able to show that the mean-squared-error of the LASSO problem is equivalent to the one achieved by a properly defined "Approximate Message Passing" (AMP) algorithm. Following this connection and after evaluating the error of the AMP algorithm, they obtain an explicit expression for the mean squared error of the LASSO algorithm in an asymptotic setting. In [142], Maleki et al. proposes Complex AMP, and characterizes the performance of LASSO for sparse signals with complex entries. In [193], Stojnic's approach relies on results on Gaussian processes [111, 112] to derive sharp bounds for the *worst case NSE* of the ℓ_1-constrained LASSO problem in (3.3). Our approach in this chapter builds on the framework proposed by Stojnic, but extends the results in multiple directions as noted in the next section.

3.0.5 Contributions

This section summarizes our main contributions. In short, this chapter:

- *generalizes* the results of [193] on the constrained LASSO for arbitrary convex functions; proves that the worst case NSE is achieved when the noise level $\sigma \to 0$, and derives sharp bounds for it.

- *extends* the analysis to the NSE of the more challenging ℓ_2-LASSO; provides bounds as a function of the penalty parameter λ, which are sharp when $\sigma \to 0$.
- identifies a *connection* between the ℓ_2-LASSO to the ℓ_2^2-LASSO; proposes a formula for precisely calculating the NSE of the latter when $\sigma \to 0$.
- provides simple *recipes* for the optimal tuning of the penalty parameters λ and τ in the ℓ_2 and ℓ_2^2-LASSO problems.
- analyzes the regime in which *stable* estimation of $\mathbf{x}_0$ fails.

3.0.6 Motivating Examples

Before going into specific examples, it is instructive to consider the scenario where $f(\cdot) = 0$. This reduces the problem to a regular least-squares estimation problem, the analysis of which is easy to perform. When $m < n$, the system is underdetermined, and one cannot expect $\mathbf{x}^*$ to be a good estimate. When $m \geq n$, the estimate can be given by $\mathbf{x}^* = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$. In this case, the normalized mean-squared-error takes the form,

$$\frac{\mathbb{E}\,\|\mathbf{x}^* - \mathbf{x}_0\|^2}{\sigma^2} = \frac{\mathbb{E}[\mathbf{z}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-2}\mathbf{A}^T\mathbf{z}]}{\sigma^2} = \mathbb{E}[\mathrm{tr}\,(()\,\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-2}\mathbf{A}^T)] = \mathbb{E}[\mathrm{tr}\,(()\,(\mathbf{A}^T\mathbf{A})^{-1})].$$

$\mathbf{A}^T\mathbf{A}$ is a Wishart matrix and its inverse is well studied. In particular, when $m \geq n+2$, we have $\mathbb{E}[(\mathbf{A}^T\mathbf{A})^{-1}] = \frac{\mathbf{I}_n}{m-n-1}$ (see [173]). Hence,

$$\frac{\mathbb{E}\,\|\mathbf{x}^* - \mathbf{x}_0\|^2}{\sigma^2} = \frac{n}{m-n-1}. \tag{3.6}$$

How does this result change when a nontrivial convex function $f(\cdot)$ is introduced?

Our message is simple: *when $f(\cdot)$ is an arbitrary convex function, the LASSO error formula is obtained by simply replacing the ambient dimension n in (3.6) with a summary parameter* $\mathbf{D}(cone(\partial f(\mathbf{x}_0)))$ *or* $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. These parameters are defined as the expected squared-distance of a standard normal vector in $\mathbb{R}^n$ to the conic hull of the subdifferential $\mathrm{cone}(\partial f(\mathbf{x}_0))$ and to the scaled subdifferential $\lambda\partial f(\mathbf{x}_0)$, respectively. They summarize the effect of the structure of the signal $\mathbf{x}_0$ and choice of the function $f(\cdot)$ on the estimation error.

To get a flavor of the (simple) nature of our results, we briefly describe how they apply in three commonly encountered settings, namely the "sparse signal", "low-rank matrix" and "block-sparse signal" estimation problems. For simplicity of exposition, let us focus on the C-LASSO estimator in (3.3). A more elaborate

discussion, including estimation via ℓ_2-LASSO and ℓ_2^2-LASSO, can be found in Section 3.3.4. The following statements are true with high probability in $\mathbf{A}, \mathbf{v}$ and hold under mild assumptions.

1. **Sparse signal estimation:** Assume $\mathbf{x}_0 \in \mathbb{R}^n$ has k nonzero entries. In order to estimate $\mathbf{x}_0$, use the Constrained-LASSO and pick ℓ_1-norm for $f(\cdot)$. Let $m > 2k(\log \frac{n}{k} + 1)$. Then,

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \lesssim \frac{2k(\log \frac{n}{k} + 1)}{m - 2k(\log \frac{n}{k} + 1)}. \tag{3.7}$$

2. **Low-rank matrix estimation:** Assume $\mathbf{X}_0 \in \mathbb{R}^{d \times d}$ is a rank r matrix, $n = d \times d$. This time, $\mathbf{x}_0 \in \mathbb{R}^n$ corresponds to vectorization of $\mathbf{X}_0$ and $f(\cdot)$ is chosen as the nuclear norm $\|\cdot\|_\star$ (sum of the singular values of a matrix) [94, 178]. Hence, we observe $\mathbf{y} = \mathbf{A} \cdot \mathrm{vec}(\mathbf{X}_0) + \mathbf{z}$ and solve,

$$\min_{\mathbf{X} \in \mathbb{R}^{d \times d}} \|\mathbf{y} - \mathbf{A} \cdot \mathrm{vec}(\mathbf{X})\|_2 \quad \text{subject to} \quad \|\mathbf{X}\|_\star \leq \|\mathbf{X}_0\|_\star$$

Let $m > 6dr$. Denote the LASSO estimate by $\mathbf{X}_c^*$ and use $\|\cdot\|_F$ for the Frobenius norm of a matrix. Then,

$$\frac{\|\mathbf{X}_c^* - \mathbf{X}_0\|_F^2}{\sigma^2} \lesssim \frac{6dr}{m - 6dr}. \tag{3.8}$$

3. **Block sparse estimation:** Let $n = t \times b$ and assume the entries of $\mathbf{x}_0 \in \mathbb{R}^n$ can be grouped into t known blocks of size b so that only k of these t blocks are nonzero. To induce the structure, the standard approach is to use the $\ell_{1,2}$ norm which sums up the ℓ_2 norms of the blocks, [91, 175, 191, 195]. In particular, denoting the subvector corresponding to i'th block of a vector $\mathbf{x}$ by $\mathbf{x}_i$, the $\ell_{1,2}$ norm is equal to $\|\mathbf{x}\|_{1,2} = \sum_{i=1}^t \|\mathbf{x}_i\|_2$. Assume $m > 4k(\log \frac{t}{k} + b)$. Then,

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \lesssim \frac{4k(\log \frac{t}{k} + b)}{m - 4k(\log \frac{t}{k} + b)}. \tag{3.9}$$

Note how (3.7)-(3.9) are similar in nature to (3.6).

3.1 Our Approach

In this section we introduce the main ideas that underlie our approach. This will also allow us to introduce important concepts from convex geometry required for the statements of our main results in Section 3.2. The details of most of the technical discussion in this introductory section are deferred to later sections. To

keep the discussion concise, we focus our attention on the ℓ_2-LASSO.

3.1.1 First-Order Approximation

Recall the ℓ_2-LASSO problem introduced in (3.4):

$$\mathbf{x}^*_{\ell_2} = \arg\min_{\mathbf{x}} \{ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda f(\mathbf{x}) \}. \tag{3.10}$$

A key idea behind our approach is using the linearization of the convex structure inducing function $f(\cdot)$ around the vector of interest $\mathbf{x}_0$ [22, 182]:

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}_0) + \sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T(\mathbf{x}-\mathbf{x}_0). \tag{3.11}$$

$\partial f(\mathbf{x}_0)$ denotes the subdifferential of $f(\cdot)$ at $\mathbf{x}_0$ and is always a compact and convex set [182]. Throughout, we assume that $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$, hence, $\partial f(\mathbf{x}_0)$ does not contain the origin. From convexity of $f(\cdot)$, $f(\mathbf{x}) \geq \hat{f}(\mathbf{x})$, for all $\mathbf{x}$. What is more, when $\|\mathbf{x}-\mathbf{x}_0\|_2$ is sufficiently small, then $\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$. We substitute $f(\cdot)$ in (3.10) by its first-order approximation $\hat{f}(\cdot)$, to get a corresponding "*Approximated LASSO*" problem. To write the approximated problem in an easy-to-work-with format, recall that $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} = \mathbf{A}\mathbf{x}_0 + \sigma\mathbf{v}$, for $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_m)$ and change the optimization variable from $\mathbf{x}$ to $\mathbf{w} = \mathbf{x} - \mathbf{x}_0$:

$$\hat{\mathbf{w}}_{\ell_2}(\lambda, \sigma, \mathbf{A}, \mathbf{v}) = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w} \right\}. \tag{3.12}$$

We will often drop all or part of the arguments $\lambda, \sigma, \mathbf{A}, \mathbf{v}$ above, when it is clear from the context. We denote $\hat{\mathbf{w}}_{\ell_2}$ for the optimal solution of the approximated problem in (3.12) and $\mathbf{w}^*_{\ell_2} = \mathbf{x}^*_{\ell_2} - \mathbf{x}_0$ for the optimal solution of the original problem in (3.10)[2]. Also, denote the optimal cost achieved in (3.11) by $\hat{\mathbf{w}}_{\ell_2}$, as $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$.

Taking advantage of the simple characterization of $\hat{f}(\cdot)$ via the subdifferential $\partial f(\mathbf{x}_0)$, we are able to *precisely* analyze the optimal cost and the normalized squared error of the resulting approximated problem. The approximation is tight when $\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|_2 \to 0$ and we later show that this is the case when the noise level $\sigma \to 0$. This fact allows us to translate the results obtained for the Approximated LASSO problem to corresponding *precise* results for the Original LASSO problem, in the small noise variance regime.

[2]We follow this conventions throughout the chapter: use the symbol "ˆ" over variables that are associated with the approximated problems. To distinguish, use the symbol " * " for the variables associated with the original problem .

3.1.2 Importance of $\sigma \to 0$

In this chapter, we focus on the precise characterization of the NSE. While we show that the first order characteristics of the function, i.e. $\partial f(\mathbf{x}_0)$, suffice to provide sharp and closed-form bounds for small noise level σ, we believe that higher order terms are required for such precise results when σ is arbitrary. On the other hand, we empirically observe that the worst case NSE for the LASSO problem is achieved when $\sigma \to 0$. While we do not have a proof for the validity of this statement for the ℓ_2- and ℓ_2^2-LASSO, we *do prove* that this is indeed the case for the C-LASSO problem. Interestingly, the same phenomena has been observed and proved to be true for related estimation problems, for example for the proximal denoising problem (2.1) in [70,78,167] and, closer to the present chapter, for the LASSO problem with ℓ_1 penalization (see Donoho et al. [82]).

Summarizing, for the C-LASSO problem, we derive a formula that sharply characterizes its NSE for the small σ regime and we show that the same formula upper bounds the NSE when σ is arbitrary. Proving the validity of this last statement for the ℓ_2- and ℓ_2^2-LASSO would ensure that our corresponding NSE formulae for small σ provide upper bounds to the NSE for arbitrary σ.

3.1.3 Gaussian Min-Max Theorem

Perhaps the most important technical ingredient of the analysis presented in this chapter is Gaussian Min-Max Theorem [112]. For the purposes of our analysis, we will make use of Proposition 2.7. Here, it suffices to observe that the Gordon's original statement Lemma 2.2 is (almost) directly applicable to the LASSO problem in (3.12). First, write $\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 = \max_{\|\mathbf{a}\|_2=1} \mathbf{a}^T [\mathbf{A}, -\mathbf{v}] \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix}$ and take function $\psi(\cdot)$ in the lemma to be $\sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}$. Then, the optimization problem in the left hand side of (2.6) takes the format of the LASSO problem in (3.12), except for the "distracting" factor $\|\mathbf{x}\|_2 g$. Proposition 2.7 takes care of this extra term without affecting the essence of the probabilistic statement of Lemma 2.2. Details being postponed to the later sections (cf. Section 3.4), Corollary 3.1 below summarizes the result of applying Proposition 2.7 to the LASSO problem.

Corollary 3.1 (Lower Key Optimization) *Let* $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$, $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *and* $h \sim \mathcal{N}(0,1)$ *be independent of each other. Define the following optimization problem:*

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}} \left\{ \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} + \max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w} \right\}. \tag{3.13}$$

Then, for any $c \in \mathbb{R}$:

$$\mathbb{P}\left(\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v}) \geq c \right) \geq 2 \cdot \mathbb{P}\left(\mathscr{L}(\mathbf{g},\mathbf{h}) - h\sigma \geq c \right) - 1.$$

Corollary 3.1 establishes a probabilistic connection between the LASSO problem and the minimization (3.13). In the next section, we argue that the latter is much easier to analyze than the former. Intuitively, the main reason is that instead of an $m \times n$ matrix, (3.13) only involves two vectors of sizes $m \times 1$ and $n \times 1$. Even more, those vectors have independent standard normal entries and are independent of each other, which greatly facilitates probabilistic statements about the value of $\mathscr{L}(\mathbf{g},\mathbf{h})$. Due to its central role in our analysis, we often refer to problem (3.13) as "*key optimization*" or "*lower key optimization*". The term "lower" is attributed to the fact that analysis of (3.13) results in a probabilistic lower bound for the optimal cost of the LASSO problem.

3.1.4 Analyzing the Key Optimization

3.1.4.1 Deterministic Analysis

First, we perform the deterministic analysis of $\mathscr{L}(\mathbf{g},\mathbf{h})$ for fixed $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$. In particular, we reduce the optimization in (3.13) to a *scalar* optimization. To see this, perform the optimization over a fixed ℓ_2-norm of $\mathbf{w}$ to equivalently write

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\|_2 - \max_{\|\mathbf{w}\|_2 = \alpha} \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} (\mathbf{h} - \mathbf{s})^T \mathbf{w} \right\}.$$

The maximin problem that appears in the objective function of the optimization above has a simple solution. It can be shown that

$$\begin{aligned} \max_{\|\mathbf{w}\|_2 = \alpha} \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} (\mathbf{h} - \mathbf{s})^T \mathbf{w} &= \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \max_{\|\mathbf{w}\|_2 = \alpha} (\mathbf{h} - \mathbf{s})^T \mathbf{w} \\ &= \alpha \min_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \|\mathbf{h} - \mathbf{s}\|_2. \end{aligned}$$

This reduces (3.13) to a scalar optimization problem over α, for which one can compute the optimal value $\hat{\alpha}$ and the corresponding optimal cost. The result is summarized in Lemma 3.1 below. For the statement of the lemma, for any vector $\mathbf{v} \in \mathbb{R}^n$ define its projection and its distance to a convex and closed set $\mathscr{C} \in \mathbb{R}^n$ as

$$\text{Proj}(\mathbf{v},\mathscr{C}) := \text{argmin}_{\mathbf{s}\in\mathscr{C}} \|\mathbf{v}-\mathbf{s}\|_2 \quad \text{and} \quad \text{dist}(\mathbf{v},\mathscr{C}) := \|\mathbf{v}-\text{Proj}(\mathbf{v},\mathscr{C})\|_2.$$

Lemma 3.1 *Let* $\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})$ *be a minimizer of the problem in* (3.13)*. If* $\|\mathbf{g}\|_2 > dist(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$*, then,*

$$a) \quad \hat{\mathbf{w}}(\mathbf{g},\mathbf{h}) = \sigma \frac{\mathbf{h} - Proj(\mathbf{h},\lambda\partial f(\mathbf{x}_0))}{\sqrt{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))}},$$

$$b) \quad \|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|_2^2 = \sigma^2 \frac{dist^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))}{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))},$$

$$c) \quad \mathscr{L}(\mathbf{g},\mathbf{h}) = \sigma\sqrt{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))}.$$

3.1.4.2 Probabilistic Analysis

Of interest is making probabilistic statements about $\mathscr{L}(\mathbf{g},\mathbf{h})$ and the norm of its minimizer $\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|_2$. Lemma 3.1 provided closed form deterministic solutions for both of them, which only involve the quantities $\|\mathbf{g}\|_2^2$ and $\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$. For $\mathbf{g}\sim\mathscr{N}(0,\mathbf{I}_m)$ and $\mathbf{h}\sim\mathscr{N}(0,\mathbf{I}_n)$, standard results on Gaussian concentration show that, these quantities concentrate nicely around their means $\mathbb{E}\left[\|\mathbf{g}\|_2^2\right] = m$ and $\mathbb{E}\left[\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))\right]$ $=: \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, respectively. Combining these arguments with Lemma 3.1, we conclude with Lemma 3.2 below.

Lemma 3.2 (Probabilistic Result) *Assume that* $(1-\varepsilon_L)m \geq \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \geq \varepsilon_L m$ *for some constant* $\varepsilon_L > 0$*. Define*[3]*,*

$$\eta = \sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))} \quad \text{and} \quad \gamma = \frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}.$$

Then, for any $\varepsilon > 0$*, there exists a constant* $c > 0$ *such that, for sufficiently large* m*, with probability* $1 - \exp(-cm)$*,*

$$|\mathscr{L}(\mathbf{g},\mathbf{h}) - \sigma\eta| \leq \varepsilon\sigma\eta, \quad \text{and} \quad \left|\frac{\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|_2^2}{\sigma^2} - \gamma\right| \leq \varepsilon\gamma.$$

[3]Observe that the dependence of η and γ on λ, m and $\partial f(\mathbf{x}_0)$, is implicit in this definition.

Remark: In Lemma 3.2, the condition "$(1-\varepsilon_L)m \geq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$" ensures that $\|\mathbf{g}\|_2 > \text{dist}(\mathbf{h}, \lambda\partial f(\mathbf{x}_0))$ (cf. Lemma 3.1) with high probability over the realizations of $\mathbf{g}$ and $\mathbf{h}$.

3.1.5 The Predictive Power of Gaussian Min-Max Theorem

Let us recap the last few steps of our approach. Application of the "modified Gordon's Lemma" Proposition 2.7 to the approximated LASSO problem in (3.12) introduced the simpler lower key optimization (3.13). Without much effort, we found in Lemma 3.2 that its cost $\mathscr{L}(\mathbf{g},\mathbf{h})$ and the normalized squared norm of its minimizer $\frac{\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|_2^2}{\sigma^2}$ concentrate around $\sigma\eta$ and γ, respectively. This brings the following question:

- *To what extent do such results on $\mathscr{L}(\mathbf{g},\mathbf{h})$ and $\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})$ translate to useful conclusions about $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ and $\hat{\mathbf{w}}_{\ell 2}(\mathbf{A},\mathbf{v})$?*

Application of Proposition 2.7 as performed in Corollary 3.1 when combined with Lemma 3.2, provide a preliminary answer to this question: $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ is lower bounded by $\sigma\eta$ with overwhelming probability. Formally,

Lemma 3.3 (Lower Bound) *Assume $(1-\varepsilon_L)m \geq \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \geq \varepsilon_L m$ for some constant $\varepsilon_L > 0$ and m is sufficiently large. Then, for any $\varepsilon > 0$, there exists a constant $c > 0$ such that, with probability $1-\exp(-cm)$,*

$$\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v}) \geq (1-\varepsilon)\sigma\eta.$$

But is that all? A major part of our technical analysis in the remainder of this chapter involves showing that the connection between the LASSO problem and the simple optimization (3.13) is much *deeper* than Lemma 3.3 predicts. In short, under certain conditions on λ and m (similar in nature to those involved in the assumption of Lemma 3.3), we prove that the followings are true:

- Similar to $\mathscr{L}(\mathbf{g},\mathbf{h})$, the optimal cost $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ of the approximated ℓ_2-LASSO concentrates around $\sigma\eta$.
- Similar to $\frac{\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|_2^2}{\sigma^2}$, the NSE of the approximated ℓ_2-LASSO $\frac{\|\hat{\mathbf{w}}_{\ell_2}(\mathbf{A},\mathbf{v})\|_2^2}{\sigma^2}$ concentrates around γ.

In some sense, $\mathscr{L}(\mathbf{g},\mathbf{h})$ "predicts" $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ and $\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|_2$ "predicts" $\|\hat{\mathbf{w}}_{\ell_2}(\mathbf{A},\mathbf{v})\|_2$, which attributes Proposition 2.7 (or more precisely to the lower key optimization) a "predictive power". This power is not necessarily restricted to the two examples above. In Section 3.7, we extend the applicability of this idea to prove that worst case NSE of the C-LASSO is achieved when $\sigma \to 0$. Finally, in Section 3.12 we rely on this predictive power of Proposition 2.7 to motivate our claims regarding the ℓ_2^2-LASSO.

The main idea behind the framework that underlies the proof of the above claims was originally introduced by Stojnic in his recent work [193] in the context of the analysis of the ℓ_1-constrained LASSO. While the fundamentals of the approach remain similar, we significantly extend the existing results in multiple directions by analyzing the more involved ℓ_2-LASSO and ℓ_2^2-LASSO problems and by generalizing the analysis to arbitrary convex functions. A synopsis of the framework is provided in the next section, while the details are deferred to later sections.

3.1.6 Synopsis of the Technical Framework

We highlight the main steps of the technical framework.

1. Apply Proposition 2.7 to $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ to find a *high-probability lower bound* for it. (cf. Lemma 3.3)
2. Apply Proposition 2.7 to the *dual* of $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ to find a *high-probability upper bound* for it.
3. Both lower and upper bounds can be made arbitrarily close to $\sigma\eta$. Hence, $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ concentrates with high probability around $\sigma\eta$ as well.
4. Assume $\frac{\|\hat{\mathbf{w}}_{\ell_2}\|_2^2}{\sigma^2}$ deviates from γ. A third application of Proposition 2.7 shows that such a deviation would result in a *significant increase* in the optimal cost, namely $\hat{\mathscr{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ would be significantly larger than $\sigma\eta$.
5. From the previous step, conclude that $\frac{\|\hat{\mathbf{w}}_{\ell_2}\|_2^2}{\sigma^2}$ concentrates with high probability around γ.

3.1.7 Gaussian Squared Distance and Related Quantities

The Gaussian squared distance to the λ-scaled set of subdifferential of $f(\cdot)$ at $\mathbf{x}_0$,

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) := \mathbb{E}\left[\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))\right], \tag{3.14}$$

has been key to our discussion above. Here, we explore some of its useful properties and introduce some other relevant quantities that altogether capture the (convex) geometry of the problem. Given a set $\mathscr{C}\in\mathbb{R}^n$, denote its conic hull by $\text{cone}(\mathscr{C})$. Also, denote its polar cone by $\mathscr{C}^\circ$, which is the closed and convex set $\{\mathbf{u}\in\mathbb{R}^n \big| \mathbf{u}^T\mathbf{v}\leq 0 \text{ for all } \mathbf{v}\in\mathscr{C}\}$.

Let $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$. Then, define,

$$\mathbf{C}(\lambda \partial f(\mathbf{x}_0)) := \mathbb{E}\left[(\mathbf{h} - \text{Proj}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)))^T \text{Proj}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \right], \tag{3.15}$$

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) := \mathbb{E}\left[\text{dist}^2(\mathbf{h}, \text{cone}(\partial f(\mathbf{x}_0))) \right]. \tag{3.16}$$

From the previous discussion, it has become clear how $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ appears in the analysis of the NSE of the ℓ_2-LASSO. $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ replaces $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ in the case of C-LASSO. This correspondence is actually not surprising as the approximated C-LASSO problem can be written in the format of the problem in (3.12) by replacing $\lambda \partial f(\mathbf{x}_0)$ with $\text{cone}(\partial f(\mathbf{x}_0))$. While $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ is the only quantity that appears in the analysis of the C-LASSO, the analysis of the ℓ_2-LASSO requires considering not only $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ but also $\mathbf{C}(\lambda \partial f(\mathbf{x}_0))$. $\mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ appears in the analysis during the second step of the framework described in Section 3.1.6. In fact, $\mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ is closely related to $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ as the following lemma shows.

Lemma 3.4 ([4]) *Suppose $\partial f(\mathbf{x}_0)$ is nonempty and does not contain the origin. Then,*

1. *$\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ is a strictly convex function of $\lambda \geq 0$, and is differentiable for $\lambda > 0$.*
2. $\frac{\partial \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{\partial \lambda} = -\frac{2}{\lambda} \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$.

As a last remark, the quantities $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ also play a crucial role in the analysis of the Noiseless CS and the Proximal Denoising problems. Without going into details, we mention that it has been recently proved in [4][4] that the noiseless compressed sensing problem (1.4) exhibits a transition from "failure" to "success" around $m \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. Also, [49, 70, 167] shows that $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ are equal to the worst case normalized mean-squared-error of the proximal denoisers (2.1) and (3.2) respectively. Recall from Proposition 2.6 that $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ is close to the optimally tuned distance $\min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ under mild assumptions (also see [4, 101, 167]).

3.2 Main Results

This section provides the formal statements of our main results. A more elaborate discussion follows in Section 3.3.

[4]Recall from Chapter 2 that $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ is same as the statistical dimension of the $(\text{cone}(\partial f(\mathbf{x}_0)))^\circ$, or equivalently (see Lemma 3.11) of the descent cone of $f(\cdot)$ at $\mathbf{x}_0$.

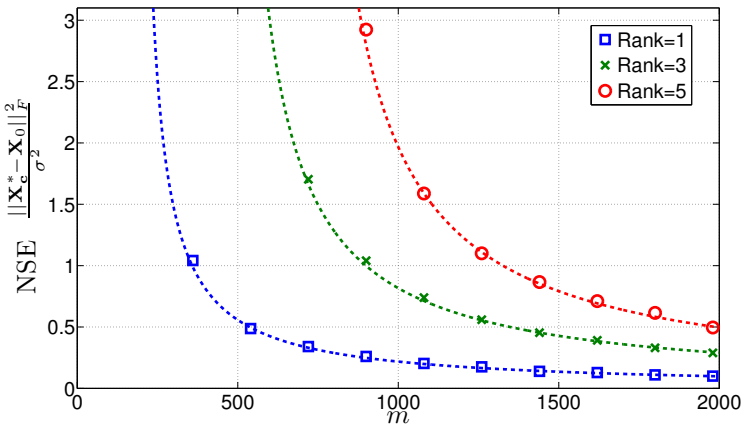


Figure 3.1: We have considered the Constrained-LASSO with nuclear norm minimization and fixed the signal to noise ratio $\frac{\|\mathbf{X}_0\|_F^2}{\sigma^2}$ to 10^5. Size of the underlying matrices are 40×40 and their ranks are $1, 3$ and 5. Based on [78, 163], we estimate $\mathbf{D}_f(\mathbf{X}_0, \mathbb{R}^+) \approx 179,450$ and 663 respectively. As the rank increases, the corresponding $\mathbf{D}_f(\mathbf{X}_0, \mathbb{R}^+)$ increases and the normalized squared error increases.

3.2.1 Setup

Before stating our results, we repeat our basic assumptions on the model of the LASSO problem. Recall the definitions of the three versions of the LASSO problem as given in (3.3), (3.4) and (3.5). Therein, assume:

- $\mathbf{A} \in \mathbb{R}^{m \times n}$ has independent standard normal entries,
- $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$,
- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuous,
- $\partial f(\mathbf{x}_0)$ does *not* contain the origin.

The results to be presented hold with high probability over the realizations of the measurement matrix $\mathbf{A}$ and the noise vector $\mathbf{v}$. Finally, recall the definitions of the quantities $\mathbf{D}(\lambda \partial f(\mathbf{x}_0)), \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ in (3.14), (3.15) and (3.16), respectively.

3.2.2 C-LASSO

Theorem 3.1 (NSE of C-LASSO) *Assume that m is sufficiently large and there exists a constant $\varepsilon_L > 0$ such that, $(1 - \varepsilon_L)m \geq \mathbf{D}(cone(\partial f(\mathbf{x}_0))) \geq \varepsilon_L m$. For any $\varepsilon > 0$, there exists a constant $C = C(\varepsilon, \varepsilon_L) > 0$*

such that, with probability $1-\exp(-Cm)$,

$$\frac{\|\mathbf{x}_c^*-\mathbf{x}_0\|_2^2}{\sigma^2} \leq (1+\varepsilon)\frac{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}, \tag{3.17}$$

Furthermore, there exists a deterministic number $\sigma_0>0$ *(i.e. independent of* $\mathbf{A},\mathbf{v}$*) such that, if* $\sigma\leq\sigma_0$*, with the same probability,*

$$\left|\frac{\|\mathbf{x}_c^*-\mathbf{x}_0\|_2^2}{\sigma^2}\times\frac{m-\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}-1\right|<\varepsilon. \tag{3.18}$$

3.2.3 ℓ_2-LASSO

Definition 3.1 ($\mathscr{R}_{\mathbf{ON}}$) *Suppose* $m>\min_{\lambda\geq 0}\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$*. Define* $\mathscr{R}_{ON}$ *as follows,*

$$\mathscr{R}_{ON}=\{\lambda>0\mid m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))>\max\{0,\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}\}.$$

Remark: *Section 3.8 fully characterizes* $\mathscr{R}_{ON}$ *and shows that it is an open interval.*

Theorem 3.2 (NSE of ℓ_2-LASSO in $\mathscr{R}_{\mathbf{ON}}$) *Assume there exists a constant* $\varepsilon_L>0$ *such that* $(1-\varepsilon_L)m\geq\max\{\mathbf{D}(\lambda\partial f(\mathbf{x}_0)),\ \mathbf{D}(\lambda\partial f(\mathbf{x}_0))+\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}$ *and* $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))\geq\varepsilon_L m$*. Further, assume that* m *is sufficiently large. Then, for any* $\varepsilon>0$*, there exists a constant* $C=C(\varepsilon,\varepsilon_L)>0$ *and a deterministic number* $\sigma_0>0$ *(i.e. independent of* $\mathbf{A},\mathbf{v}$*) such that, whenever* $\sigma\leq\sigma_0$*, with probability* $1-\exp(-C\min\{m,\frac{m^2}{n}\})$*,*

$$\left|\frac{\|\mathbf{x}_{\ell_2}^*-\mathbf{x}_0\|_2^2}{\sigma^2}\times\frac{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}-1\right|<\varepsilon. \tag{3.19}$$

Nonasymptotic bounds: In Section 3.10, we find a simple and non-asymptotic (which does not require $m,\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ to be large) bound for the ℓ_2-lasso which is strikingly close to what one would expect from Theorem 3.2. This new bound holds for all penalty parameters $\lambda\geq 0$ and all noise levels $\sigma>0$ and gives the bound $\frac{\|\mathbf{x}_{\ell_2}^*-\mathbf{x}_0\|_2^2}{\sigma^2}\lesssim\frac{4\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{(\sqrt{m}-\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))})^2}$ (see Theorem 3.7). Observe that, the difference between this and $\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ is only a factor of 4 in the regime $m\gg\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$.

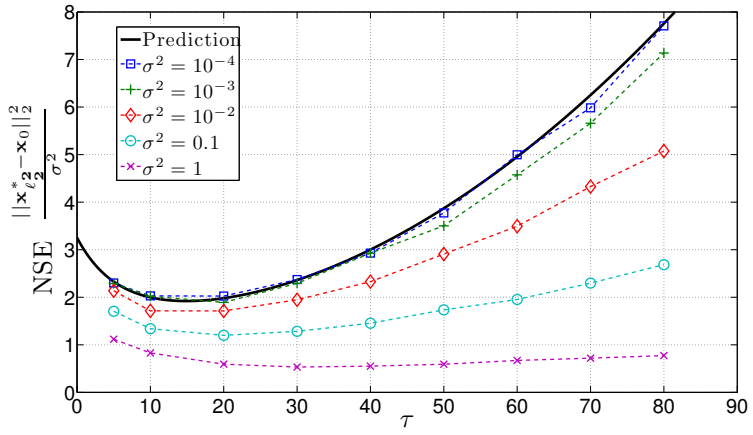


Figure 3.2: We considered ℓ_2^2-LASSO problem, for a k sparse signal of size $n = 1000$. We let $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$ and normalize the signal power by setting $\|\mathbf{x}_0\|_2 = 1$. τ is varied from 0 to 80 and the signal-to-noise ratio (SNR) $\frac{\|\mathbf{x}_0\|_2^2}{\sigma^2}$ is varied from 1 to 10^4. We observe that, for high SNR ($\sigma^2 \leq 10^{-3}$), the analytical prediction matches with simulation. Furthermore, the lower SNR curves are upper bounded by the high SNR curves. This behavior is fully consistent with what one would expect from Theorem 3.1 and Formula 1.

3.2.4 ℓ_2^2-LASSO

Definition 3.2 (Mapping Function) *For any $\lambda \in \mathscr{R}_{ON}$, define*

$$map(\lambda) = \lambda \cdot \frac{m - \mathbf{D}_f(\mathbf{x}_0, \lambda) - \mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}. \tag{3.20}$$

Theorem 3.3 (Properties of map$(\cdot)$) *Assume $m > \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. The function $map(\cdot) : \mathscr{R}_{ON} \to \mathbb{R}^+$ is strictly increasing, continuous and bijective. Thus, its inverse function $map^{-1}(\cdot) : \mathbb{R}^+ \to \mathscr{R}_{ON}$ is well defined.*

Formula 1 (Conjecture on the NSE of ℓ_2^2-LASSO) *Assume $(1 - \varepsilon_L)m \geq \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) \geq \varepsilon_L m$ for a constant $\varepsilon_L > 0$ and m is sufficiently large. For any value of the penalty parameter $\tau > 0$, we claim that, the expression,*

$$\frac{\mathbf{D}_f(\mathbf{x}_0, map^{-1}(\tau))}{m - \mathbf{D}_f(\mathbf{x}_0, map^{-1}(\tau))},$$

provides a good prediction of the NSE $\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|_2^2}{\sigma^2}$ for sufficiently small σ. Furthermore, we believe that the same expression upper bounds the NSE for arbitrary values of σ.

3.2.5 Converse Results

Definition 3.3 *A function $f(\cdot):\mathbb{R}^n\to\mathbb{R}$ is called Lipschitz continuous if there exists a constant $L>0$ such that, for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, we have $|f(\mathbf{x})-f(\mathbf{y})|\leq L\|\mathbf{x}-\mathbf{y}\|_2$.*

Remark: Any norm in $\mathbb{R}^n$ is Lipschitz continuous [186].

Theorem 3.4 (Failure of Robust Recovery) *Let $f(\cdot)$ be a Lipschitz continuous convex function Assume $m<\mathbf{D}(cone(\partial f(\mathbf{x}_0)))$. Then, for any $C_{max}>0$, there exists a positive number $\sigma_0:=\sigma_0(m,n,f,\mathbf{x}_0,C_{max})$ such that, if $\sigma\leq\sigma_0$, with probability $1-8\exp(-\frac{(\mathbf{D}(cone(\partial f(\mathbf{x}_0)))-m)^2}{4n})$, we have,*

$$\frac{\|\mathbf{x}^*_{\ell_2}(\mathbf{A},\mathbf{z})-\mathbf{x}_0\|_2^2}{\sigma^2}\geq C_{max},\quad \text{and}\quad \frac{\|\mathbf{x}^*_{\ell_2^2}(\mathbf{A},\mathbf{z})-\mathbf{x}_0\|_2^2}{\sigma^2}\geq C_{max}. \tag{3.21}$$

3.2.6 Remarks

A detailed discussion of the results follows in Section 3.3. Before this, the following remarks are in place.

- Known results in the noiseless CS problem (1.4) quantify the minimum number of measurements required for successful recovery of the signal of interest. Our Theorems 3.1 and 3.2 hold in the regime where this minimum number of measurements required grows proportional to the actual number of measurements m. As Theorem 3.4 shows, when m is less than the minimum number of measurements required, then the LASSO programs fails to stably estimate $\mathbf{x}_0$.
- In Theorem 3.2, the exponent in the probability expression grows as $\min\{m,\frac{m^2}{n}\}$. This implies that, we require m to grow at least linearly in $\sqrt{n}$.
- Theorem 3.1 suggests that the NSE of the Constrained-LASSO is maximized as $\sigma\to 0$. While we believe, the same statement is also valid for the ℓ_2- and ℓ_2^2-LASSO, we do not have a proof yet. Thus, Theorem 3.2 and Formula 1 lack this guarantee.
- As expected the NSE of the ℓ_2-LASSO depends on the particular choice of the penalty parameter λ. Theorem 3.2 sharply characterizes the NSE (in the small σ regime) for all values of the penalty parameter $\lambda\in\mathscr{R}_{\text{ON}}$. In Section 3.3 we elaborate on the behavior of the NSE for other values of the penalty parameter. Yet, the set of values $\mathscr{R}_{\text{ON}}$ is the most interesting one for several reasons, including but not limited to the following:

(a) The optimal penalty parameter λ_{best} that minimizes the NSE is in $\mathscr{R}_{\text{ON}}$.

(b) The function map$(\cdot)$ defined in Definition 3.2 proposes a bijective mapping from $\mathscr{R}_{\text{ON}}$ to $\mathbb{R}^+$. The inverse of this function effectively maps any value of the penalty parameter τ of the ℓ_2^2-LASSO to a particular value in $\mathscr{R}_{\text{ON}}$. Following this mapping, the exact characterization of the NSE of the ℓ_2-LASSO for $\lambda \in \mathscr{R}_{\text{ON}}$, translates (see Formula 1) to a prediction of the NSE of the ℓ_2^2-LASSO for any $\tau \in \mathbb{R}^+$.

- We don't have a rigorous proof of Formula 1. Yet, we provide partial justification and explain the intuition behind it in Section 3.12. Section 3.12 also shows that, when $m > \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, ℓ_2^2-LASSO will stably recover $\mathbf{x}_0$ for any value of $\tau > 0$, which is consistent with Formula 1. See also the discussion in Section 3.3. We, also, present numerical simulations that support the validity of the claim.
- Theorem 3.4 proves that both in the ℓ_2- and ℓ_2^2-LASSO problems, the estimation error does not grow proportionally to the noise level σ, when the number of measurements is not large enough. This result can be seen as a corollary of Theorem 1 of [4]. A result of similar nature holds for the C-LASSO, as well. For the exact statement of this result and the proofs see Section 3.13.

3.2.7 Organization of the Chapter

Section 3.3 contains a detailed discussion on our results and on their interpretation. Sections 3.4 and 3.5 contain the technical details of the framework as it was summarized in Section 3.1.6. In Sections 3.6 and 3.7, we prove the two parts of Theorem 3.1 on the NSE of the C-LASSO. Section 3.8 analyzes the ℓ_2-LASSO and Section 3.9 proves Theorem 3.2 regarding the NSE over $\mathscr{R}_{\text{ON}}$. Section 3.12 discusses the mapping between ℓ_2 and ℓ_2^2-LASSO, proves Theorem 3.3 and motivates Formula 1. In Section 3.13 we focus on the regime where robust estimation fails and prove Theorem 3.4. Simulation results presented in Section 3.14 support our analytical predictions. Finally, directions for future work are discussed in Section 3.15. Some of the technical details are deferred to the end of the chapter as "Further Proofs".

3.3 Discussion of the Results

This section contains an extended discussion on the results of this work. We elaborate on their interpretation and implications.

3.3.1 C-LASSO

We are able to characterize the estimation performance of the Constrained-LASSO in (3.3) solely based on $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. Whenever $m > \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, for sufficiently small σ, we prove that,

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \approx \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}. \tag{3.22}$$

Furthermore, (3.22) holds for arbitrary values of σ when $\approx$ is replaced with $\lesssim$. Observe in (3.22) that as m approaches $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, the NSE increases and when $m = \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, NSE $= \infty$. This behavior is not surprising as when $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, one cannot even recover $\mathbf{x}_0$ from noiseless observations via (1.4) hence it is futile to expect noise robustness. For purposes of illustration, notice that (3.22) can be further simplified for certain regimes as follows:

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \approx \begin{cases} 1 & \text{when} \quad m = 2\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))), \\ \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m} & \text{when} \quad m \gg \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))). \end{cases}$$

3.3.1.1 Relation to Proximal Denoising

We want to compare the NSE of the C-LASSO in (3.3) to the MSE risk of the constrained proximal denoiser in (3.2). For a fair comparison, the average signal power $\mathbb{E}[\|\mathbf{A}\mathbf{x}_0\|_2^2]$ in (3.3) should be equal to $\|\mathbf{x}_0\|_2^2$. This is the case for example when $\mathbf{A}$ has independent $\mathcal{N}(0, \frac{1}{m})$ entries. This is equivalent to amplifying the noise variance to $m\sigma^2$ while still normalizing the error term $\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2$ by σ^2. Thus, in this case, the formula (3.22) for the NSE is multiplied by m to result in $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \cdot \frac{m}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$ (see Section 3.3.5 for further explanation). Now, let us compare this with the results known for proximal denoising. There [49, 167], it is known that the normalized MSE is maximized when $\sigma \to 0$ and is equal to $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. Hence, we can conclude that the NSE of the LASSO problem is amplified compared to the corresponding quantity of proximal denoising by a factor of $\frac{m}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} > 1$. This factor can be interpreted as the penalty paid in the estimation error for using linear measurements.

3.3.2 ℓ_2-LASSO

Characterization of the NSE of the ℓ_2-LASSO is more involved than that of the NSE of the C-LASSO. For this problem, choice of λ naturally plays a critical role. We characterize three distinct "regions of operation"

of the ℓ_2-LASSO, depending on the particular value of λ.

3.3.2.1 Regions Of Operation

First, we identify the regime in which the ℓ_2-LASSO can robustly recover $\mathbf{x}_0$. In this direction, the number of measurements should be large enough to guarantee at least noiseless recovery in (1.4), which is the case when $m > \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ [4,50]. To translate this requirement in terms of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, recall Proposition 2.6 and Lemma 3.4, and define λ_{best} to be the *unique* minimizer of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ over $\lambda \in \mathbb{R}^+$. We, then, write the regime of interest as $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$.

Next, we identify three important values of the penalty parameter λ, needed to describe the distinct regions of operation of the estimator.

a) λ_{best} : We show that λ_{best} is optimal in the sense that the NSE is minimized for this particular choice of the penalty parameter. This also explains the term "best" we associate with it.

b) λ_{max} : Over $\lambda \geq \lambda_{\text{best}}$, the equation $m = \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ has a unique solution. We denote this solution by λ_{max}. For values of λ larger than λ_{max}, we have $m \leq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$.

c) λ_{crit} : Over $0 \leq \lambda \leq \lambda_{\text{best}}$, if $m \leq n$, the equation $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) = \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ has a unique solution which we denote λ_{crit}. Otherwise, it has no solution and $\lambda_{\text{crit}} := 0$.

Based on the above definitions, we recognize the three distinct regions of operation of the ℓ_2-LASSO, as follows,

a) $\mathscr{R}_{\text{ON}} = \{\lambda \in \mathbb{R}^+ \big| \lambda_{\text{crit}} < \lambda < \lambda_{\text{max}}\}$.

b) $\mathscr{R}_{\text{OFF}} = \{\lambda \in \mathbb{R}^+ \big| \lambda \leq \lambda_{\text{crit}}\}$.

c) $\mathscr{R}_{\infty} = \{\lambda \in \mathbb{R}^+ \big| \lambda \geq \lambda_{\text{max}}\}$.

See Figure 3.4 for an illustration of the definitions above and Section 3.8 for the detailed proofs of the statements.

3.3.2.2 Characterizing the NSE in each Region

Our main result on the ℓ_2-LASSO is for the region $\mathscr{R}_{ON}$ as stated in Theorem 3.2. We also briefly discuss on our observations regarding $\mathscr{R}_{\text{OFF}}$ and $\mathscr{R}_{\infty}$:

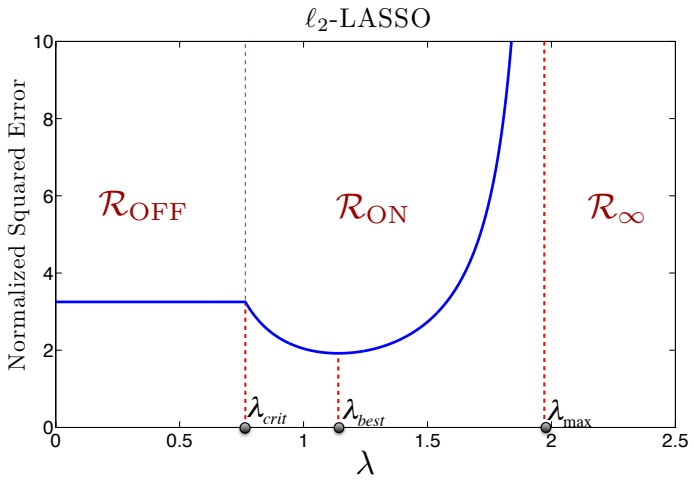


Figure 3.3: We consider the ℓ_1-penalized ℓ_2-LASSO problem for a k sparse signal in $\mathbb{R}^n$. x-axis is the penalty parameter λ. For $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$, we have $\lambda_{\text{crit}} \approx 0.76$, $\lambda_{\text{best}} \approx 1.14$, $\lambda_{\max} \approx 1.97$.

- $\mathscr{R}_{\text{OFF}}$: For $\lambda \in \mathscr{R}_{\text{OFF}}$, we empirically observe that the LASSO estimate $\mathbf{x}^*_{\ell_2}$ satisfies $\mathbf{y} = \mathbf{A}\mathbf{x}^*_{\ell_2}$ and the optimization (3.4) reduces to:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}, \tag{3.23}$$

  which is the standard approach to solving the noiseless linear inverse problems (recall (1.4)). We prove that this reduction is indeed true for values of λ sufficiently small (see Lemma 3.25), while our empirical observations suggest that the claim is valid for all $\lambda \in \mathscr{R}_{\text{OFF}}$. Proving the validity of the claim would show that when $\sigma \to 0$, the NSE is $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{crit}})}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{crit}})}$, for all $\lambda \in \mathscr{R}_{\text{OFF}}$. Interestingly, this would also give the NSE formula for the particularly interesting problem (3.23). Simulation results in Section 3.14 validate the claim.

- $\mathscr{R}_{\text{ON}}$: Begin with observing that $\mathscr{R}_{\text{ON}}$ is a nonempty and open interval. In particular, $\lambda_{\text{best}} \in \mathscr{R}_{\text{ON}}$ since $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})$. We prove that for all $\lambda \in \mathscr{R}_{\text{ON}}$ and σ is sufficiently small,

$$\frac{\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|_2^2}{\sigma^2} \approx \frac{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{m - \mathbf{D}(\lambda \partial f(\mathbf{x}_0))}. \tag{3.24}$$

  Also, empirical observations suggest that 3.24 holds for arbitrary σ when $\approx$ replaced with $\lesssim$. Finally, we should note that the NSE formula $\frac{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}$ is a convex function of λ over $\mathscr{R}_{\text{ON}}$.

- $\mathscr{R}_{\infty}$: Empirically, we observe that the stable recovery of $\mathbf{x}_0$ is not possible for $\lambda \in \mathscr{R}_{\infty}$.

3.3.2.3 Optimal Tuning of the Penalty Parameter

It is not hard to see that the formula in (3.24) is strictly increasing in $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. Thus, when $\sigma \to 0$, the NSE achieves its minimum value when the penalty parameter is set to λ_{best}. Now, recall that $\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and compare the formulae in (3.22) and (3.24), to conclude that the C-LASSO and ℓ_2-LASSO can be related by choosing $\lambda = \lambda_{\text{best}}$. In particular, we have,

$$\frac{\|\mathbf{x}^*_{\ell_2}(\lambda_{\text{best}}) - \mathbf{x}_0\|_2^2}{\sigma^2} \approx \frac{\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})}{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})} \approx \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} \approx \frac{\|\mathbf{x}^*_c - \mathbf{x}_0\|_2^2}{\sigma^2}. \tag{3.25}$$

3.3.3 ℓ_2^2-LASSO

3.3.3.1 Connection to ℓ_2-LASSO

We propose a mapping between the penalty parameters λ of the ℓ_2-LASSO program (3.4) and τ of the ℓ_2^2-LASSO program (3.5), for which the NSE of the two problems behaves the same. The mapping function was defined in Definition 3.2. Observe that $\text{map}(\lambda)$ is well-defined over the region $\mathscr{R}_{\text{ON}}$, since $m > \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) > \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ for all $\lambda \in \mathscr{R}_{\text{ON}}$. Theorem 3.3 proves that $\text{map}(\cdot)$ defines a bijective mapping from $\mathscr{R}_{\text{ON}}$ to $\mathbb{R}^+$. Other useful properties of the mapping function include the following:

- $\text{map}(\lambda_{\text{crit}}) = 0$,
- $\lim_{\lambda \to \lambda_{\max}} \text{map}(\lambda) = \infty$,

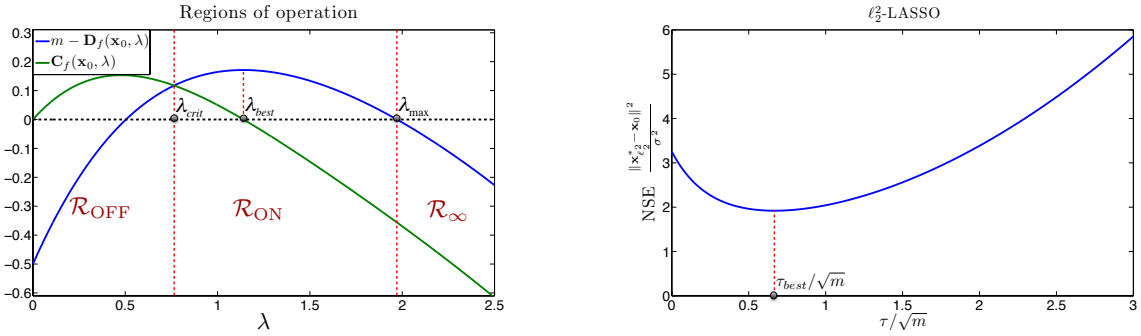
Section 3.12 proves these properties and more, and contains a short technical discussion that motivates the proposed mapping function.

3.3.3.2 Proposed Formula

We use the mapping function in (3.20) to translate our results on the NSE of the ℓ_2-LASSO over $\mathscr{R}_{\text{ON}}$ (see formula (3.24)) to corresponding results on the ℓ_2^2-LASSO for $\tau \in \mathbb{R}^+$. Assume $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})$. We suspect that for any $\tau > 0$,

$$\frac{\mathbf{D}_f(\mathbf{x}_0, \text{map}^{-1}(\tau))}{m - \mathbf{D}_f(\mathbf{x}_0, \text{map}^{-1}(\tau))},$$

accurately characterizes $\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|_2^2}{\sigma^2}$ for sufficiently small σ, and upper bounds $\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|_2^2}{\sigma^2}$ for arbitrary σ.



(a) (b)

Figure 3.4: We consider the exact same setup of Figure 3.3. a) We plot $m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ as a function of λ to illustrate the important penalty parameters $\lambda_{\text{crit}}, \lambda_{\text{best}}, \lambda_{\max}$ and the regions of operation $\mathscr{R}_{\text{OFF}}, \mathscr{R}_{\text{ON}}, \mathscr{R}_{\infty}$. b) We plot the ℓ_2^2-LASSO error as a function of $\frac{\tau}{\sqrt{m}}$ by using the map$(\cdot)$ function. The normalization is due to the fact that τ grows linearly in $\sqrt{m}$.

3.3.3.3 A rule of thumb for the optimal penalty parameter

Formula 1 provides a simple recipe for computing the optimal value of the penalty parameter, which we call τ_{best}. Recall that λ_{best} minimizes the error in the ℓ_2-LASSO. Then, the proposed mapping between the two problems, suggests that $\tau_{best} = \text{map}(\lambda_{\text{best}})$. To evaluate map$(\lambda_{\text{best}})$ we make use of Lemma 3.4 and the fact that $\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda} = -\frac{2}{\lambda}\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ for all $\lambda \geq 0$. Combine this with the fact that λ_{best} is the unique minimizer of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, to show that $\mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{best}}) = 0$, and to conclude with,

$$\tau_{\text{best}} = \lambda_{\text{best}}\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})}. \tag{3.26}$$

As a last comment, (3.26) simplifies even further if one uses the fact $\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, which is valid under reasonable assumptions, [4, 101, 167]. In this case, $\tau_{best} \approx \lambda_{\text{best}}\sqrt{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$.

3.3.4 Closed Form Calculations of the Formulae

Table 3.2 summarizes the formulae for the NSE of the three versions of the LASSO problem. While simple and concise, it may appear to the reader that the formulae are rather abstract, because of the presence of $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ ($\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ is also implicitly involved in the calculation of map$^{-1}(\cdot)$) which were introduced to capture the convex geometry of the problem. However, as discussed here, for certain critical regularizers $f(\cdot)$, one can calculate (tight) upper bounds or even explicit formulas for these quantities. For example, for the estimation of a k-sparse signal $\mathbf{x}_0$ with $f(\cdot) = \|\cdot\|_1$, it has been shown that

	Normalized Squared Error
C-LASSO	$\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$
ℓ_2-LASSO	$\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ for $\lambda \in \mathscr{R}_{\text{ON}}$
ℓ_2^2-LASSO	$\frac{\mathbf{D}_f(\mathbf{x}_0,\text{map}^{-1}(\tau))}{m-\mathbf{D}_f(\mathbf{x}_0,\text{map}^{-1}(\tau))}$ for $\tau \in \mathbb{R}^+$

Table 3.2: Summary of formulae for the NSE.

$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \lesssim 2k(\log\frac{n}{k}+1)$. Substituting this into the formula for the NSE of the C-LASSO results in the "closed-form" upper bound given in (3.7), i.e. one expressed only in terms of m,n and k. Analogous results have been derived [50,101,163,191] for other well-known signal models as well, including low rankness (see (3.8)) and block-sparsity (see (3.9)). The first row of Table 3.3 summarizes some of the results for $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ found in the literature (see [50,101]). The second row provides our closed form results on $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ when λ is sufficiently large. The reader will observe that, by setting λ to its lower bound in the second row, one approximately obtains the corresponding result in the first row. For a related discussion on $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and closed form bounds, the reader is referred to [101]. The derivation of these results can be found in Section A.7 of the Appendix. In the same section, we also provide exact formulas for $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ for the same signal models. Based on those formulas and Table 3.3, one simply needs to substitute $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ or $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ with their corresponding value to reach the error bounds. We should emphasize that, examples are not limited to the ones discussed here (see for instance [50]).

	k-**sparse,** $\mathbf{x}_0 \in \mathbb{R}^n$	**Rank** r**,** $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$	k-**block sparse,** $\mathbf{x}_0 \in \mathbb{R}^{tb}$
$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$	$2k(\log\frac{n}{k}+1)$	$6dr$	$4k(\log\frac{t}{k}+b)$
$\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$	$(\lambda^2+3)k$ for $\lambda \geq \sqrt{2\log\frac{n}{k}}$	$\lambda^2 r+2d(r+1)$ for $\lambda \geq 2\sqrt{d}$	$(\lambda^2+b+2)k$ for $\lambda \geq \sqrt{b}+\sqrt{2\log\frac{t}{k}}$

Table 3.3: Closed form upper bounds for $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ ([50,101]) and $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ corresponding to (3.7), (3.8) and (3.9).

It follows from this discussion, that establishing new and tighter analytic bounds for $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ for more regularizers f is certainly an interesting direction for future research. In the case where such analytic bounds do not already exist in literature or are hard to derive, one can numerically estimate $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ once there is an available characterization of the set of subdifferentials $\partial f(\mathbf{x}_0)$. More in detail, it is not hard to show that, when $\mathbf{h} \sim \mathcal{N}(0,\mathbf{I}_n)$, $\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$

concentrates nicely around $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ (see Lemma A.3) . Hence to compute $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$:

(a) draw a vector $\mathbf{h}\sim\mathcal{N}(0,\mathbf{I}_n)$,

(b) return the solution of the convex program $\min_{\mathbf{s}\in\partial f(\mathbf{x}_0)}\|\mathbf{h}-\lambda\mathbf{s}\|_2^2$.

Computing $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ can be built on the same recipe by writing $\text{dist}^2(\mathbf{h},\text{cone}(\partial f(\mathbf{x}_0)))$ as $\min_{\lambda\geq 0,\mathbf{s}\in\partial f(\mathbf{x}_0)}\|\mathbf{h}-\lambda\mathbf{s}\|_2^2$.

Summing up, our proposed formulae for the NSE of the LASSO problems can be effectively calculated, either analytically or numerically.

3.3.5 Translating the Results

Until this point, we have considered the scenario, in which the measurement matrix $\mathbf{A}$ has independent standard normal entries, and the noise vector $\mathbf{z}$ is equal to $\sigma\mathbf{v}$ with $\mathbf{v}\sim\mathcal{N}(0,\mathbf{I}_m)$. In related literature, the entries of $\mathbf{A}$ are often assumed to have variance $\frac{1}{m}$ or $\frac{1}{n}$, [13, 14, 33]. For example, a variance of $\frac{1}{m}$ ensures that in expectation $\|\mathbf{Ax}\|_2^2$ is same as $\|\mathbf{x}\|_2^2$. Hence, it is important to understand, how our setting can be translated to those. To distinguish our setup from the "non-unit variance" setup, we introduce the "non-unit variance" variables $\mathbf{A}',\sigma',\lambda'$ and τ'. Let entries of $\mathbf{A}'$ have variance $\frac{1}{m}$ and consider the ℓ_2-LASSO problem with these new variables, which can be equivalently written as,

$$\min_{\mathbf{x}}\|\mathbf{A}'\mathbf{x}_0+\sigma'\mathbf{v}-\mathbf{A}'\mathbf{x}\|_2+\lambda' f(\mathbf{x}).$$

Multiplying the objective with $\sqrt{m}$, we obtain,

$$\min_{\mathbf{x}}\|\sqrt{m}\mathbf{A}'\mathbf{x}_0+\sqrt{m}\sigma'\mathbf{v}-\sqrt{m}\mathbf{A}'\mathbf{x}\|_2+\sqrt{m}\lambda' f(\mathbf{x}).$$

Observe that, $\sqrt{m}\mathbf{A}'$ is now statistically identical to $\mathbf{A}$. Hence, Theorem 3.2 is applicable under the mapping $\sigma\leftarrow\sqrt{m}\sigma'$ and $\lambda\leftarrow\sqrt{m}\lambda'$. Consequently, the NSE formula for the new setting for $\sqrt{m}\lambda'\in\mathscr{R}_{\text{ON}}$ can be given as,

$$\frac{\|\mathbf{x}^*_{\ell_2}-\mathbf{x}_0\|_2^2}{(\sqrt{m}\sigma')^2}=\frac{\|\mathbf{x}^*_{\ell_2}-\mathbf{x}_0\|_2^2}{\sigma^2}\lesssim\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}=\frac{\mathbf{D}_f(\mathbf{x}_0,\sqrt{m}\lambda')}{m-\mathbf{D}_f(\mathbf{x}_0,\sqrt{m}\lambda')}.$$

Identical arguments for the Constrained-LASSO and ℓ_2^2-LASSO results in the following NSE formulas,

$$\frac{\|\mathbf{x}^*_c-\mathbf{x}_0\|_2^2}{m\sigma'^2}\lesssim\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}\quad\text{and}\quad\frac{\|\mathbf{x}^*_{\ell_2^2}-\mathbf{x}_0\|_2^2}{m\sigma'^2}\lesssim\frac{\mathbf{D}_f(\mathbf{x}_0,\text{map}^{-1}(\sqrt{m}\tau'))}{m-\mathbf{D}_f(\mathbf{x}_0,\text{map}^{-1}(\sqrt{m}\tau'))}.$$

In general, reducing the signal power $\|\mathbf{A}\mathbf{x}_0\|_2^2$ by a factor of m, amplifies the proposed NSE upper bound by m times and the penalty parameters should be mapped as $\tau \longleftrightarrow \sqrt{m}\tau'$ and $\lambda \longleftrightarrow \sqrt{m}\lambda'$.

3.4 Applying Gaussian Min-Max Theorem

First, we introduce the basic notation that is used throughout the technical analysis of our results. Some additional notation, specific to the subject of each particular section is introduced later therein. To make explicit the variance of the noise vector $\mathbf{z}$, we denote $\mathbf{z} = \sigma\mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(0, \mathbf{I}_m)$. Also, we reserve the variables $\mathbf{h}$ and $\mathbf{g}$ to denote i.i.d. Gaussian vectors in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively. In similar flavor, reserve the variable $\mathbf{s}$ to describe the subgradients of f at $\mathbf{x}_0$. Finally, the Euclidean unit ball and unit sphere are respectively denoted as

$$\mathscr{B}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq 1\} \quad \text{and} \quad \mathscr{S}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1\}.$$

3.4.1 Introducing the Error Vector

For each candidate solution $\mathbf{x}$ of the LASSO algorithm, denote $\mathbf{w} = \mathbf{x} - \mathbf{x}_0$. Solving for $\mathbf{w}$ is clearly equivalent to solving for $\mathbf{x}$, but simplifies considerably the presentation of the analysis. Under this notation, $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2$. Furthermore, it is convenient to subtract the constant factor $\lambda f(\mathbf{x}_0)$ from the objective function of the LASSO problem and their approximations. In this direction, define the following *"perturbation" functions*:

$$f_p(\mathbf{w}) = f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0), \tag{3.27}$$

$$\hat{f}_p(\mathbf{w}) = \hat{f}(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0) = \sup_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}. \tag{3.28}$$

Then, the ℓ_2-LASSO will write as

$$\mathbf{w}^*_{\ell_2} = \arg\min_{\mathbf{w}} \{\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \lambda f_p(\mathbf{w})\}. \tag{3.29}$$

and the C-LASSO as

$$\mathbf{w}_c^* = \arg\min_{\mathbf{w}} \ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2$$
$$\text{s.t.} \quad f_p(\mathbf{w}) \leq 0.$$

or, equivalently,

$$\mathbf{w}_c^* = \arg\min_{\mathbf{w}} \left\{ \ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \max_{\lambda \geq 0} \lambda f_p(\mathbf{w}) \ \right\}. \tag{3.30}$$

3.4.2 The Approximate LASSO Problem

In Section 3.1, and in particular in (3.12) we introduced the approximated ℓ_2-LASSO problem. We repeat the definition here, and also, we define accordingly the approximate C-LASSO. The approximated ℓ_2-LASSO writes:

$$\hat{\mathbf{w}}_{\ell_2} = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \lambda \hat{f}_p(\mathbf{w}) \right\}. \tag{3.31}$$

Similarly, the approximated C-LASSO writes

$$\hat{\mathbf{w}}_c \ = \ \arg\min_{\mathbf{w}} \left\{ \ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \max_{\lambda \geq 0} \lambda \hat{f}_p(\mathbf{w}) \ \right\}. \tag{3.32}$$

Denote $\hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v})$ and $\hat{\mathscr{F}}_{\ell 2}(\mathbf{A},\mathbf{v})$ the optimal costs of problems (3.32) and (3.31), respectively. Note our convention to use the symbol "$\hat{}$" over variables that are associated with the approximate problems. To distinguish, we use the symbol "*" for the variables associated with the original problems.

3.4.3 Simplifying the LASSO objective through Gaussian Min-Max Theorem

Section 3.1.6 introduced the technical framework. Key feature in this framework is the application of Gordon's Theorem. In particular, we apply the "modified Gordon's Lemma" Proposition 2.7 three times: once each for the purposes of the lower bound, the upper bound and the deviation analysis. Each application results in a corresponding simplified problem, which we call *"key optimization"*. The analysis is carried out for that latter one as opposed to the original and more complex LASSO problem. In this Section, we show the details of applying Gordon's Theorem and we identify the corresponding key optimizations. Later, in

Section 3.5, we focus on the approximate LASSO problem and we show that in that case, the key optimizations are amenable to detailed analysis.

To avoid unnecessary repetitions, we treat the original and approximate versions of both the C-LASSO and the ℓ_2-LASSO, in a common framework, by defining the following problem:

$$\mathscr{F}(\mathbf{A},\mathbf{v}) = \min_{\mathbf{w}} \{ \ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + p(\mathbf{w}) \ \}, \tag{3.33}$$

where $p : \mathbf{R}^n \to \mathbf{R} \cup \infty$ is a proper convex function [182]. Choose the penalty function $p(\cdot)$ in the generic formulation (3.33) accordingly to end up with (3.29), (3.30), (3.31) or (3.32). To retrieve (3.30) and (3.32), choose $p(\mathbf{w})$ as the indicator function of the sets $\{\mathbf{w}|f_p(\mathbf{w}) \leq 0\}$ and $\left\{\mathbf{w}|\hat{f}_p(\mathbf{w}) \leq 0\right\}$ [23].

3.4.3.1 Lower Bound

The following corollary is a direct application of Proposition 2.7 to $\mathscr{F}(\mathbf{A},\mathbf{v})$ in (3.33).

Corollary 3.2 *Let* $\mathbf{g} \sim \mathscr{N}(0,\mathbf{I}_m)$, $\mathbf{h} \sim \mathscr{N}(0,\mathbf{I}_n)$ *and* $h \sim \mathscr{N}(0,1)$ *and assume all* $\mathbf{g},\mathbf{h},h$ *are independently generated. Let*

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}} \left\{ \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} + p(\mathbf{w}) \right\}. \tag{3.34}$$

Then, for any $c \in \mathbb{R}$*:*

$$\mathbb{P}\left(\ \mathscr{F}(\mathbf{A},\mathbf{v}) \geq c \ \right) \geq 2 \cdot \mathbb{P}\left(\ \mathscr{L}(\mathbf{g},\mathbf{h}) - h\sigma \geq c \ \right) - 1.$$

Proof: Notice that $\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 = \|\mathbf{A}_\mathbf{v}\mathbf{w}_\sigma\|_2$, where $\mathbf{A}_\mathbf{v} := [\mathbf{A} \ \ -\mathbf{v}]$ is a matrix with i.i.d. standard normal entries of size $m \times (n+1)$ and $\mathbf{w}_\sigma = [\mathbf{w}^T \ \sigma]^T \in \mathbb{R}^{n+1}$. Apply Proposition 2.7, with $\mathbf{x} = \mathbf{w}_\sigma$, $\Phi_1 = \{\mathbf{w}_\sigma \big| \mathbf{w} \in \mathbb{R}^n\}$, $\Phi_2 = \mathscr{S}^{m-1}$, $\mathbf{G} = \mathbf{A}_\mathbf{v}$, $\psi(\mathbf{w}_\sigma) = p(\mathbf{w})$. Further perform the trivial optimizations over $\mathbf{a}$ on both sides of the inequality. Namely, $\max_{\|\mathbf{a}\|_2=1} \mathbf{a}^T\mathbf{A}_\mathbf{z}[\mathbf{w}^T \ \sigma]^T = \|\mathbf{A}_\mathbf{z}\mathbf{w}_\sigma\|_2$ and, $\max_{\|\mathbf{a}\|_2=1} \mathbf{g}^T\mathbf{a} = \|\mathbf{g}\|_2$. ■

3.4.3.2 Upper Bound

Similar to the lower bound derived in the previous section, we derive an upper bound for $\mathscr{F}(\mathbf{A},\mathbf{v})$. For this, we need to apply Gordon's Theorem to $-\mathscr{F}(\mathbf{A},\mathbf{v})$ and use the dual formulation of it. Lemma A.3 in the

Appendix shows that the dual of the minimization in (3.33) can be written as

$$-\mathscr{F}(\mathbf{A},\mathbf{v}) = \min_{\|\boldsymbol{\mu}\|_2\leq 1} \max_{\mathbf{w}} \left\{ \boldsymbol{\mu}^T (\mathbf{A}\mathbf{w} - \sigma\mathbf{v}) - p(\mathbf{w}) \right\}. \tag{3.35}$$

Proposition 2.7 requires the set over which maximization is performed to be compact. We thus apply Proposition 2.7 to the restricted problem,

$$\min_{\|\boldsymbol{\mu}\|_2\leq 1} \max_{\|\mathbf{w}\|_2\leq C_{up}} \left\{ \boldsymbol{\mu}^T (\mathbf{A}\mathbf{w} - \sigma\mathbf{v}) - p(\mathbf{w}) \right\}.$$

Notice, that this still gives a valid lower bound to $-\mathscr{F}(\mathbf{A},\mathbf{v})$ since the optimal cost of this latter problem is no larger than $-\mathscr{F}(\mathbf{A},\mathbf{v})$. In Section 3.5, we will choose C_{up} so that the resulting lower bound is as tight as possible.

Corollary 3.3 *Let* $\mathbf{g}\sim\mathscr{N}(0,\mathbf{I}_m)$, $\mathbf{h}\sim\mathscr{N}(0,\mathbf{I}_n)$ *and* $h\sim\mathscr{N}(0,1)$ *and assume all* $\mathbf{g},\mathbf{h},h$ *are independently generated. Let,*

$$\mathscr{U}(\mathbf{g},\mathbf{h}) = -\min_{\|\boldsymbol{\mu}\|_2\leq 1} \max_{\|\mathbf{w}\|_2\leq C_{up}} \left\{ \sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\, \mathbf{g}^T\boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2 \mathbf{h}^T\mathbf{w} - p(\mathbf{w}) \right\}. \tag{3.36}$$

Then, for any $c\in\mathbb{R}$*:*

$$\mathbb{P}\left(\mathscr{F}(\mathbf{A},\mathbf{v}) \leq c \right) \geq 2\cdot\mathbb{P}\left(\mathscr{U}(\mathbf{g},\mathbf{h}) - \min_{0\leq\alpha\leq 1} \alpha\sigma h \leq c \right) - 1.$$

Proof: Similar to the proof of Corollary 3.2 write $\|\sigma\mathbf{v}-\mathbf{A}\mathbf{w}\|_2 = \|\mathbf{A}_{\mathbf{v}}\mathbf{w}_\sigma\|_2$. Then, apply the modified Gordon's Theorem 2.7, with $\mathbf{x}=\boldsymbol{\mu}$, $\alpha=\mathbf{w}_\sigma$, $\Phi_1=\mathscr{B}^{m-1}$, $\Phi_2=\left\{\mathbf{w}_\sigma \mid \frac{1}{C_{up}}\mathbf{w}\in\mathscr{B}^{n-1}\right\}$, $\mathbf{G}=\mathbf{A}_{\mathbf{v}}$, $\psi(\mathbf{w}_\sigma)=p(\mathbf{w})$, to find that for any $c\in\mathbb{R}$:

$$\begin{aligned}\mathbb{P}\left(-\mathscr{F}(\mathbf{A},\mathbf{v}) \geq -c \right) &\geq 2\cdot\mathbb{P}\left(\min_{\|\boldsymbol{\mu}\|_2\leq 1} \max_{\|\mathbf{w}\|_2\leq C_{up}} \{ \sqrt{C_{up}^2+\sigma^2}\, \mathbf{g}^T\boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2\mathbf{h}^T\mathbf{w} - p(\mathbf{w}) + \|\boldsymbol{\mu}\|_2\sigma h\} \geq -c \right) - 1 \\ &\geq 2\mathbb{P}\left(-\mathscr{U}(\mathbf{g},\mathbf{h}) + \min_{\|\mu\|_2\leq 1} \|\boldsymbol{\mu}\|_2\sigma h \geq -c \right) - 1.\end{aligned}$$

■

3.4.3.3 Deviation Analysis

Of interest in the deviation analysis of the LASSO problem (cf. Step 4 in Section 3.1.6) is the analysis of a restricted version of the *LASSO* problem, namely

$$\min_{\|\mathbf{w}\|_2 \in S_{dev}} \{\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + p(\mathbf{w})\} \tag{3.37}$$

where

$$S_{dev} := \left\{ \ell \mid \left| \frac{\ell}{C_{dev}} - 1 \right| \geq \delta_{dev} \right\}.$$

$\delta_{dev} > 0$ is any arbitrary small constant and $C_{dev} > 0$ a constant that will be chosen carefully for the purpose of the deviation analysis . We establish a high probability lower bound for (3.37). As usual, we apply Proposition 2.7 to our setup, to conclude the following.

Corollary 3.4 *Let* $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$, $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *and* $h \sim \mathcal{N}(0,1)$ *and assume all* $\mathbf{g}, \mathbf{h}, h$ *are independently generated. Let*

$$\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\|\mathbf{w}\|_2 \in S_{dev}} \left\{ \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \|\mathbf{g}\|_2 - \mathbf{h}^T \mathbf{w} + p(\mathbf{w}) \right\}. \tag{3.38}$$

Then, for any $c \in \mathbb{R}$*:*

$$\mathbb{P}\left(\min_{\|\mathbf{w}\|_2 \in S_{dev}} \{\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + p(\mathbf{w})\} \geq c \right) \geq 2 \cdot \mathbb{P}\left(\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) - h\sigma \geq c \right) - 1.$$

Proof: Follows from Proposition 2.7 following exactly the same steps as in the proof of Corollary 3.2. ■

The reader will observe that $\mathscr{L}$ is a special case of $\mathscr{L}_{dev}$ where $S_{dev} = \mathbb{R}^+$.

3.4.3.4 Summary

We summarize the results of Corollaries 3.2, 3.3 and 3.4 in Lemma 3.5. Adding to a simple summary, we perform a further simplification of the corresponding statements. In particular, we discard the "distracting" term σh in Corollaries 3.2 and 3.4, as well as the term $\min_{0 \leq \alpha \leq 1} \alpha\sigma h$ in Corollary 3.3. Recall the definitions of the key optimizations $\mathscr{L}$, $\mathscr{U}$ and $\mathscr{L}_{dev}$ in (3.34), (3.36) and (3.38).

Lemma 3.5 *Let* $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$ *and* $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *be independently generated. Then, for any positive constant* $\varepsilon > 0$*, the following are true:*

1. $\mathbb{P}\left(\mathscr{F}(\mathbf{A}, \mathbf{v}) \geq c \right) \geq 2\,\mathbb{P}\left(\mathscr{L}(\mathbf{g}, \mathbf{h}) - \sigma\varepsilon\sqrt{m} \geq c \right) - 4\exp\left(-\frac{\varepsilon^2 m}{2}\right) - 1.$
2. $\mathbb{P}\left(\mathscr{F}(\mathbf{A}, \mathbf{v}) \leq c \right) \geq 2\,\mathbb{P}\left(\mathscr{U}(\mathbf{g}, \mathbf{h}) + \sigma\varepsilon\sqrt{m} \leq c \right) - 4\exp\left(-\frac{\varepsilon^2 m}{2}\right) - 1.$
3. $\mathbb{P}\left(\min_{\|\mathbf{w}\|_2 \in S_{dev}} \{\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + p(\mathbf{w})\} \geq c \right) \geq 2\,\mathbb{P}\left(\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) - \sigma\varepsilon\sqrt{m} \geq c \right) - 4\exp\left(-\frac{\varepsilon^2 m}{2}\right) - 1.$

Proof: For $h \sim \mathcal{N}(0,1)$ and all $\varepsilon > 0$,

$$\mathbb{P}\left(|h| \leq \varepsilon\sqrt{m}\right) \geq 1 - 2\exp(-\frac{\varepsilon^2 m}{2}). \tag{3.39}$$

Thus,

$$\begin{aligned}\mathbb{P}\left(\mathscr{L}(\mathbf{g}, \mathbf{h}) - h\sigma \geq c \right) &\geq \mathbb{P}\left(\mathscr{L}(\mathbf{g}, \mathbf{h}) - \varepsilon\sigma\sqrt{m} \geq c\,,\; h \leq \varepsilon\sqrt{m}\right) \\ &\geq \mathbb{P}(\mathscr{L}(\mathbf{g}, \mathbf{h}) - \varepsilon\sigma\sqrt{m} \geq c) - 2\exp(-\frac{\varepsilon^2 m}{2}).\end{aligned}$$

Combine this with Corollary 3.2 to conclude with the first statement of Lemma 3.5. The proof of the third statement of the Lemma follows the exact same steps applied this time to Corollary 3.4. For the second statement write,

$$\begin{aligned}\mathbb{P}\left(\mathscr{U}(\mathbf{g}, \mathbf{h}) - \min_{\|\mu\|_2 \leq 1} \|\boldsymbol{\mu}\|_2 \sigma h \leq c \right) &\geq \mathbb{P}\left(\mathscr{U}(\mathbf{g}, \mathbf{h}) + \sigma|h| \leq c \right) \\ &\geq \mathbb{P}\left(\mathscr{U}(\mathbf{g}, \mathbf{h}) + \varepsilon\sigma\sqrt{m} \leq c\,,\; |h| \leq \varepsilon\sqrt{m} \right),\end{aligned}$$

and use (3.39) as above. To conclude, combine with the statement of Corollary 3.3. ■

3.5 After Gordon's Theorem: Analyzing the Key Optimizations

3.5.1 Preliminaries

This section is devoted to the analysis of the three key optimizations introduced in the previous section. In particular, we focus on the approximated C-LASSO and ℓ_2-LASSO problems, for which a detailed

such analysis is tractable. Recall that the approximated C-LASSO and ℓ_2-LASSO are obtained from the generic optimization in (3.33) when substituting $p(\mathbf{w}) = \max_{\lambda\geq 0}\max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)}\mathbf{s}^T\mathbf{w} = \max_{\mathbf{s}\in\text{cone}(\partial f(\mathbf{x}_0))}\mathbf{s}^T\mathbf{w}$ and $p(\mathbf{w}) = \max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)}\mathbf{s}^T\mathbf{w}$, respectively. Considering this and recalling the definitions in (3.34), (3.36) and (3.38), we will be analyzing the following *key optimizations*,

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} + \max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{3.40a}$$

$$\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) = -\min_{\|\boldsymbol{\mu}\|_2\leq 1}\max_{\|\mathbf{w}\|_2=C_{up}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\,\mathbf{g}^T\boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2\mathbf{h}^T\mathbf{w} - \max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{3.40b}$$

$$\mathscr{L}_{dev}(\mathbf{g},\mathbf{h}) = \min_{\|\mathbf{w}\|_2\in S_{dev}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} + \max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{3.40c}$$

where $\mathscr{C}$ is taken to be either $\text{cone}(\partial f(\mathbf{x}_0))$ or $\lambda\partial f(\mathbf{x}_0)$, corresponding to the C-LASSO and ℓ_2-LASSO, respectively. Notice that in (3.40b) we have constrained the feasible set of the inner maximization to the scaled sphere rather than ball. Following our discussion, in Section 3.4.3.2 this does not affect the validity of Lemma 3.5, while it facilitates our derivations here.

To be consistent with the definitions in (3.40), which treat the key optimizations of the C-LASSO and ℓ_2-LASSO under a common framework with introducing a generic set $\mathscr{C}$, we also define

$$\hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) = \min_{\mathbf{w}}\left\{\|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\|_2 + \max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{3.41}$$

to correspond to (3.32) and (3.31), when setting $\mathscr{C} = \text{cone}(\partial f(\mathbf{x}_0))$ and $\mathscr{C} = \lambda\partial f(\mathbf{x}_0)$, respectively.

3.5.2 Some Notation

Recall the definitions in Section 2.2. Additionally, define the correlation induced by a convex and closed set $\mathscr{C}\in\mathbb{R}^n$ by,

$$\text{corr}(\mathbf{x},\mathscr{C}) := \langle \text{Proj}(\mathbf{x},\mathscr{C}), \Pi(\mathbf{x},\mathscr{C})\rangle\,.$$

Now, let $\mathbf{h}\sim\mathscr{N}(0,\mathbf{I}_n)$. The following quantities are of central interest throughout the chapter:

$$\mathbf{D}(\mathscr{C}) := \mathbb{E}\left[\,\text{dist}^2(\mathbf{h},\mathscr{C})\,\right], \tag{3.42a}$$

$$\mathbf{P}(\mathscr{C}) := \mathbb{E}\left[\,\|\text{Proj}(\mathbf{h},\mathscr{C})\|_2^2\,\right], \tag{3.42b}$$

$$\mathbf{C}(\mathscr{C}) := \mathbb{E}\left[\,\text{corr}(\mathbf{h},\mathscr{C})\,\right], \tag{3.42c}$$

where the $\mathbb{E}[\cdot]$ is over the distribution of the Gaussian vector $\mathbf{h}$. It is easy to verify that $n = \mathbf{D}(\mathscr{C}) + \mathbf{P}(\mathscr{C}) + 2\mathbf{C}(\mathscr{C})$. Under this notation,

$$\begin{aligned}\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) &= \mathbf{D}(\lambda\partial f(\mathbf{x}_0)),\\ \mathbf{C}(\lambda\partial f(\mathbf{x}_0)) &= \mathbf{C}(\lambda\partial f(\mathbf{x}_0)),\\ \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) &= \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))).\end{aligned}$$

On the same lines, define $\mathbf{P}(\lambda\partial f(\mathbf{x}_0)) := \mathbf{P}(\lambda\partial f(\mathbf{x}_0))$.

3.5.3 Analysis

We perform a detailed analysis of the three key optimization problems $\mathscr{L}$, $\hat{\mathscr{U}}$ and $\mathscr{L}_{dev}$. For each one of them we summarize the results of the analysis in Lemmas 3.6, 3.7 and 3.8 below. Each lemma includes three statements. First, we reduce the corresponding key optimization problem to a scalar optimization. Next, we compute the optimal value of this optimization in a deterministic setup. We convert this into a probabilistic statement in the last step, which is directly applicable in Lemma 3.5. Eventhough, we are eventually interested only in this last probabilistic statement, we have decided to include all three steps in the statement of the lemmas in order to provide some further intuition into how they nicely build up to the desired result. All proofs of the lemmas are deferred to Section A.4 in the Appendix.

3.5.3.1 Lower Key Optimization

Lemma 3.6 (Properties of $\mathscr{L}$) *Let $\mathbf{g} \sim \mathscr{N}(0, \mathbf{I}_m)$ and $\mathbf{h} \sim \mathscr{N}(0, \mathbf{I}_n)$ and*

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}} \left\{ \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} + \max_{\mathbf{s}\in\mathscr{C}} \mathbf{s}^T\mathbf{w} \right\}, \tag{3.43}$$

Denote $\hat{\mathbf{w}}_{low}(\mathbf{g},\mathbf{h})$ its optimal value. The following are true:

1. Scalarization: $\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\alpha\geq 0} \left\{ \sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2 - \alpha \cdot dist(\mathbf{h},\mathscr{C}) \right\}$
2. Deterministic result: *If* $\|\mathbf{g}\|_2^2 > dist(\mathbf{h},\mathscr{C})^2$, *then,*

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \sigma\sqrt{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\mathscr{C})},$$

and,

$$\|\hat{\mathbf{w}}_{low}(\mathbf{g},\mathbf{h})\|_2^2 = \sigma^2 \frac{dist^2(\mathbf{h},\mathscr{C})}{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\mathscr{C})}.$$

3. Probabilistic result: *Assume that* $m \geq \mathbf{D}(\mathscr{C}) + \varepsilon_L m$ *for some* $\varepsilon_L \geq 0$*. Then, for any* $\varepsilon > 0$*, there exist* $c_1, c_2 > 0$ *such that, for sufficiently large m,*

$$\mathbb{P}\left(\mathscr{L}(\mathbf{g},\mathbf{h}) \geq (1-\varepsilon)\sigma\sqrt{m - \mathbf{D}(\mathscr{C})}\right) \geq 1 - c_1 \exp(-c_2 m).$$

3.5.3.2 Upper Key Optimization

Lemma 3.7 (Properties of $\hat{\mathscr{U}}$) *Let* $\mathbf{g} \sim \mathscr{N}(0, \mathbf{I}_m)$*,* $\mathbf{h} \sim \mathscr{N}(0, \mathbf{I}_n)$ *and*

$$\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) = -\min_{\|\mu\|_2 \leq 1} \max_{\|\mathbf{w}\|_2 = C_{up}} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2 \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathscr{C}} \mathbf{s}^T \mathbf{w} \right\}. \tag{3.44}$$

The following hold true:

1. Scalarization: $\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) = -\min_{0 \leq \alpha \leq 1} \left\{ -\alpha \cdot \sqrt{C_{up}^2 + \sigma^2}\, \|\mathbf{g}\|_2 + C_{up} dist(\alpha \mathbf{h}, \mathscr{C}) \right\}$.

2. Deterministic result: *If* $\mathbf{h} \notin \mathscr{C}$ *and*

$$C_{up} dist(\mathbf{h},\mathscr{C}) + C_{up} \frac{corr(\mathbf{h},\mathscr{C})}{dist(\mathbf{h},\mathscr{C})} < \sqrt{C_{up}^2 + \sigma^2} \|\mathbf{g}\|_2, \tag{3.45}$$

then,

$$\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) = \sqrt{C_{up}^2 + \sigma^2} \|\mathbf{g}\|_2 - C_{up} dist(\mathbf{h},\mathscr{C}). \tag{3.46}$$

3. Probabilistic result: *Assume* $m \geq \max\{\mathbf{D}(\mathscr{C}), \mathbf{D}(\mathscr{C}) + \mathbf{C}(\mathscr{C})\} + \varepsilon_L m$ *for some* $\varepsilon_L > 0$*. Set*

$$C_{up} = \sigma \sqrt{\frac{\mathbf{D}(\mathscr{C})}{m - \mathbf{D}(\mathscr{C})}}.$$

Then, for any $\varepsilon > 0$*, there exist* $c_1, c_2 > 0$ *such that for sufficiently large* $\mathbf{D}(\mathscr{C})$*,*

$$\mathbb{P}\left(\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) \leq (1+\varepsilon)\sigma\sqrt{m - \mathbf{D}(\mathscr{C})}\right) \geq 1 - c_1 \exp\left(-c_2 \gamma(m,n)\right).$$

where $\gamma(m,n) = m$ *if* $\mathscr{C}$ *is a cone and* $\gamma(m,n) = \min\left\{m, \frac{m^2}{n}\right\}$ *otherwise.*

3.5.3.3 Deviation Key Optimization

Lemma 3.8 (Properties of $\mathscr{L}_{dev}$) *Let $\mathbf{g} \sim \mathscr{N}(0, \mathbf{I}_m)$ and $\mathbf{h} \sim \mathscr{N}(0, \mathbf{I}_n)$ and*

$$\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\|\mathbf{w}\|_2 \in S_{dev}} \left\{ \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \|\mathbf{g}\|_2 - \mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \mathscr{C}} \mathbf{s}^T \mathbf{w} \right\}, \tag{3.47}$$

where

$$S_{dev} := \left\{ \ell \mid \left| \frac{\ell}{C_{dev}} - 1 \right| \geq \delta_{dev} \right\},$$

$\delta_{dev} > 0$ is any arbitrary small constant and $C_{dev} > 0$. The following are true:

1. Scalarization: $\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\alpha \in S_{dev}} \left\{ \sqrt{\alpha^2 + \sigma^2} \|\mathbf{g}\|_2 - \alpha \cdot dist(\mathbf{h}, \mathscr{C}) \right\}$.

2. Deterministic result: *If*

$$\frac{\sigma \cdot dist(\mathbf{h}, \mathscr{C})}{\sqrt{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h}, \mathscr{C})}} \notin S_{dev}, \tag{3.48}$$

   then,

$$\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) = \sqrt{(1 \pm \delta_{dev})^2 C_{dev}^2 + \sigma^2} \|\mathbf{g}\|_2 - (1 \pm \varepsilon) C_{dev} dist(\mathbf{h}, \mathscr{C}).$$

3. Probabilistic result: *Assume $(1 - \varepsilon_L)m > \mathbf{D}(\mathscr{C}) > \varepsilon_L m$, for some $\varepsilon_0 > 0$ and set*

$$C_{dev} = \sigma \sqrt{\frac{\mathbf{D}(\mathscr{C})}{m - \mathbf{D}(\mathscr{C})}}.$$

   Then, for all $\delta_{dev} > 0$ there exists $t > 0$ and $c_1, c_2 > 0$ such that,

$$\mathbb{P}\left(\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) \geq (1 + t)\sigma \sqrt{m - \mathbf{D}(\mathscr{C})} \right) \geq 1 - c_1 \exp(-c_2 m). \tag{3.49}$$

3.5.4 Going Back: From the Key Optimizations to the Squared Error of the LASSO

Application of Gaussian Min-Max Theorem to $\hat{\mathscr{F}}(\mathbf{A}, \mathbf{v})$ introduced the three key optimizations in Lemma 3.5. Next, in Lemmas 3.6, 3.7 and 3.8 we carried out the analysis of those problems. Here, we combine the results of the four Lemmas mentioned above in order to evaluate $\hat{\mathscr{F}}(\mathbf{A}, \mathbf{v})$ and to compute an exact value for

the norm of its optimizer $\hat{\mathbf{w}}(\mathbf{A},\mathbf{v})$. Lemma 3.9 below formally states the results of the analysis and the proof of it follows.

Lemma 3.9 *Assume $m \geq \max\{\mathbf{D}(\mathscr{C}),\mathbf{D}(\mathscr{C})+\mathbf{C}(\mathscr{C})\}+\varepsilon_L m$ and $\mathbf{D}(\mathscr{C}) \geq \varepsilon_L m$ for some $\varepsilon_L > 0$. Also, assume m is sufficiently large and let $\gamma(m,n) = m$ if $\mathscr{C}$ is a cone and $\min\{m, \frac{m^2}{n}\}$ else. Then, the following statements are true.*

1. *For any $\varepsilon > 0$, there exist constants $c_1, c_2 > 0$ such that*

$$\left|\hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) - \sigma\sqrt{m-\mathbf{D}(\mathscr{C})}\right| \leq \varepsilon\sigma\sqrt{m-\mathbf{D}(\mathscr{C})}, \tag{3.50}$$

   with probability $1-c_1\exp(-c_2\gamma(m,n))$.

2. *For any $\delta_{dev} > 0$ and all $\mathbf{w} \in \mathscr{C}$ satisfying*

$$\left|\|\mathbf{w}\|_2 - \sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}\right| \geq \delta_{dev}\sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}, \tag{3.51}$$

   there exists constant $t(\delta_{dev}) > 0$ and $c_1, c_2 > 0$ such that

$$\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \max_{\mathbf{s}\in\mathscr{C}} \mathbf{s}^T\mathbf{w} \geq \hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) + t\sigma\sqrt{m}, \tag{3.52}$$

   with probability $1-c_1\exp(-c_2\gamma(m,n))$.

3. *For any $\delta > 0$, there exist constants $c_1, c_2 > 0$ such that*

$$\left|\|\hat{\mathbf{w}}(\mathbf{A},\mathbf{v})\|_2 - \sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}\right| \leq \delta\sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}, \tag{3.53}$$

   with probability $1-c_1\exp(-c_2\gamma(m,n))$.

Proof:

We prove each one of the three statements of Theorem 3.9 sequentially. Assume the regime where $m \geq \max\{\mathbf{D}(\mathscr{C}),\mathbf{D}(\mathscr{C})+\mathbf{C}(\mathscr{C})\}+\varepsilon_L m$ and $\mathbf{D}(\mathscr{C}) \geq \varepsilon_L m$ for some $\varepsilon_L > 0$ and also m is sufficiently large.

1. *Proof of* (3.50)*:* Consider any $\varepsilon' > 0$. First, we establish a high probability lower bound for $\hat{\mathscr{F}}(\mathbf{A},\mathbf{v})$. From Lemma 3.6,

$$\mathscr{L}(\mathbf{g},\mathbf{h}) \geq (1-\varepsilon')\sigma\sqrt{m-\mathbf{D}(\mathscr{C})},$$

with probability $1-\exp(-\mathscr{O}(m))$. Combine this with the first statement of Lemma 3.5 to conclude that

$$\hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) \geq (1-\varepsilon')\sigma\sqrt{m-\mathbf{D}(\mathscr{C})} - \varepsilon'\sigma\sqrt{m}, \tag{3.54}$$

with the same probability.

Similarly, for a high probability upper bound for $\hat{\mathscr{F}}(\mathbf{A},\mathbf{v})$ we have from Lemma 3.7, that

$$\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) \leq (1+\varepsilon')\sigma\sqrt{m-\mathbf{D}(\mathscr{C})},$$

with probability $1-\exp(-\mathscr{O}(\gamma(m,n)))$. Combine this with the second statement of Lemma 3.5 to conclude that

$$\hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) \leq (1+\varepsilon')\sigma\sqrt{m-\mathbf{D}(\mathscr{C})} + \varepsilon'\sigma\sqrt{m}, \tag{3.55}$$

with the same probability. To conclude the proof of (3.50) fix any positive constant $\varepsilon > 0$, and observe that by choosing $\varepsilon' = \varepsilon\frac{\sqrt{\varepsilon_L}}{1+\sqrt{\varepsilon_L}}$ in (3.54) and (3.55) we ensure that $\varepsilon'\left(1+\frac{\sqrt{m}}{\sqrt{m-\mathbf{D}(\mathscr{C})}}\right) \leq \varepsilon$. It then follows from (3.54) and (3.55) that there exist $c_1, c_2 > 0$ such that

$$\left|\frac{\hat{\mathscr{F}}(\mathbf{A},\mathbf{v})}{\sigma\sqrt{m-\mathbf{D}(\mathscr{C})}} - 1\right| \leq \varepsilon, \tag{3.56}$$

with probability $1-c_1\exp(-c_2\gamma(m,n))$.

2. *Proof of* (3.52)*:* Fix any $\delta_{dev} > 0$. In accordance to its definition in previous sections define the set

$$S_{dev} = \left\{\ell \mid \left|\ell - \sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}\right| \leq \delta_{dev}\sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}\right\}.$$

Clearly, for all $\mathbf{w}$ such that $\|\mathbf{w}\|_2 \in S_{dev}$ we have,

$$\|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\|_2+\max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w} \geq \min_{\|\mathbf{w}\|_2\in S_{dev}}\left\{\|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\|_2+\max_{s\in\mathscr{C}}s^T\mathbf{w}\right\}.$$

Combining this with the third statement of Lemma 3.5, it suffices for the proof of (3.52) to show that there exists constant $t(\delta_{dev})>0$ such that

$$\mathscr{L}_{dev}(\mathbf{g},\mathbf{h}) \geq \hat{\mathscr{F}}(\mathbf{A},\mathbf{v})+2t\sigma\sqrt{m}, \tag{3.57}$$

with probability $1-\exp(-\mathscr{O}(m))$.

To show (3.57), start from Lemma 3.8 which gives that here exists $t'(\delta_{dev})>0$, such that

$$\mathscr{L}_{dev}(\mathbf{g},\mathbf{h}) \geq (1+t')\sigma\sqrt{m-\mathbf{D}(\mathscr{C})}, \tag{3.58}$$

with probability $1-\exp(-\mathscr{O}(m))$. Furthermore, from the first statement of Lemma 3.9,

$$\hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) \leq (1+\frac{t'}{2})\sigma\sqrt{m-\mathbf{D}(\mathscr{C})}, \tag{3.59}$$

with probability $1-\exp(-\mathscr{O}(\gamma(m,n)))$. Finally, choose $t=\frac{t'}{4}\sqrt{\varepsilon_L}$ to ensure that

$$2t\sigma\sqrt{m} \leq \frac{t'}{2}\sigma\sqrt{m-\mathbf{D}(\mathscr{C})}. \tag{3.60}$$

Combine (3.58), (3.59) and (3.60) to conclude that (3.57) indeed holds with the desired probability.

3. *Proof of* (3.53)*:* The third statement of Lemma 3.9 is a simple consequence of its second statement. Fix any $\varepsilon>0$. The proof is by contradiction. Assume that $\hat{\mathbf{w}}(\mathbf{A},\mathbf{v})$ does not satisfy (3.53). It then satisfies (3.51) for $\delta_{dev}=\varepsilon$. Thus, it follows from the second statement of Lemma 3.9, that there exists $t(\varepsilon)>0$ such that

$$\hat{\mathscr{F}}(\mathbf{A},\mathbf{v}) \geq \hat{\mathscr{F}}(\mathbf{A},\mathbf{v})+t\sigma\sqrt{m}, \tag{3.61}$$

with probability $1-\exp(-\mathscr{O}(\gamma(m,n)))$. This is a contradiction and completes the proof.

■

3.6 The NSE of the C-LASSO

In this section, we prove the second statement of Theorem 3.1, namely (3.18). We restate the theorem here for ease of reference.

Theorem 3.5 *Assume there exists a constant $\varepsilon_L > 0$ such that, $(1-\varepsilon_L)m \geq \mathbf{D}(cone(\partial f(\mathbf{x}_0))) \geq \varepsilon_L m$ and m is sufficiently large. For any $\varepsilon > 0$, there exists a constant $C = C(\varepsilon, \varepsilon_L) > 0$ such that, with probability $1-\exp(-Cm)$,*

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \leq (1+\varepsilon)\frac{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(cone(\partial f(\mathbf{x}_0)))}. \tag{3.62}$$

Furthermore, there exists a deterministic number $\sigma_0 > 0$ (i.e. independent of $\mathbf{A}, \mathbf{v}$) such that, if $\sigma \leq \sigma_0$, with the same probability,

$$\left| \frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \times \frac{m - \mathbf{D}(cone(\partial f(\mathbf{x}_0)))}{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))} - 1 \right| < \varepsilon. \tag{3.63}$$

First, in Section 3.6.1 we focus on the approximated C-LASSO and prove that its NSE concentrates around $\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$ for arbitrary values of σ. Later in Section 3.6.2, we use that result and fundamental properties of the approximated problem to prove (3.18), i.e. that the NSE of the original problem concentrates around the same quantity for small enough σ.

3.6.1 Approximated C-LASSO Problem

Recall the definition of the approximated C-LASSO problem in (3.32). As it has been argued previously, this is equivalent to the generic problem (3.41) with $\mathscr{C} = \text{cone}\{\partial f(\mathbf{x}_0)\}$. Hence, to calculate its NSE we will simply apply the results we obtained throughout Section 3.5. We first start by mapping the generic formulation in Section 3.5 to the C-LASSO.

Lemma 3.10 *Let $\mathscr{C} = cone\{\partial f(\mathbf{x}_0)\}$. Then,*

- *$corr(\mathbf{h}, \mathscr{C}) = 0$, for all $\mathbf{h} \in \mathbb{R}^n$,*
- *$\mathbf{C}(\mathscr{C}) = 0$.*

Proof: The first statement is a direct consequence of Moreau's decomposition theorem (Fact 2.1) applied on the closed and convex cone $\text{cone}\{\partial f(\mathbf{x}_0)\}$. The second statement follows easily after taking the expectation in both sides of the equality in the second statement. ■

With this mapping, we can directly apply Lemma 3.9, where $\mathscr{C}$ is a cone, to conclude with the desired result. The following corollary summarizes the result.

Corollary 3.5 *Assume* $(1-\varepsilon_L)m \geq \mathbf{D}(cone(\partial f(\mathbf{x}_0))) \geq \varepsilon_L m$, *for some* $\varepsilon_L > 0$. *Also, assume* m *is sufficiently large. Then, for any constants* $\varepsilon_1, \varepsilon_2 > 0$, *there exist constants* $c_1, c_2 > 0$ *such that with probability* $1 - c_1 \exp(-c_2 m)$,

$$\left| \frac{\hat{\mathscr{F}}_c(\mathbf{A}, \mathbf{v})}{\sigma\sqrt{m - \mathbf{D}(cone(\partial f(\mathbf{x}_0)))}} - 1 \right| \leq \varepsilon_1,$$

and

$$\left| \frac{\|\hat{\mathbf{w}}_c(\mathbf{A}, \mathbf{v})\|_2^2}{\sigma^2} - \frac{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(cone(\partial f(\mathbf{x}_0)))} \right| \leq \varepsilon_2.$$

3.6.2 Original C-LASSO Problem

In this section we prove (3.18). For the proof we rely on Corollary 3.5. First, we require the introduction of some useful concepts from convex analysis.

3.6.2.1 Tangent Cone and Cone of the Subdifferential

Consider any *convex* set $\mathscr{C} \subset \mathbb{R}^n$ and $\mathbf{x}^* \in \mathscr{C}$. We define the set of feasible directions in $\mathscr{C}$ at $\mathbf{x}^*$ as

$$F_{\mathscr{C}}(\mathbf{x}^*) := \{\mathbf{u} \mid (\mathbf{x}^* + \mathbf{u}) \in \mathscr{C}\}.$$

The tangent cone of $\mathscr{C}$ at $\mathbf{x}^*$ is defined as

$$\mathscr{T}_{\mathscr{C}}(\mathbf{x}^*) := \mathrm{Cl}\left(\mathrm{cone}(F_{\mathscr{C}}(\mathbf{x}^*))\right),$$

where $Cl(\cdot)$ denotes the closure of a set. By definition, tangent cone $\mathscr{T}_{\mathscr{C}}(\mathbf{x}^*)$ and feasible set $F_{\mathscr{C}}(\mathbf{x}^*)$ should be *close* to each other around a small neighborhood of 0. The following proposition is a corollary of Proposition F.1 of [167] and shows that the elements of tangent cone, that are close to the origin, can be *uniformly* approximated by the elements of the feasible set.

Proposition 3.1 (Approximating the tangent cone, [167]) *Let* $\mathscr{C}$ *be a closed convex set and* $\mathbf{x}^* \in \mathscr{C}$. *For*

any $\delta > 0$*, there exists* $\varepsilon > 0$ *such that*

$$dist(\mathbf{u}, F_{\mathscr{C}}(\mathbf{x}^*)) \leq \delta \|\mathbf{u}\|_2,$$

for all $\mathbf{u} \in \mathscr{T}_{\mathscr{C}}(\mathbf{x}^*)$ *with* $\|\mathbf{u}\|_2 \leq \varepsilon$.

Assume $\mathscr{C}$ is the descent set of f at $\mathbf{x}_0$, namely, $\mathscr{C} = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ for some convex function $f(\cdot)$. In this case, we commonly refer to $\mathscr{T}_{\mathscr{C}}(\mathbf{x}_0)$ as the "tangent cone of $f(\cdot)$ at $\mathbf{x}_0$" and denote it by $\mathscr{T}_f(\mathbf{x}_0)$. Under the condition that $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$, the following lemma relates $\mathscr{T}_f(\mathbf{x}_0)$ to the cone of the subdifferential.

Lemma 3.11 ([182]) *Assume* $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ *is convex and* $\mathbf{x}_0 \in \mathbb{R}^n$ *is not a minimizer of it . Then,*

$$(\mathscr{T}_f(\mathbf{x}_0))^\circ = cone(\partial f(\mathbf{x}_0)).$$

3.6.2.2 Proof of Theorem 3.1: Small σ regime

We prove here the second part of Theorem 3.1, namely (3.18). For a proof of (3.17) see Section 3.7. For the purposes of the proof, we will use $\mathscr{C} = \{\mathbf{x} \big| f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$. Recall that we denote the minimizers of the C-LASSO and approximated C-LASSO by $\mathbf{w}_c^*$ and $\hat{\mathbf{w}}_c$, respectively. Also, for convenience denote

$$\eta_c = \frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}.$$

Recalling the definition of the approximated C-LASSO problem in (3.32), we may write

$$\begin{aligned}
\hat{\mathbf{w}}_c &= \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \max_{\lambda \geq 0} \lambda \hat{f}_p(\mathbf{w}) \right\} \\
&= \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \max_{\mathbf{s} \in \text{cone}(\partial f(\mathbf{x}_0))} \mathbf{s}^T \mathbf{w} \right\} \\
&= \arg\min_{\mathbf{w} \in \mathscr{T}_{\mathscr{C}}(\mathbf{x}_0)} \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2,
\end{aligned}$$

where for the last equality we have used Lemma 3.11. Hence,

$$\hat{\mathbf{w}}_c \in \mathscr{T}_{\mathscr{C}}(\mathbf{x}_0). \tag{3.64}$$

At the same time, clearly,

$$\mathbf{w}_c^* \in F_{\mathscr{C}}(\mathbf{x}_0). \tag{3.65}$$

After Corollary 3.5, $\|\hat{\mathbf{w}}_c\|_2^2$ concentrates around $\sigma^2\eta_c$. We will argue that, in the small noise regime, we can translate our results to the original problem in a smooth way. Assume that the statements of Corollary 3.5, hold with high probability for some arbitrary $\varepsilon_1, \varepsilon_2 > 0$. It suffices to prove that for any $\varepsilon_3 > 0$ there exists $\sigma_0 > 0$ such that

$$\left| \frac{\|\mathbf{w}_c^*\|_2^2}{\sigma^2} - \eta_c \right| \leq \varepsilon_3, \tag{3.66}$$

for all $\sigma < \sigma_0$. To begin with, fix a $\delta > 0$, the value of which is to be determined later in the proof. As an immediate implication of Proposition 3.1, there exists σ_0 such that

$$\operatorname{dist}(\mathbf{w}, F_{\mathscr{C}}(\mathbf{x}_0)) \leq \delta \|\mathbf{w}\|_2 \tag{3.67}$$

for all $\mathbf{w} \in \mathscr{T}_{\mathscr{C}}(\mathbf{x}_0)$ satisfying $\|\mathbf{w}\|_2 \leq C = C(\sigma_0, \varepsilon_2) := \sigma_0\sqrt{(1+\varepsilon_2)\eta_c}$.

Now, fix any $\sigma < \sigma_0$. We will make use of the fact that the following three events hold with high probability.

- Using Corollary 3.5, with high probability $\hat{\mathbf{w}}_c$ satisfies,

$$\|\hat{\mathbf{w}}_c\|_2 \leq \sigma\sqrt{(1+\varepsilon_2)\eta_c} \leq C. \tag{3.68}$$

- $\mathbf{A}$ has independent standard normal entries. Hence, its spectral norm satisfies $\|\mathbf{A}\| \leq 2(\sqrt{n}+\sqrt{m})$ with probability $1 - \exp(-\mathscr{O}(\max\{m,n\}))$, [213].
- Using (3.52) of Lemma 3.9 with $\mathscr{C} = \operatorname{cone}(\partial f(\mathbf{x}_0))$, there exists a constant $t = t(\varepsilon_3)$ so that for all $\mathbf{w}$ satisfying $|\frac{\|\mathbf{w}\|_2^2}{\sigma^2} - \eta_c| \geq \varepsilon_3$, we have,

$$\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|_2 + \max_{\mathbf{s}\in\operatorname{cone}(\partial f(\mathbf{x}_0))} \mathbf{s}^T\mathbf{w} \geq \hat{\mathscr{F}}_c(\mathbf{A}, \mathbf{v}) + t(\varepsilon_3)\sigma\sqrt{m}. \tag{3.69}$$

Consider the projection of $\hat{\mathbf{w}}_c$ on the set of feasible directions $F_{\mathscr{C}}(\mathbf{x}_0)$,

$$\mathbf{p}(\hat{\mathbf{w}}_c) := \text{Proj}(\hat{\mathbf{w}}_c, F_{\mathscr{C}}(\mathbf{x}_0)) = \hat{\mathbf{w}}_c - \blacksquare(\hat{\mathbf{w}}_c, F_{\mathscr{C}}(\mathbf{x}_0)). \tag{3.70}$$

First, we show that $\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\|_2$ is not much larger than the objective of the approximated problem, namely $\hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v})$. Indeed,

$$\begin{aligned}
\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\|_2 &\leq \|\mathbf{A}\hat{\mathbf{w}}_c - \sigma\mathbf{v}\|_2 + \|\mathbf{A}\hat{\mathbf{w}}_c - \mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c)\|_2 \\
&\leq \hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v}) + \|\mathbf{A}\|\text{dist}(\hat{\mathbf{w}}_c, F_{\mathscr{C}}(\mathbf{x}_0)) \\
&\leq \hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v}) + \|\mathbf{A}\|\sigma\delta\sqrt{(1+\varepsilon_2)\eta_c} \\
&\leq \hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v}) + 2(\sqrt{m}+\sqrt{n})\sigma\delta\sqrt{(1+\varepsilon_2)\eta_c}.
\end{aligned} \tag{3.71}$$

The first inequality is an application of the triangle inequality and the second one follows from (3.70). For the third inequality, we have used (3.64) and combined (3.67) with (3.68).

Next, we show that if (3.66) was not true then a suitable choice of δ would make $\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\|_2$ much larger than the optimal $\hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v})$ than (3.71) allows. Therefore, concluding a desired contradiction. More precisely, assuming (3.66) does not hold, we have

$$\begin{aligned}
\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\|_2 &\geq \|\mathbf{A}\mathbf{w}_c^* - \sigma\mathbf{v}\|_2 \\
&\geq \hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v}) + t(\varepsilon_3)\sigma\sqrt{m}.
\end{aligned} \tag{3.72}$$

The first inequality above follows since $\mathbf{p}(\hat{\mathbf{w}}_c) \in F_{\mathscr{C}}(\mathbf{x}_0)$ and from the optimality of $\mathbf{w}_c^* \in F_{\mathscr{C}}(\mathbf{x}_0)$. To get the second inequality, recall that (3.66) is not true. Also, from (3.65), $\max_{\mathbf{s}\in\text{cone}(\partial f(\mathbf{x}_0))} \mathbf{s}^T\mathbf{w}_c^* = \max_{\mathbf{s}\in(\mathscr{T}(\mathbf{x}_0))^\circ} \mathbf{s}^T\mathbf{w}_c^* = 0$. Combine these and invoke (3.69).

To conclude, choose σ_0 sufficiently small to ensure $\delta < \frac{t(\varepsilon_3)\sqrt{m}}{2(\sqrt{m}+\sqrt{n})\sqrt{(1+\varepsilon_2)\eta_c}}$ and combine (3.71) and (3.72) to obtain the following contradiction.

$$\begin{aligned}
\hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v}) + 2(\sqrt{m}+\sqrt{n})\delta\sigma\sqrt{(1+\varepsilon_2)\eta_c} &\geq \|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\|_2 \\
&\geq \hat{\mathscr{F}}_c(\mathbf{A},\mathbf{v}) + t(\varepsilon_3)\sigma\sqrt{m}.
\end{aligned}$$

σ_0 is a deterministic number that is a function of $m, n, f, \mathbf{x}_0, \varepsilon_3$.

3.7 Constrained-LASSO Analysis for Arbitrary σ

In Section 3.6 we proved the first part of Theorem 3.1, which refers to the case where $\sigma \to 0$. Here, we complete the proof of the Theorem by showing (3.17), which is to say that the worst case NSE of the C-LASSO problem is achieved as $\sigma \to 0$. In other words, we prove that our exact bounds for the small σ regime upper bound the squared error, for arbitrary values of the noise variance. The analysis relies, again, on the proper application of the "modified Gordon's Lemma" Proposition 2.7.

3.7.1 Notation

We begin with describing some notation used throughout this section. First, we denote

$$\text{dist}_{\mathbb{R}^+}(\mathbf{h}) := \text{dist}(\mathbf{h}, \text{cone}(\partial f(\mathbf{x}_0))).$$

Also, recall the definitions of the "perturbation" functions $f_p(\cdot)$ and $\hat{f}_p(\cdot)$ in (3.27) and (3.28). Finally, we will be making use of the following functions:

$$\begin{aligned}
&\mathscr{F}(\mathbf{w};\mathbf{A},\mathbf{v}) := \|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\|_2, \\
&\mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}) := \sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w}, \qquad (3.73)\\
&L(\alpha;a,b) := \sqrt{\alpha^2+\sigma^2}a - \alpha b. \qquad (3.74)
\end{aligned}$$

Using this notation, and denoting the optimal cost of the (original) C-LASSO (see (3.3)) as $\mathscr{F}_c^*(\mathbf{A},\mathbf{v})$, we write

$$\mathscr{F}_c^*(\mathbf{A},\mathbf{v}) = \min_{f_p(\mathbf{w})\le 0} \mathscr{F}(\mathbf{w};\mathbf{A},\mathbf{v}) = \mathscr{F}(\mathbf{w}_c^*;\mathbf{A},\mathbf{v}). \qquad (3.75)$$

3.7.2 Lower Key Optimization

As a first step in our proof, we apply Proposition 2.7 to the original C-LASSO problem in (3.75). Recall, that application of Corollary 3.2 to the approximated problem resulted in the following key optimization:

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \min_{\hat{f}_p(\mathbf{w})\le 0} \left\{ \sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} \right\} = \min_{\hat{f}_p(\mathbf{w})\le 0} \mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}). \qquad (3.76)$$

Denote the minimizer of (3.76), as $\hat{\mathbf{w}}_{low}$. Using Corollary 3.2, the lower key optimization corresponding to the original C-LASSO has the following form:

$$\mathscr{L}^*(\mathbf{g},\mathbf{h}) = \min_{f_p(\mathbf{w})\leq 0} \left\{ \sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} \right\} = \min_{f_p(\mathbf{w})\leq 0} \mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}). \tag{3.77}$$

Recall that in both (3.76) and (3.77), $\mathbf{g}\in\mathbb{R}^m$ and $\mathbf{h}\in\mathbb{R}^n$. In Lemma 3.6 in Section 3.5 we solved explicitly for the optimizer $\hat{\mathbf{w}}_{low}$ of problem (3.76). In a similar nature, Lemma 3.12 below identifies a critical property of the optimizer $\mathbf{w}^*_{low}$ of the key optimization (3.77): $\|\mathbf{w}^*\|_2$ is no larger than $\|\hat{\mathbf{w}}_{low}\|_2$.

Lemma 3.12 *Let $\mathbf{g}\in\mathbb{R}^m,\mathbf{h}\in\mathbb{R}^n$ be given and $\|\mathbf{g}\|_2 > dist_{\mathbb{R}^+}(\mathbf{h})$. Denote the minimizer of the problem (3.77) as $\mathbf{w}^*_{low} = \mathbf{w}^*_{low}(\mathbf{g},\mathbf{h})$. Then,*

$$\frac{\|\mathbf{w}^*_{low}\|_2^2}{\sigma^2} \leq \frac{dist_{\mathbb{R}^+}(\mathbf{h})^2}{\|\mathbf{g}\|_2^2 - dist_{\mathbb{R}^+}(\mathbf{h})^2} = \frac{\|\hat{\mathbf{w}}_{low}\|_2^2}{\sigma^2}. \tag{3.78}$$

For the proof of Lemma 3.12, we require the following result on the tangent cone of the feasible set of (3.77).

Lemma 3.13 *Let $f(\cdot):\mathbb{R}^n\to\mathbb{R}$ be a convex function and $\mathbf{x}_0\in\mathbb{R}^n$ that is not a minimizer of $f(\cdot)$. Consider the set $\mathscr{C} = \{\mathbf{w}\big| f(\mathbf{x}_0+\mathbf{w})\leq f(\mathbf{x}_0)\}$. Then, for all $\mathbf{w}^*\in\mathscr{C}$,*

$$\mathscr{T}_{\mathscr{C}}(\mathbf{w}^*)^\circ = \begin{cases} cone(\partial f(\mathbf{x}_0+\mathbf{w}^*)) & \text{if } f(\mathbf{x}_0+\mathbf{w}^*) = f(\mathbf{x}_0), \\ \{0\} & \text{if } f(\mathbf{x}_0+\mathbf{w}^*) < f(\mathbf{x}_0). \end{cases} \tag{3.79}$$

Proof: We need to characterize the feasible set $F_{\mathscr{C}}(\mathbf{w}^*)$.

Suppose $f(\mathbf{x}_0+\mathbf{w}^*) < f(\mathbf{x}_0)$. Since $f(\cdot)$ is continuous, for all directions $\mathbf{u}\in\mathbb{R}^n$, there exists sufficiently small $\varepsilon>0$ such that $f(\mathbf{x}_0+\mathbf{w}^*+\varepsilon\mathbf{u})\in\mathscr{C}$. Hence, $\mathscr{T}_{\mathscr{C}}(\mathbf{w}^*) = \text{cone}(\text{Cl}(F_{\mathscr{C}}(\mathbf{w}^*))) = \mathbb{R}^n \implies (\mathscr{T}_{\mathscr{C}}(\mathbf{w}^*))^\circ = \{0\}$ in this case.

Now, assume $f(\mathbf{x}_0+\mathbf{w}^*) = f(\mathbf{x}_0)$. Then, $F_{\mathscr{C}}(\mathbf{w}^*) = \{\mathbf{u}\big| f(\mathbf{x}_0+\mathbf{w}^*+\mathbf{u})\leq f(\mathbf{x}_0) = f(\mathbf{x}_0+\mathbf{w}^*)\} = F_{\mathscr{C}'}(\mathbf{x}_0+\mathbf{w}^*)$, where $F_{\mathscr{C}'}(\mathbf{x}_0+\mathbf{w}^*)$ denotes the set of feasible directions in $\mathscr{C}' := \{\mathbf{x}|f(\mathbf{x})\leq f(\mathbf{x}_0+\mathbf{w}^*)\}$ at $\mathbf{x}_0+\mathbf{w}^*$. Thus, $\mathscr{T}_{\mathscr{C}}(\mathbf{w}^*) = \mathscr{T}_{\mathscr{C}'}(\mathbf{x}_0+\mathbf{w}^*) = \text{cone}(\partial f(\mathbf{x}_0+\mathbf{w}^*))^\circ$, where the last equality follows from Lemma 3.11, and the fact that $\mathbf{x}_0+\mathbf{w}^*$ is not a minimizer of $f(\cdot)$ as $f(\mathbf{x}_0) = f(\mathbf{x}_0+\mathbf{w}^*)$. ■

Proof: [Proof of Lemma 3.12]

We first show that, $\mathbf{w}^*_{low}$ exists and is finite. From the convexity of $f(\cdot)$, $\hat{f}_p(\mathbf{w}) \leq f_p(\mathbf{w})$, thus, every feasible solution of (3.77) is also feasible for (3.76). This implies that $\mathscr{L}^*(\mathbf{g},\mathbf{h}) \geq \mathscr{L}(\mathbf{g},\mathbf{h})$. Also, from Lemma 3.6, $\mathscr{L}(\mathbf{g},\mathbf{h}) = \sigma\sqrt{\|\mathbf{g}\|_2^2 - \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})^2}$. Combining,

$$\mathscr{L}^*(\mathbf{g},\mathbf{h}) \geq \sigma\sqrt{\|\mathbf{g}\|_2^2 - \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})^2} > 0. \tag{3.80}$$

Using the scalarization result of Lemma 3.6 with $\mathscr{C} = \mathrm{cone}(\partial f(\mathbf{x}_0))$, for any $\alpha \geq 0$,

$$\min_{\substack{\hat{f}_p(\mathbf{w})\leq 0 \\ \|\mathbf{w}\|_2=\alpha}} \mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}) = L(\alpha, \|\mathbf{g}\|_2, \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})).$$

Hence, using Lemma A.10 in the appendix shows that, when $\|\mathbf{g}\|_2 > \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})$,

$$\lim_{C\to\infty} \min_{\substack{\|\mathbf{w}\|_2\geq C \\ f_p(\mathbf{w})\leq 0}} \mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}) = \lim_{C\to\infty} \min_{\alpha\geq C} L(\alpha, \|\mathbf{g}\|_2, \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})) = \infty.$$

Combining this with (3.80) shows that $\mathscr{L}^*(\mathbf{g},\mathbf{h})$ is strictly positive, and that $\|\mathbf{w}^*_{low}\|_2$ and $\mathbf{w}^*_{low}$ is finite.

The minimizer $\mathbf{w}^*_{low}$ satisfies the KKT optimality conditions of (3.77) [17]:

$$\frac{\mathbf{w}^*_{low}}{\sqrt{\|\mathbf{w}^*_{low}\|_2^2 + \sigma^2}}\|\mathbf{g}\|_2 = \mathbf{h} - \mathbf{s}^*,$$

or, equivalently,

$$\mathbf{w}^*_{low} = \sigma\frac{\mathbf{h} - \mathbf{s}^*}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} - \mathbf{s}^*\|_2^2}}, \tag{3.81}$$

where, from Lemma 3.13,

$$\mathbf{s}^* \in \begin{cases} \mathrm{cone}\left(\partial f(\mathbf{x}_0 + \mathbf{w}^*_{low})\right) & \text{if } f_p(\mathbf{w}_{low}) = 0, \\ \{0\} & \text{if } f_p(\mathbf{w}_{low}) < 0. \end{cases} \tag{3.82}$$

First, consider the scenario in (3.82) where $f_p(\mathbf{w}^*_{low}) < 0$ and $\mathbf{s}^* = 0$. Then, from (3.81) $\mathbf{h} = c_h\mathbf{w}^*_{low}$ for some constant $c_h > 0$. But, from feasibility constraints, $\mathbf{w}^*_{low} \in \mathscr{T}_f(\mathbf{x}_0)$, hence, $\mathbf{h} \in \mathscr{T}_f(\mathbf{x}_0) \implies \|\mathbf{h}\|_2 = \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})$ which implies equality in (3.78).

Otherwise, $f(\mathbf{x}_0+\mathbf{w}^*_{low}) = f(\mathbf{x}_0)$ and $\mathbf{s}^* \in \text{cone}\left(\partial f(\mathbf{x}_0+\mathbf{w}^*_{low})\right)$. For this case, we argue that $\|\mathbf{h}-\mathbf{s}^*\|_2 \leq \text{dist}_{\mathbb{R}^+}(\mathbf{h})$. To begin with, there exists scalar $\theta > 0$ such that $\theta\mathbf{s}^* \in \partial f(\mathbf{x}_0+\mathbf{w}^*_{low})$. Convexity of $f(\cdot)$, then, implies that,

$$f(\mathbf{x}_0+\mathbf{w}^*_{low}) = f(\mathbf{x}_0) \geq f(\mathbf{x}_0+\mathbf{w}^*_{low}) - \langle \theta\mathbf{s}^*, \mathbf{w}^*_{low}\rangle \implies \langle \mathbf{s}^*, \mathbf{w}^*_{low}\rangle \geq 0. \tag{3.83}$$

Furthermore, $\mathbf{w}^*_{low} \in \mathscr{T}_f(\mathbf{x}_0)$ and $\mathbf{s}_0 := \text{Proj}(\mathbf{h}, \text{cone}(\partial f(\mathbf{x}_0)))$, thus

$$\langle \mathbf{w}^*_{low}, \mathbf{s}_0\rangle \leq 0. \tag{3.84}$$

Combine (3.83) and (3.84), and further use (3.81) to conclude that

$$\langle \mathbf{w}^*_{low}, \mathbf{s}^*-\mathbf{s}_0\rangle \geq 0 \implies \langle \mathbf{h}-\mathbf{s}^*, \mathbf{s}^*-\mathbf{s}_0\rangle \geq 0.$$

We may then write,

$$(\text{dist}_{\mathbb{R}^+}(\mathbf{h}))^2 = \|(\mathbf{h}-\mathbf{s}^*)+(\mathbf{s}^*-\mathbf{s}_0)\|_2^2 \geq \|\mathbf{h}-\mathbf{s}^*\|_2^2, \tag{3.85}$$

and combine with the fact that the function $f(x,y) = \frac{x}{\sqrt{y^2-x^2}}, x \geq 0, y > 0$ is nondecreasing in the regime $x < y$, to complete the proof. ■

3.7.3 Upper Key Optimization

In this section we find a high probability upper bound for $\mathscr{F}^*_c(\mathbf{A}, \mathbf{v})$. Using Corollary 3.3 of Section 3.4.3.2, application of Proposition 2.7 to the dual of the C-LASSO results in the following key optimization:

$$\mathscr{U}^*(\mathbf{g}, \mathbf{h}) = \max_{\|\mu\|_2 \leq 1} \left\{ \min_{\substack{f_p(\mathbf{w}) \leq 0 \\ \|\mathbf{w}\|_2 \leq \sigma C_{up}}} \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\mu^T\mathbf{g} - \|\mu\|_2\mathbf{h}^T\mathbf{w} \right\}, \tag{3.86}$$

where

$$C_{up} = 2\sqrt{\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}}.$$

Normalizing the inner terms in (3.86) by $\|\mu\|_2$ for $\mu \neq 0$, this can be equivalently be written as,

$$\begin{aligned}\mathscr{U}^*(\mathbf{g},\mathbf{h}) &= \max_{\|\mu\|_2\leq 1}\left\{\|\mu\|_2 \min_{\substack{f_p(\mathbf{w})\leq 0\\ \|\mathbf{w}\|_2\leq\sigma C_{up}}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2-\mathbf{h}^T\mathbf{w}\right\}\right\}\\ &= \max\left\{0,\ \min_{\substack{f_p(\mathbf{w})\leq 0\\ \|\mathbf{w}\|_2\leq\sigma C_{up}}}\mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h})\right\}\\ &= \max\left\{0,\ \mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})\right\},\end{aligned} \tag{3.87}$$

where we additionally defined

$$\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h}) := \min_{\substack{f_p(\mathbf{w})\leq 0\\ \|\mathbf{w}\|_2\leq\sigma C_{up}}}\mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}). \tag{3.88}$$

Observe the similarity of the upper key optimization (3.87) to the lower key optimization (3.77). The next lemma proves that $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ and $\mathscr{U}^*(\mathbf{g},\mathbf{h})$ are Lipschitz functions.

Lemma 3.14 (Lipschitzness of $\mathscr{U}^*(\mathbf{g},\mathbf{h})$) *$\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ and, consequently, $\mathscr{U}^*(\mathbf{g},\mathbf{h})$ are Lipschitz with Lipschitz constants at most $2\sigma\sqrt{C_{up}^2+1}$.*

Proof: First, we prove that $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ is Lipschitz. Given pairs $(\mathbf{g}_1,\mathbf{h}_1),(\mathbf{g}_2,\mathbf{h}_2)$, denote $\mathbf{w}_1$ and $\mathbf{w}_2$ the corresponding optimizers in problem (3.88). W.l.o.g., assume that $\mathscr{L}^*_{up}(\mathbf{g}_1,\mathbf{h}_1)\geq\mathscr{L}^*_{up}(\mathbf{g}_2,\mathbf{h}_2)$. Then,

$$\begin{aligned}\mathscr{L}^*_{up}(\mathbf{g}_1,\mathbf{h}_1)-\mathscr{L}^*_{up}(\mathbf{g}_2,\mathbf{h}_2) &= \mathscr{L}(\mathbf{w}_1;\mathbf{g}_1,\mathbf{h}_2)-\mathscr{L}(\mathbf{w}_2;\mathbf{g}_2,\mathbf{h}_2)\\ &\leq \mathscr{L}(\mathbf{w}_2;\mathbf{g}_1,\mathbf{h}_1)-\mathscr{L}(\mathbf{w}_2;\mathbf{g}_2,\mathbf{h}_2)\\ &= \sqrt{\|\mathbf{w}_2\|_2^2+\sigma^2}(\|\mathbf{g}_1\|_2-\|\mathbf{g}_2\|_2)-(\mathbf{h}_1-\mathbf{h}_2)^T\mathbf{w}_2\\ &\leq \sqrt{\sigma^2C_{up}^2+\sigma^2}\|\mathbf{g}_1-\mathbf{g}_2\|_2+\|\mathbf{h}_1-\mathbf{h}_2\|_2\sigma C_{up},\end{aligned} \tag{3.89}$$

where, we have used the fact that $\|\mathbf{w}_2\|_2\leq\sigma C_{up}$. From (3.89), it follows that $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ is indeed Lipschitz and

$$|\mathscr{L}^*_{up}(\mathbf{g}_1,\mathbf{h}_1)-\mathscr{L}^*_{up}(\mathbf{g}_2,\mathbf{h}_2)|\leq 2\sigma\sqrt{C_{up}^2+1}\sqrt{\|\mathbf{g}_1-\mathbf{g}_2\|_2^2+\|\mathbf{h}_1-\mathbf{h}_2\|_2^2}.$$

To prove that $\mathscr{U}^*(\mathbf{g},\mathbf{h})$ is Lipschitz with the same constant, assume w.l.o.g that $\mathscr{U}^*(\mathbf{g}_1,\mathbf{h}_1) \geq \mathscr{U}^*(\mathbf{g}_2,\mathbf{h}_2)$. Then, from (3.87),

$$|\mathscr{U}^*(\mathbf{g}_1,\mathbf{h}_1) - \mathscr{U}^*(\mathbf{g}_2,\mathbf{h}_2)| \leq |\mathscr{L}^*_{up}(\mathbf{g}_1,\mathbf{h}_1) - \mathscr{L}^*_{up}(\mathbf{g}_2,\mathbf{h}_2)|.$$

■

3.7.4 Matching Lower and Upper key Optimizations

Comparing (3.77) to (3.87), we have already noted that the lower and upper key optimizations have similar forms. The next lemma proves that their optimal costs match, in the sense that they concentrate with high probability over the same quantity, namely $\mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]$.

Lemma 3.15 *Let $\mathbf{g} \sim \mathscr{N}(0,\mathbf{I}_m), \mathbf{h} \sim \mathscr{N}(0,\mathbf{I}_n)$ be independent vectors. Assume $(1-\varepsilon_0)m \geq \mathbf{D}(cone(\partial f(\mathbf{x}_0)))$ $\geq \varepsilon_0 m$ for some constant $\varepsilon_0 > 0$ and m sufficiently large. For any $\varepsilon > 0$, there exists $c > 0$ such that, with probability $1-\exp(-cm)$, we have,*

1. $|\mathscr{U}^*(\mathbf{g},\mathbf{h}) - \mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]| \leq \varepsilon\sigma\sqrt{m}$.
2. $|\mathscr{L}^*(\mathbf{g},\mathbf{h}) - \mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]| \leq \varepsilon\sigma\sqrt{m}$.

In Lemma 3.14 we proved that $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ is Lipschitz. Gaussian concentration of Lipschitz functions (see Lemma 2.4) implies, then, that $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ concentrates with high probability around its mean $\mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]$. According to Lemma 3.15, under certain conditions implied by its assumptions, $\mathscr{U}^*(\mathbf{g},\mathbf{h})$ and $\mathscr{L}^*(\mathbf{g},\mathbf{h})$ also concentrate around the same quantity $\mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]$. The way to prove this fact is by showing that when these conditions hold, $\mathscr{U}^*(\mathbf{g},\mathbf{h})$ and $\mathscr{L}^*(\mathbf{g},\mathbf{h})$ are equal to $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ with high probability. Once we have shown that, we require the following result to complete the proof.

Lemma 3.16 *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{h} \sim \mathscr{N}(0,\mathbf{I}_n)$. Assume f_1 is L-Lipschitz and, $\mathbb{P}(f_1(\mathbf{g}) = f_2(\mathbf{g})) > 1-\varepsilon$. Then, for all $t > 0$,*

$$\mathbb{P}\left(|f_2(\mathbf{g}) - \mathbb{E}[f_1(\mathbf{g})]| \leq t\right) > 1 - \varepsilon - 2\exp\left(-\frac{t^2}{2L^2}\right).$$

Proof: From standard concentration result on Lipschitz functions (see Lemma 2.4), for all $t > 0$, $|f_1(\mathbf{g}) - \mathbb{E}[f_1(\mathbf{g})]| < t$ with probability $1-2\exp(-\frac{t^2}{2L^2})$. Also, by assumption $f_2(\mathbf{g}) = f_1(\mathbf{g})$ with probability

$1-\varepsilon$. Combine those facts to complete the proof as follows,

$$\begin{aligned}\mathbb{P}(|f_2(\mathbf{g})-\mathbb{E}[f_1(\mathbf{g})]|\leq t) &\geq \mathbb{P}(|f_2(\mathbf{g})-\mathbb{E}[f_1(\mathbf{g})]|\leq t \mid f_1(\mathbf{g})=f_2(\mathbf{g}))\,\mathbb{P}(f_1(\mathbf{g})=f_2(\mathbf{g}))\\ &= \mathbb{P}(|f_1(\mathbf{g})-\mathbb{E}[f_1(\mathbf{g})]|\leq t)\,\mathbb{P}(f_1(\mathbf{g})=f_2(\mathbf{g}))\\ &\geq \left(1-2\exp(-\frac{t^2}{2L^2})\right)(1-\varepsilon).\end{aligned}$$

■

Now, we complete the proof of Lemma 3.15 using the result of Lemma 3.16. *Proof:* [Proof of Lemma 3.15] We prove the two statements of the lemma in the order they appear.

1. First, we prove that under the assumptions of the lemma, $\mathscr{U}^*=\mathscr{L}^*_{up}$ w.h.p.. By (3.87), it suffices to show that $\mathscr{L}^*_{up}\geq 0$ w.h.p.. Constraining the feasible set of a minimization problem cannot result in a decrease in its optimal cost, hence,

$$\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})\geq \mathscr{L}^*(\mathbf{g},\mathbf{h})\geq \mathscr{L}(\mathbf{g},\mathbf{h}). \tag{3.90}$$

where recall $\mathscr{L}(\mathbf{g},\mathbf{h})$ is the lower key optimization of the approximated C-LASSO (see (3.76)). From Lemma 3.6, since $m\geq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))+\varepsilon_0 m$, we have that

$$\mathscr{L}(\mathbf{g},\mathbf{h})\geq (1-\varepsilon)\sigma\sqrt{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}\geq 0,$$

with $1-\exp(-\mathscr{O}(m))$. Combine this with (3.90) to find that $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})\geq 0$ or $\mathscr{U}^*=\mathscr{L}^*_{up}$ with probability $1-\exp(-\mathscr{O}(m))$. Furthermore, from Lemma 3.14, $\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ is Lipschitz with constant $L=2\sigma\sqrt{C_{up}^2+1}$. We now apply Lemma 3.16 setting $f_1=\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$, $f_2=\mathscr{U}^*$ and $t=\varepsilon\sigma\sqrt{m}$, to find that

$$|\mathscr{U}^*(\mathbf{g},\mathbf{h})-\mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]|\leq \varepsilon\sqrt{m},$$

with probability $1-\exp(-\mathscr{O}(m))$. In writing the exponent in the probability as $\mathscr{O}(m)$, we made use of the fact that $C_{up}=2\sqrt{\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}}$ is bounded below by a constant, since $(1-\varepsilon_0)m\geq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))\geq \varepsilon_0 m$.

2. As in the first statement, we apply Lemma 3.16, this time setting $f_1=\mathscr{L}^*_{up}$, $f_2=\mathscr{L}^*$ and $t=\varepsilon\sigma\sqrt{m}$. The result is immediate after application of the lemma, but first we need to show that $\mathscr{L}^*(\mathbf{g},\mathbf{h})=\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})$ w.h.p.. We will show equivalently that the minimizer $\mathbf{w}^*_{low}$ of (3.77) satisfies $\mathbf{w}^*_{low}\in S_{up}$. From Lemma 3.12, $\|\mathbf{w}^*_{low}\|_2\leq \frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\|\mathbf{g}\|_2-\text{dist}_{\mathbb{R}^+}(\mathbf{h})}$. On the other hand, using standard concentration arguments (Lemma A.2),

with probability $1-\exp(-\mathscr{O}(m))$, $\frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\|\mathbf{g}\|_2-\text{dist}_{\mathbb{R}^+}(\mathbf{h})} \leq \frac{2\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} = C_{up}$. Combining these completes the proof. ■

3.7.5 Deviation Bound

Resembling the approach developed in Section 3.5, we show that if we restrict the norm of the error vector $\|\mathbf{w}\|_2$ in (3.75) as follows

$$\|\mathbf{w}\|_2 \in S_{dev} := \left\{ \ell \middle| \ell \geq (1+\varepsilon_{dev})\sigma \sqrt{\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}} \right\}, \tag{3.91}$$

then, this results in a significant increase in the cost of C-LASSO. To lower bound the deviated cost, we apply Corollary 3.4 of Section 3.4.3.3 to the restricted original C-LASSO, which yields the following key optimization

$$\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) = \min_{\substack{f_p(\mathbf{w})\leq 0 \\ \|\mathbf{w}\|_2 \in S_{dev}}} \mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h}). \tag{3.92}$$

Lemma 3.17 *Let $\mathbf{g} \sim \mathscr{N}(0,\mathbf{I}_m), \mathbf{h} \sim \mathscr{N}(0,\mathbf{I}_n)$. Assume $(1-\varepsilon_L)m > \mathbf{D}(cone(\partial f(\mathbf{x}_0))) > \varepsilon_L m$ and m is sufficiently large. Then, there exists a constant $\delta_{dev} = \delta_{dev}(\varepsilon_{dev}) > 0$ such that, with probability $1-\exp(-\mathscr{O}(m))$, we have,*

$$\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})] \geq \sigma\delta_{dev}\sqrt{m}. \tag{3.93}$$

As common, our analysis begins with a deterministic result, which builds towards the proof of the probabilistic statement in Lemma 3.17.

3.7.5.1 Deterministic Result

For the statement of the deterministic result, we introduce first some notation. In particular, denote

$$\eta_d := \sigma \sqrt{\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}},$$

and, for fixed $\mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$,

$$\eta_s = \eta_s(\mathbf{g},\mathbf{h}) := \sigma \frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\sqrt{\|\mathbf{g}\|_2^2 - \text{dist}_{\mathbb{R}^+}(\mathbf{h})^2}}.$$

Also, recall the definition of the scalar function $L(\alpha;a,b)$ in (3.74).

Lemma 3.18 *Let $\mathbf{g}\in\mathbb{R}^m$ and $\mathbf{h}\in\mathbb{R}^n$ be such that $\|\mathbf{g}\|_2 > dist_{\mathbb{R}^+}(\mathbf{h})$ and $\eta_s(\mathbf{g},\mathbf{h}) \le (1+\varepsilon_{dev})\eta_d$. Then,*

$$\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathscr{L}^*(\mathbf{g},\mathbf{h}) \ge L((1+\varepsilon_{dev})\eta_d;\|\mathbf{g}\|_2,dist_{\mathbb{R}^+}(\mathbf{h})) - L(\eta_s(\mathbf{g},\mathbf{h});\|\mathbf{g}\|_2,dist_{\mathbb{R}^+}(\mathbf{h})) \tag{3.94}$$

Proof: First assume that $\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) = \infty$. Since $\mathscr{L}^*(\mathbf{g},\mathbf{h}) \le \mathscr{L}(\mathbf{0};\mathbf{g},\mathbf{h}) = \sigma\|\mathbf{g}\|_2$ and the right hand side of (3.94) is finite, we can easily conclude with the desired result.

Hence, in the following assume that $\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) < \infty$ and denote $\mathbf{w}^*_{dev}$ the minimizer of the restricted problem (3.92). From feasibility constraints, we have $f_p(\mathbf{w}_{dev}) \le 0$ and $\|\mathbf{w}^*_{dev}\|_2 \in S_{dev}$. Define $\bar{\mathbf{w}}_{dev} = c\mathbf{w}^*_{dev}$ where $c := \frac{\eta_s}{\|\mathbf{w}^*_{dev}\|_2}$. Notice, $\|\mathbf{w}^*_{dev}\|_2 \ge (1+\varepsilon_{dev})\eta_d \ge \eta_s(\mathbf{g},\mathbf{h})$, thus, $c \le 1$. Then, from convexity of $f(\cdot)$,

$$f_p(\bar{\mathbf{w}}_{dev}) = f_p(c\mathbf{w}^*_{dev}) \le cf_p(\mathbf{w}^*_{dev}) + (1-c)\underbrace{f_p(\mathbf{0})}_{=0} \le 0.$$

This shows that $\bar{\mathbf{w}}_{dev}$ is feasible for the minimization (3.77). Hence,

$$\mathscr{L}(\bar{\mathbf{w}}_{dev},\mathbf{g},\mathbf{h}) \ge \mathscr{L}^*(\mathbf{g},\mathbf{h}).$$

Starting with this, we write,

$$\begin{aligned}
\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathscr{L}^*(\mathbf{g},\mathbf{h}) &\ge \mathscr{L}(\mathbf{w}^*_{dev};\mathbf{g},\mathbf{h}) - \mathscr{L}(\bar{\mathbf{w}}_{dev};\mathbf{g},\mathbf{h}) \\
&= (\sqrt{\|\mathbf{w}^*_{dev}\|_2^2+\sigma^2} - \sqrt{\|\bar{\mathbf{w}}_{dev}\|_2^2+\sigma^2})\|\mathbf{g}\|_2 - \mathbf{h}^T(\mathbf{w}^*_{dev} - \bar{\mathbf{w}}_{dev}) \\
&= (\sqrt{\|\mathbf{w}^*_{dev}\|_2^2+\sigma^2} - \sqrt{\|\bar{\mathbf{w}}_{dev}\|_2^2+\sigma^2})\|\mathbf{g}\|_2 - (1-c)\mathbf{h}^T\mathbf{w}^*_{dev}.
\end{aligned} \tag{3.95}$$

Since, $f_p(\mathbf{w}^*_{dev}) \le 0$, $\mathbf{w}^*_{dev} \in \mathscr{T}_f(\mathbf{x}_0)$. Hence, and using Moreau's decomposition Theorem (see Fact 2.1), we have

$$\begin{aligned}
\mathbf{h}^T\mathbf{w}^*_{dev} &= \langle \mathrm{Proj}(\mathbf{h},\mathscr{T}_f(\mathbf{x}_0)),\mathbf{w}^*_{dev}\rangle + \underbrace{\langle \mathrm{Proj}(\mathbf{h},(\mathscr{T}_f(\mathbf{x}_0))^\circ),\mathbf{w}^*_{dev}\rangle}_{\le 0} \\
&\le \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})\|\mathbf{w}^*_{dev}\|_2.
\end{aligned} \tag{3.96}$$

Use (3.96) in (3.95), to write

$$\begin{aligned}\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h})-\mathscr{L}^*(\mathbf{g},\mathbf{h}) &\geq (\sqrt{\|\mathbf{w}^*_{dev}\|_2^2+\sigma^2}-\sqrt{\|\bar{\mathbf{w}}_{dev}\|_2^2+\sigma^2})-\frac{\|\mathbf{w}^*_{dev}\|_2-\eta_s}{\|\mathbf{w}^*_{dev}\|_2}\text{dist}_{\mathbb{R}^+}(\mathbf{h})\|\mathbf{w}^*_{dev}\|_2\\ &= (\sqrt{\|\mathbf{w}^*_{dev}\|_2^2+\sigma^2}-\sqrt{\eta_s^2+\sigma^2})\|\mathbf{g}\|_2-(\|\mathbf{w}^*_{dev}\|_2-\eta_s)\text{dist}_{\mathbb{R}^+}(\mathbf{h})\\ &= L(\|\mathbf{w}^*_{dev}\|_2,\|\mathbf{g}\|_2,\text{dist}_{\mathbb{R}^+}(\mathbf{h}))-L(\eta_s,\|\mathbf{g}\|_2,\text{dist}_{\mathbb{R}^+}(\mathbf{h}))\\ &\geq L((1+\varepsilon)\eta_d,\|\mathbf{g}\|_2,\text{dist}_{\mathbb{R}^+}(\mathbf{h}))-L(\eta_s,\|\mathbf{g}\|_2,\text{dist}_{\mathbb{R}^+}(\mathbf{h})).\end{aligned}$$

The last inequality above follows from the that $L(\alpha;\|\mathbf{g}\|_2,\text{dist}_{\mathbb{R}^+}(\mathbf{h}))$ is convex in α and minimized at η_s (see Lemma A.10) and, also, $\|\mathbf{w}^*_{dev}\|_2 \geq (1+\varepsilon_{dev})\eta_d \geq \eta_s$. ■

3.7.5.2 Probabilistic result

We now prove the main result of the section, Lemma 3.17.

Proof: [Proof of Lemma 3.17] The proof is based on the results of Lemma 3.18. First, we show that under the assumptions of Lemma 3.17, the assumptions of Lemma 3.18 hold w.h.p.. In this direction, using standard concentration arguments provided in Lemmas A.5 and A.3, we find that,

1. $\|\mathbf{g}\|_2 \geq \text{dist}_{\mathbb{R}^+}(\mathbf{h})$,
2. $\frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\sqrt{\|\mathbf{g}\|_2^2-\text{dist}_{\mathbb{R}^+}(\mathbf{h})^2}} \leq (1+\varepsilon_{dev})\frac{m}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$.
3. For any constant $\varepsilon > 0$,

$$|\|\mathbf{g}\|_2^2-m| \leq \varepsilon m \quad \text{and} \quad |(\text{dist}_{\mathbb{R}^+}(\mathbf{h}))^2-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))| < \varepsilon m, \tag{3.97}$$

all with probability $1-\exp(-\mathscr{O}(m))$. It follows from the first two statements that Lemma 3.18 is applicable and we can use (3.94). Thus, it suffinces to find a lower bound for the right hand side of (3.94).

Lemma A.10 in the Appendix analyzes in detail many properties of the scalar function $L(\alpha;a,b)$, which appears in (3.94). Here, we use the sixth statement of that Lemma (in a similar manner to the proof of Lemma 3.8). In particular, apply Lemma A.10 with the following mapping:

$$\sqrt{m} \iff a,\ \sqrt{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))} \iff b,\ \|\mathbf{g}\|_2 \iff a',\ \text{dist}_{\mathbb{R}^+}(\mathbf{h}) \iff b'$$

Application of the lemma is valid since (3.97) is true, and gives that with probability $1-\exp\left(-\mathscr{O}\left(m\right)\right)$,

$$L((1+\varepsilon)\eta_d, \|\mathbf{g}\|_2, \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})) - L(\eta_s, \|\mathbf{g}\|_2, \mathrm{dist}_{\mathbb{R}^+}(\mathbf{h})) \geq 2\sigma\delta_{dev}\sqrt{m}$$

for some constant δ_{dev}. Combining this with Lemma 3.18, we may conclude

$$\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathscr{L}^*(\mathbf{g},\mathbf{h}) \geq 2\sigma\delta_{dev}\sqrt{m}. \tag{3.98}$$

On the other hand, from Lemma 3.15,

$$|\mathscr{L}^*(\mathbf{g},\mathbf{h}) - \mathbb{E}[\mathscr{L}^*_{up}(\mathbf{g},\mathbf{h})]| \leq \sigma\delta_{dev}\sqrt{m} \tag{3.99}$$

with the desired probability. Union bounding over (3.98) and (3.99), we conclude with the desired result. ■

3.7.6 Merging Upper Bound and Deviation Results

This section combines the previous sections and finalizes the proof of Theorem 3.1 by showing the second statement. Recall the definition (3.3) of the original C-LASSO problem and also the definition of the set S_{dev} in (3.91).

Lemma 3.19 *Assume there exists a constant ε_L such that, $(1-\varepsilon_L)m \geq \mathbf{D}(cone(\partial f(\mathbf{x}_0))) \geq \varepsilon_L m$. Further assume, m is sufficiently large. The following hold:*

1. *For any $\varepsilon_{up} > 0$, there exists $c_{up} > 0$ such that, with probability $1-\exp(-c_{up}m)$, we have,*

$$\mathscr{F}^*_c(\mathbf{A},\mathbf{v}) \leq \mathbb{E}[\mathscr{L}^*_{up}(f,\mathbf{g},\mathbf{h})] + \varepsilon_{up}\sigma\sqrt{m} \tag{3.100}$$

2. *There exists constants $\delta_{dev} > 0, c_{dev} > 0$, such that, for sufficiently large m, with probability $1-\exp(-c_{dev}m)$, we have,*

$$\min_{\|\mathbf{w}\|_2 \in S_{dev},\ f_p(\mathbf{w}) \leq 0} \mathscr{F}(\mathbf{w};\mathbf{A},\mathbf{v}) \geq \mathbb{E}[\mathscr{L}^*_{up}(f,\mathbf{g},\mathbf{h})] + \delta_{dev}\sigma\sqrt{m} \tag{3.101}$$

3. For any $\varepsilon_{dev} > 0$, there exists $c > 0$ such that, with probability $1 - \exp(-cm)$,

$$\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2 \leq \sigma^2(1+\varepsilon_{dev})\frac{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(cone(\partial f(\mathbf{x}_0)))}.$$

Proof: We prove the statements of the lemma in the order that they appear.

1. For notational simplicity denote $\xi = \mathbb{E}[\mathscr{L}_{up}^*(\mathbf{g},\mathbf{h})]$. We combine second statement of Lemma 3.5 with Lemma 3.15. For any constant ε_{up}, we have,

$$\begin{aligned}\mathbb{P}(\mathscr{F}_c^*(\mathbf{A},\mathbf{v}) \leq \xi + 2\sigma\varepsilon_{up}\sqrt{m}) &\geq 2\mathbb{P}(\mathscr{U}^*(\mathbf{g},\mathbf{h}) + \sigma\varepsilon\sqrt{m} \leq \xi + 2\sigma\varepsilon_{up}\sqrt{m}) - 1 - \exp(-\mathscr{O}(m)) \\ &= 2\mathbb{P}(\mathscr{U}^*(\mathbf{g},\mathbf{h}) \leq \xi + \sigma\varepsilon_{up}\sqrt{m}) - 1 - \exp(-\mathscr{O}(m)) \\ &\geq 1 - \exp(-\mathscr{O}(m)),\end{aligned}$$

where we used the first statement of Lemma (3.15) to lower bound the $\mathbb{P}(\mathscr{U}^*(\mathbf{g},\mathbf{h}) \leq \xi + \sigma\varepsilon_{up}\sqrt{m})$.

2. Pick a small constant $\varepsilon > 0$ satisfying $\varepsilon < \frac{\delta_{dev}}{2}$ in the third statement of Lemma 3.5. Now, using Lemma 3.17 and this choice of ε, with probability $1 - \exp(-\mathscr{O}(m))$, we have,

$$\begin{aligned}\mathbb{P}(\min_{\mathbf{w}\in S_{dev},\ f_p(\mathbf{w})\leq 0} \mathscr{F}(\mathbf{w};\mathbf{A},\mathbf{v}) \geq \xi + \sigma\frac{\delta_{dev}}{2}\sqrt{m}) &\geq 2\mathbb{P}(\mathscr{L}_{dev}^*(\mathbf{g},\mathbf{h}) \geq \xi + \sigma\delta_{dev}\sqrt{m} - \varepsilon\sigma\sqrt{m}) - 1 - \exp(-\mathscr{O}(m)) \\ &\geq 1 - \exp(-\mathscr{O}(m)),\end{aligned}$$

where we used (3.93) of Lemma 3.17.

3. Apply Statements 1. and 2. of the lemma, choosing $\varepsilon_{up} = \frac{\delta_{dev}}{8}$. Union bounding we find that

$$\mathbb{P}(\min_{\mathbf{w}\in S_{dev},\ f_p(\mathbf{w})\leq 0} \mathscr{F}(\mathbf{w};\mathbf{A},\mathbf{v}) \geq \mathscr{F}_c^*(\mathbf{A},\mathbf{v}) + \sigma\frac{\delta_{dev}}{4}) \geq 1 - \exp(-\mathscr{O}(m)),$$

which implies with the same probability $\|\mathbf{w}_c^*\|_2 \notin S_{dev}$, i.e., $\|\mathbf{w}_c^*\|_2 \leq (1+\varepsilon_{dev})\sigma\sqrt{\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m - \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}}$.

■

3.8 ℓ_2-LASSO: Regions of Operation

The performance of the ℓ_2-regularized LASSO clearly depends on the particular choice of the parameter λ. A key contribution of this work is that we are able to fully characterize this dependence. In other words, our

analysis predicts the performance of the ℓ_2-LASSO estimator for all values $\lambda \geq 0$. To facilitate our analysis we divide the range $[0,\infty)$ of possible values of λ into three distinct regions. We call the regions $\mathscr{R}_{OFF}$, $\mathscr{R}_{ON}$ and $\mathscr{R}_{\infty}$. Each region has specific performance characteristics and the analysis is the same for all λ that belong to the same region. In this Section, we formally define those distinct regions of operation.The analysis of the value of the NSE for each one of them is then deferred to Section 3.9.

3.8.1 Properties of Distance, Projection and Correlation

For the purpose of defining the distinct regions of operation of the ℓ_2-LASSO, it is first important to explore some useful properties of the Gaussian squared distance $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, projection $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ and correlation $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$. Those quantities are closely related to each other and are of key importance to our analysis. We choose to enlist all their important properties in a single Lemma, which serves as a reference for the rest of the Section.

Lemma 3.20 *Consider fixed $\mathbf{x}_0$ and $f(\cdot)$. Let $\partial f(\mathbf{x}_0)$ be a nonempty, compact set of $\mathbb{R}^n$ that does not contain the origin. Then, the following properties hold*

1. $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))+2\mathbf{C}(\lambda\partial f(\mathbf{x}_0))+\mathbf{P}(\lambda\partial f(\mathbf{x}_0))=n$.
2. $\mathbf{D}_f(\mathbf{x}_0,0)=n$, $\mathbf{P}_f(\mathbf{x}_0,0)=0$, *and* $\mathbf{C}_f(\mathbf{x}_0,0)=0$.
3. $\lim_{\lambda\to\infty}\mathbf{D}(\lambda\partial f(\mathbf{x}_0))=\infty$, $\lim_{\lambda\to\infty}\mathbf{P}(\lambda\partial f(\mathbf{x}_0))=\infty$, *and* $\lim_{\lambda\to\infty}\mathbf{C}(\lambda\partial f(\mathbf{x}_0))=-\infty$.
4. $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$, $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ *and* $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ *are all continuous functions of* $\lambda\geq 0$.
5. $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ *is strictly convex and attains its minimum at a unique point. Denote* λ_{best} *the unique minimizer of* $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$.
6. $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ *is an increasing function for* $\lambda\geq 0$.
7. $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ *is differentiable for* $\lambda>0$. *For* $\lambda>0$,

$$\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda}=-\frac{2}{\lambda}\mathbf{C}(\lambda\partial f(\mathbf{x}_0)).$$

For $\lambda=0$, *interpret* $\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda}$ *as a right derivative.*

8.

$$\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\begin{cases}\geq 0 & ,\lambda\in[0,\lambda_{best}]\\ =0 & ,\lambda=\lambda_{best}\\ \leq 0 & ,\lambda\in[\lambda_{best},\infty)\end{cases}$$

9. $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))+\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ *is strictly decreasing for* $\lambda\in[0,\lambda_{best}]$.

Some of the statements in Lemma 3.20 are easy to prove, while others require more work. Statements 5 and 7 have been recently proved in [4]. We defer the proofs of all statements to Appendix A.6.

3.8.2 Key Values of the Penalty Parameter

We define three key values of the regularizer λ. The main work is devoted to showing that those definitions are well established.

3.8.2.1 λ_{best}

The first key parameter is λ_{best} which was defined in Lemma 3.20 to be the unique minimum of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ over $\lambda\in[0,\infty)$. The rationale behind the subscript "best" associated with this parameter is that the estimation error is minimized for that particular choice of λ. In that sense, λ_{best} is the optimal penalty parameter. We formally prove this fact in Section 3.9, where we explicitly calculate the NSE. In what follows, we assume that $\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{best}})<m$ to ensure that there exists $\lambda\geq 0$ for which estimation of $\mathbf{x}_0$ is robust. Also, observe that, $\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{best}})\leq\mathbf{D}_f(\mathbf{x}_0,0)=n$.

3.8.2.2 λ_{max}

The second key parameter λ_{max} is defined as the unique $\lambda\geq\lambda_{\text{best}}$ that satisfies $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))=m$. We formally repeat this definition in the following Lemma.

Lemma 3.21 *Suppose* $\mathbf{D}_f(\mathbf{x}_0,\lambda_{best})<m$ *and consider the following equation over* $\lambda\geq\lambda_{best}$*:*

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0))=m,\quad \lambda\geq\lambda_{best}. \tag{3.102}$$

Equation (3.102) *has a unique solution, which we denote* λ_{max}.

Proof: We make use of Lemma 3.20. First, we show that equation (3.102) has at most one solution: $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is a strictly convex function of $\lambda \geq 0$ and thus strictly increasing for $\lambda \geq \lambda_{\text{best}}$. Next, we show that (3.102) has at least one solution. From assumption, $\mathbf{D}(\mathbf{x}_0, \lambda_{\text{best}}) < m$. Also, $\lim_{\lambda\to\infty}\mathbf{D}(\mathbf{x}_0, \lambda_{\text{best}}) = \infty$. Furthermore, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is continuous in λ. Combining those facts and using the intermediate value theorem we conclude with the desired result. ■

3.8.2.3 λ_{crit}

The third key parameter λ_{crit} is defined to be the unique $\lambda \leq \lambda_{\text{best}}$ that satisfies $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) = \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ when $m \leq n$ or to be 0 when $m > n$. We formally repeat this definition in the following Lemma.

Lemma 3.22 *Suppose* $\mathbf{D}(\mathbf{x}_0, \lambda_{best}) < m$ *and consider the following equation over* $0 \leq \lambda \leq \lambda_{best}$*:*

$$m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) = \mathbf{C}(\lambda\partial f(\mathbf{x}_0)), \quad 0 \leq \lambda \leq \lambda_{best}. \tag{3.103}$$

- *If* $m \leq n$*, then* (3.103) *has a unique solution, which we denote as* λ_{crit}*.*
- *If* $m > n$*, then* (3.103) *has no solution. Then* $\lambda_{crit} = 0$*.*

Proof: We repeatedly make use of Lemma 3.20. For convenience define the function

$$g(\lambda) = \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0)),$$

for $\lambda \in [0, \lambda_{\text{best}})$. The function $g(\lambda)$ has the following properties over $\lambda \in [0, \lambda_{\text{best}}]$:

- it is strictly decreasing,
- $g(0) = n$,
- $g(\lambda_{\text{best}}) = \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) < m$.

If $m \leq n$, from the intermediate value Theorem it follows that (3.103) has at least one solution. This solution is unique since $g(\lambda)$ is strictly decreasing.

If $m > n$, since $g(\lambda) \leq n$ for all $\lambda \in [0, \lambda_{\text{best}}]$, it is clear that (3.103) has no solution. ■

3.8.3 Regions of Operation: $\mathscr{R}_{OFF}$, $\mathscr{R}_{ON}$, $\mathscr{R}_{\infty}$

Having defined the key parameters λ_{best},λ_{crit} and λ_{max}, we are now ready to define the three distinct regions of operation of the ℓ_2-LASSO problem.

Definition 3.4 *Define the following regions of operation for the ℓ_2-LASSO problem:*

- $\mathscr{R}_{OFF} = \{\lambda \mid 0 \leq \lambda \leq \lambda_{crit}\}$,
- $\mathscr{R}_{ON} = \{\lambda \mid \lambda_{crit} < \lambda < \lambda_{\max}\}$,
- $\mathscr{R}_{\infty} = \{\lambda \mid \lambda \geq \lambda_{\max}\}$.

Remark: The definition of $\mathscr{R}_{\text{ON}}$ in Definition 3.4 is consistent to the Definition in 3.1. In other words, $\lambda_{\text{crit}} \leq \lambda \leq \lambda_{\max}$ if and only if $m \geq \max\{\mathbf{D}(\lambda\partial f(\mathbf{x}_0)), \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}$. This follows after combining Lemmas 3.21 and 3.22 with the Lemma 3.23 below.

Lemma 3.23 *The following hold:*

1. $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ *for all* $\lambda \in \mathscr{R}_{OFF}$ *if* $\lambda_{crit} \neq 0$.
2. $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) > \max\{0, \mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}$ *for all* $\lambda \in \mathscr{R}_{ON}$,
3. $m \leq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ *for all* $\lambda \in \mathscr{R}_{\infty}$.

Proof: We prove the statements in the order they appear. We use Lemma 3.20 throughout.

1. The function $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ is strictly decreasing in $[0, \lambda_{\text{best}}]$. Thus, assuming $\lambda_{\text{crit}} \neq 0$, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0)) \geq \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{crit}}) + \mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{crit}}) = m$ for all $\lambda \in [0, \lambda_{\text{crit}}]$.

2. Since $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is strictly convex, $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is strictly concave and has a unique maximum at λ_{best}. Therefore, for all $\lambda \in [\lambda_{\text{crit}}, \lambda_{\max}]$,

$$m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \geq \max\{ \underbrace{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{crit}})}_{=\mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{crit}}) \geq 0}, \underbrace{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\max})}_{=0} \} \geq 0.$$

Furthermore, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ is strictly decreasing in $[0, \lambda_{\text{best}}]$. Thus, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0)) < \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{crit}}) + \mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{crit}}) \leq m$ for all $\lambda \in (\lambda_{\text{crit}}, \lambda_{\text{best}}]$. For $\lambda \in [\lambda_{\text{best}}, \lambda_{\max})$, we have $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) > 0 \geq \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$.

3. $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is strictly convex. Hence, $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is strictly decreasing in $[\lambda_{\text{best}}, \infty)$. This proves that $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\max}) = 0$ for all $\lambda \geq \lambda_{\max}$.

■

3.9 The NSE of the ℓ_2-LASSO

We split our analysis in three sections, one for each of the three regions $\mathscr{R}_{\text{OFF}}$, $\mathscr{R}_{\text{ON}}$ and $\mathscr{R}_{\infty}$. We start from $\mathscr{R}_{\text{ON}}$, for which the analysis is similar in nature to C-LASSO.

3.9.1 $\mathscr{R}_{ON}$

In this section we prove Theorem 3.2 which characterizes the NSE of the ℓ_2-LASSO in the region $\mathscr{R}_{\text{ON}}$. We repeat the statement of the theorem here, for ease of reference.

Theorem 3.6 (NSE of ℓ_2-LASSO in $\mathscr{R}_{\text{ON}}$) *Assume there exists a constant $\varepsilon_L > 0$ such that $(1-\varepsilon_L)m \geq \max\{\mathbf{D}(\lambda\partial f(\mathbf{x}_0)),\ \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}$ and $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \geq \varepsilon_L m$. Further, assume that m is sufficiently large. Then, for any $\varepsilon > 0$, there exists a constant $C = C(\varepsilon, \varepsilon_L) > 0$ and a deterministic number $\sigma_0 > 0$ (i.e. independent of $\mathbf{A}, \mathbf{v}$) such that, whenever $\sigma \leq \sigma_0$, with probability $1 - \exp(-C\min\{m, \frac{m^2}{n}\})$,*

$$\left| \frac{\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|_2^2}{\sigma^2} \times \frac{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} - 1 \right| < \varepsilon. \tag{3.104}$$

As usual, we first focus on the approximated ℓ_2-LASSO problem in Section 3.9.1.1. Next, in Section 3.9.1.2, we translate this result to the original ℓ_2-LASSO problem.

3.9.1.1 Approximated ℓ_2-LASSO

The approximated ℓ_2-LASSO problem is equivalent to the generic problem (3.41) after taking $\mathscr{C} = \lambda\partial f(\mathbf{x}_0)$. Hence, we simply need to apply the result of Lemma 3.9. with $\mathbf{D}(\mathscr{C})$ and $\mathbf{C}(\mathscr{C})$ corresponding to $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$. We conclude with the following result.

Corollary 3.6 *Let $m \geq \min_{\lambda\geq 0}\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and assume there exists constant $\varepsilon_L > 0$ such that $(1-\varepsilon_L)m \geq \max\{\mathbf{D}(\lambda\partial f(\mathbf{x}_0)),\ \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}$ and $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \geq \varepsilon_L m$. Further assume that m is sufficiently large. Then, for any constants $\varepsilon_1, \varepsilon_2 > 0$, there exist constants $c_1, c_2 > 0$ such that with probability $1 - c_1\exp(-c_2\min\{m, \frac{m^2}{n}\})$,*

$$\left| \frac{\hat{\mathscr{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})}{\sigma\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}} - 1 \right| \leq \varepsilon_1, \tag{3.105}$$

and

$$\left| \frac{\|\hat{\mathbf{w}}_{\ell_2}(\mathbf{A},\mathbf{v})\|_2^2}{\sigma^2} - \frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} \right| \leq \varepsilon_2. \tag{3.106}$$

3.9.1.2 Original ℓ_2-LASSO: Proof of Theorem 3.2

Next, we use Corollary 3.6 to prove Theorem 3.2. To do this, we will first relate $f(\cdot)$ and $\hat{f}(\cdot)$. The following result shows that, $f(\cdot)$ and $\hat{f}(\cdot)$ are close around a sufficiently small neighborhood of $\mathbf{x}_0$.

Proposition 3.2 (Max formula, [21,22]) *Let $f(\cdot):\mathbb{R}^n\to\mathbb{R}$ be a convex and continuous function on $\mathbb{R}^n$. Then, any point $\mathbf{x}$ and any direction $\mathbf{v}$ satisfy,*

$$\lim_{\varepsilon\to 0^+}\frac{f(\mathbf{x}+\varepsilon\mathbf{v})-f(\mathbf{x})}{\varepsilon} = \sup_{\mathbf{s}\in\partial f(\mathbf{x})}\langle \mathbf{s},\mathbf{v}\rangle .$$

In particular, the subdifferential $\partial f(\mathbf{x})$ is nonempty.

Proposition 3.2 considers a fixed direction $\mathbf{v}$, and compares $f(\mathbf{x}_0+\varepsilon\mathbf{v})$ and $\hat{f}(\mathbf{x}_0+\varepsilon\mathbf{v})$. We will need a slightly stronger version which says $\hat{f}(\cdot)$ is a good approximation of $f(\cdot)$ at all directions simultaneously. The following proposition is a restatement of Lemma 2.1.1 of Chapter VI of [116].

Proposition 3.3 (Uniform max formula) *Assume $f(\cdot):\mathbb{R}^n\to\mathbb{R}$ is convex and continuous on $\mathbb{R}^n$ and $\mathbf{x}_0\in\mathbb{R}^n$. Let $\hat{f}(\cdot)$ be the first order approximation of $f(\cdot)$ around $\mathbf{x}_0$ as defined in (3.11). Then, for any $\delta>0$, there exists $\varepsilon>0$ such that,*

$$f(\mathbf{x}_0+\mathbf{w})-\hat{f}(\mathbf{x}_0+\mathbf{w})\leq\delta\|\mathbf{w}\|_2, \tag{3.107}$$

for all $\mathbf{w}\in\mathbb{R}^n$ with $\|\mathbf{w}\|_2\leq\varepsilon$.

Recall that we denote the minimizers of the ℓ_2-LASSO and approximated ℓ_2-LASSO by $\mathbf{w}^*_{\ell_2}$ and $\hat{\mathbf{w}}_{\ell_2}$, respectively. Also, for convenience denote,

$$\eta_{\ell_2} = \frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}.$$

After Corollary 3.6, $\|\hat{\mathbf{w}}_{\ell_2}\|_2^2$ concentrates around $\sigma^2\eta_{\ell_2}$. We will argue that, in the small noise regime, we can translate our results to the original problem in a smooth way. Assume that the statements of Corollary 3.6 hold with high probability for some arbitrary $\varepsilon_1,\varepsilon_2>0$. It suffices to prove that for any $\varepsilon_3>0$ there

exists $\sigma_0 > 0$ such that

$$\left| \frac{\|\mathbf{w}^*_{\ell_2}\|_2^2}{\sigma^2} - \eta_{\ell_2} \right| \leq \varepsilon_3, \tag{3.108}$$

for all $\sigma < \sigma_0$. To begin with, fix a $\delta > 0$, the value of which is to be determined later in the proof. As an immediate implication of Proposition 3.3, there exists σ_0 such that

$$f(\mathbf{x}_0 + \mathbf{w}) - \hat{f}(\mathbf{x}_0 + \mathbf{w}) \leq \delta \|\mathbf{w}\|_2 \tag{3.109}$$

for all $\mathbf{w}$ satisfying $\|\mathbf{w}\|_2 \leq C = C(\sigma_0, \varepsilon_2) := \sigma_0 \sqrt{(1+\varepsilon_2)\eta_{\ell_2}}$. Now, fix any $\sigma < \sigma_0$. We will make use of the fact that the following three events hold with high probability.

- Using Corollary 3.6, with high probability $\hat{\mathbf{w}}_{\ell_2}$ satisfies,

$$\|\hat{\mathbf{w}}_{\ell_2}\|_2 \leq \sigma \sqrt{(1+\varepsilon_2)\eta_{\ell_2}} \leq C. \tag{3.110}$$

- Using (3.52) of Lemma 3.9 with $\mathscr{C} = \lambda \partial f(\mathbf{x}_0)$, there exists a constant $t = t(\varepsilon_3)$ so that for any $\mathbf{w}$ satisfying $|\frac{\|\mathbf{w}\|_2^2}{\sigma^2} - \eta_{\ell_2}| \geq \varepsilon_3$, we have,

$$\|\mathbf{A}\mathbf{w} - \sigma \mathbf{v}\|_2 + \max_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w} \geq \hat{\mathscr{F}}_{\ell_2}(\mathbf{A}, \mathbf{v}) + t(\varepsilon_3)\sigma\sqrt{m}. \tag{3.111}$$

Combine (3.110) with (3.109) to find that

$$\begin{aligned} \|\mathbf{A}\hat{\mathbf{w}}_{\ell_2} - \sigma \mathbf{v}\|_2 + \lambda(f(\mathbf{x}_0 + \hat{\mathbf{w}}_{\ell_2}) - f(\mathbf{x}_0)) &\leq \underbrace{\|\mathbf{A}\hat{\mathbf{w}}_{\ell_2} - \sigma \mathbf{v}\|_2 + \lambda(\hat{f}(\mathbf{x}_0 + \hat{\mathbf{w}}_{\ell_2}) - f(\mathbf{x}_0))}_{=\hat{\mathscr{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})} + \delta \|\hat{\mathbf{w}}_{\ell_2}\|_2 \\ &\leq \hat{\mathscr{F}}_{\ell_2}(\mathbf{A}, \mathbf{v}) + \delta\sigma\sqrt{(1+\varepsilon_2)\eta_{\ell_2}}. \end{aligned} \tag{3.112}$$

Now, assume that $\|\mathbf{w}^*_{\ell_2}\|_2$ does not satisfy (3.108). Then,

$$\|\mathbf{A}\hat{\mathbf{w}}_{\ell_2} - \sigma \mathbf{v}\|_2 + \lambda(f(\mathbf{x}_0 + \hat{\mathbf{w}}_{\ell_2}) - f(\mathbf{x}_0)) \geq \mathscr{F}^*_{\ell_2}(\mathbf{A}, \mathbf{v}) \tag{3.113}$$

$$\geq \|\mathbf{A}\mathbf{w}^*_{\ell_2} - \sigma \mathbf{v}\|_2 + \lambda \max_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w}^*_{\ell_2} \tag{3.114}$$

$$\geq \hat{\mathscr{F}}_{\ell_2}(\mathbf{A}, \mathbf{v}) + t(\varepsilon_3)\sigma\sqrt{m}. \tag{3.115}$$

(3.113) follows from optimality of $\mathbf{w}^*_{\ell_2}$. For (3.114) we used convexity of $f(\cdot)$ and the basic property of the subdifferential that $f(\mathbf{x}_0+\mathbf{w}) \geq f(\mathbf{x}_0)+\mathbf{s}^T\mathbf{w}$, for all $\mathbf{w}$ and $\mathbf{s} \in \partial f(\mathbf{x}_0)$. Finally, (3.115) follows from (3.111).

To complete the proof, choose $\delta < \frac{t\sqrt{m}}{\sqrt{(1+\varepsilon_2)}\eta_{\ell_2}}$. This will result in contradiction between (3.112) and (3.115). Observe that, our choice of δ and σ_0 is deterministic and depends on $m, \mathbf{x}_0, f(\cdot), \varepsilon_3$.

3.9.1.3 A Property of the NSE Formula

Theorem 3.2 shows that the asymptotic NSE formula in $\mathscr{R}_{\text{ON}}$ is $\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$. The next lemma provides a useful property of this formula as a function of λ on $\mathscr{R}_{\text{ON}}$.

Lemma 3.24 *$\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ is a convex function of λ over $\mathscr{R}_{ON}$.*

Proof: From 3.20, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is a strictly convex function of λ. Also, $\frac{x}{m-x}$ is an increasing function of x over $0 \leq x < m$ and its second derivative is $\frac{m}{(m-x)^3}$ which is strictly positive over $\mathscr{R}_{\text{ON}}$. Consequently, the asymptotic NSE formula is a composition of an increasing convex function with a convex function, and is thus itself convex [23]. ■

3.9.2 $\mathscr{R}_{OFF}$

Our analysis, unfortunately, does not extend to $\mathscr{R}_{OFF}$, and we have no proof that characterizes the NSE in this regime. On the other hand, our extensive numerical experiments (see Section 3.14) show that, in this regime, the optimal estimate $\mathbf{x}^*_{\ell_2}$ of (3.4) satisfies $\mathbf{y} = \mathbf{A}\mathbf{x}^*_{\ell_2}$. Observe that, in this case, the ℓ_2-LASSO reduces to the standard approach taken for the noiseless compressed sensing problem,

$$\min f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{3.116}$$

Here, we provide some intuition to why it is reasonable to expect this to be the case. Recall that $\lambda \in \mathscr{R}_{\text{OFF}}$ iff $0 \leq \lambda \leq \lambda_{\text{crit}}$, and so the "small" values of the penalty parameter λ are in $\mathscr{R}_{\text{OFF}}$. As λ gets smaller, $\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2$ becomes the dominant term, and ℓ_2-LASSO penalizes this term more. So, at least for sufficiently small λ, the reduction to problem (3.116) would not be surprising. Lemma 3.25 formalizes this idea for the small λ regime.

Lemma 3.25 *Assume $m \leq \alpha n$ for some constant $\alpha < 1$ and $f(\cdot)$ is a Lipschitz continuous function with Lipschitz constant $L > 0$. Then, for $\lambda < \frac{\sqrt{n}-\sqrt{m}}{L}(1-o(1))$, the solution $\mathbf{x}^*_{\ell_2}$ of ℓ_2-LASSO satisfies $\mathbf{y} = \mathbf{A}\mathbf{x}^*_{\ell_2}$, with probability $1-\exp(-\mathscr{O}(n))$. Here, $o(1)$ term is arbitrarily small positive constant.*

Proof: When $m \leq \alpha n$ for some constant $0 < \alpha < 1$, $\sqrt{n}-\sqrt{m} = \mathscr{O}(\sqrt{n})$. Then, from standard concentration results (see [213]), with probability $1-\exp(-\mathscr{O}(n))$, minimum singular value $\sigma_{min}(\mathbf{A})$ of $\mathbf{A}$ satisfies

$$\frac{\sigma_{min}(\mathbf{A}^T)}{\sqrt{n}-\sqrt{m}} \geq 1-o(1).$$

Take any $\lambda < \frac{\sqrt{n}-\sqrt{m}}{L}(1-o(1))$ and let $\mathbf{p} := \mathbf{y} - \mathbf{A}\mathbf{x}^*_{\ell_2}$. We will prove that $\|\mathbf{p}\|_2 = 0$. Denote $\mathbf{w}_2 := \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{p}$. Using (3.117), with the same probability,

$$\|\mathbf{w}_2\|_2^2 = \mathbf{p}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{p} \leq \frac{\|\mathbf{p}\|_2^2}{(\sigma_{min}(\mathbf{A}^T))^2} \leq \frac{\|\mathbf{p}\|_2^2}{((\sqrt{n}-\sqrt{m})(1-o(1)))^2}, \tag{3.117}$$

Define $\mathbf{x}_2 = \mathbf{x}^*_{\ell_2} + \mathbf{w}_2$, for which $\mathbf{y} - \mathbf{A}\mathbf{x}_2 = 0$ and consider the difference between the ℓ_2-LASSO costs achieved by the minimizer $\mathbf{x}^*_{\ell_2}$ and $\mathbf{x}_2$. From optimality of $\mathbf{x}^*_{\ell_2}$, we have,

$$\begin{aligned} 0 &\geq \|\mathbf{p}\|_2 + \lambda f(\mathbf{x}^*_{\ell_2}) - \lambda f(\mathbf{x}_2) \\ &\geq \|\mathbf{p}\|_2 - \lambda L\|\mathbf{x}^*_{\ell_2} - \mathbf{x}^*_2\|_2 = \|\mathbf{p}\|_2 - \lambda L\|\mathbf{w}_2\|_2 \qquad (3.118) \\ &\geq \|\mathbf{p}\|_2(1 - \lambda\frac{L}{(\sqrt{n}-\sqrt{m})(1-o(1))}). \qquad (3.119) \end{aligned}$$

The inequality in (3.118) follows from Lipschitzness of $f(\cdot)$, while we use (3.117) to find (3.119). For the sake of contradiction, assume that $\|\mathbf{p}\|_2 \neq 0$, then (3.119) reduces to $0 > 0$, clearly, a contradiction. ■

For an illustration of Lemma 3.25, consider the case where $f(\cdot) = \|\cdot\|_1$. ℓ_1-norm is Lipschitz with $L = \sqrt{n}$ (see [168] for related discussion). Lemma 3.25 would, then, require $\lambda < 1 - \sqrt{\frac{m}{n}}$ to be applicable. As an example, considering the setup in Figure 3.3, Lemma 3.25 would yield $\lambda < 1 - \sqrt{\frac{1}{2}} \approx 0.292$ whereas $\lambda_{\text{crit}} \approx 0.76$. While Lemma 3.25 supports our claims on $\mathscr{R}_{\text{OFF}}$, it does not say much about the exact location of the transition point, at which the ℓ_2-LASSO reduces to (3.116). We claim this point is $\lambda = \lambda_{\text{crit}}$.

3.9.3 $\mathscr{R}_\infty$

In this region $m \leq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. In this region, we expect *no* noise robustness, namely, $\frac{\|\mathbf{x}^*_{\ell_2}-\mathbf{x}_0\|_2^2}{\sigma^2} \to \infty$ as $\sigma \to 0$. In this work, we show this under a stricter assumption, namely, $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. See Theorem

3.4 and Section 3.13 for more details. Our proof method relies on results of [4] rather than Gaussian Min-Max Theorem. On the other hand, we believe, a more careful study of Gaussian comparison inequalities (Proposition 2.7) can give the desired result for the wider regime $m < \mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. We leave this as a future work.

3.10 Nonasymptotic results on ℓ_2-LASSO

The main result of this section is a closed and non asymptotic bound which approximately matches to what one would expect from Theorem 3.2. Rather remarkably, this bound holds for all $\lambda \geq 0$.

Theorem 3.7 *Assume $m \geq 2$, $\mathbf{z} \in \mathbb{R}^m$ and $\mathbf{x}_0 \in \mathbb{R}^n$ are arbitrary, and, $\mathbf{A} \in \mathbb{R}^{m \times n}$ has i.i.d $\mathcal{N}(0,1)$ entries. Fix the regularizer parameter in (3.4) to be $\lambda \geq 0$ and let $\hat{\mathbf{x}}$ be a minimizer of (3.4). Then, for any $0 < t \leq (\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))})$, with probability $1 - 5\exp(-t^2/32)$, we have,*

$$\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2 \leq 2 \frac{\|\mathbf{z}\|_2}{\sqrt{m}} \frac{\sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} + t}{\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))} - t}. \tag{3.120}$$

3.10.1 Interpretation

Theorem 3.7 provides a simple, general, non-asymptotic and (rather) sharp upper bound on the error of the regularized lasso estimator (3.4), which also takes into account the specific choice of the regularizer parameter $\lambda \geq 0$. In principle, the bound applies to any signal class that exhibits some sort of low-dimensionality (see [169] and references therein). It is non-asymptotic and is applicable in any regime of m, λ and $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$. Also, the constants involved in it are small making it rather tight[5].

The Gaussian distance term $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$ summarizes the geometry of the problem and is key in (3.120). In [4] (Proposition 4.4), it is proven that $\mathbf{D}(\lambda \partial f(\mathbf{x}_0))$, when viewed as a function of $\lambda \geq 0$, is strictly convex, differentiable for $\lambda > 0$ and achieves its minimum at a unique point. Figure 3.5 illustrates this behavior; $\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda \partial f(\mathbf{x}_0))}$ achieves its unique maximum value at some $\lambda = \lambda_{\text{best}}$, it is strictly increasing for $\lambda < \lambda_{\text{best}}$ and strictly decreasing for $\lambda > \lambda_{\text{best}}$. For the bound in (3.120) to be at all meaningful, we require $m > \min_{\lambda \geq 0} \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) = \mathbf{D}(\lambda_{\text{best}} \partial f(\mathbf{x}_0))$. This is perfectly in line with our discussion so far, and translates to the number of measurements being large enough to at least guarantee noiseless

[5] We suspect and is also supported by our simulations (e.g. Figure 3.6) that the factor of 2 in (3.120) is an artifact of our proof technique and not essential.

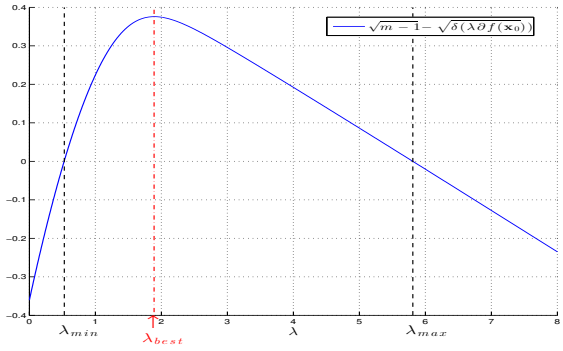


Figure 3.5: Illustration of the denominator $\sqrt{m-1}-\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ in (3.120) as a function of $\lambda \geq 0$. The bound is meaningful for $\lambda \in (\lambda_{min}, \lambda_{max})$ and attains its minimum value at λ_{best}. The y-axis is normalized by $\sqrt{n}$.

recovery [4, 50, 81, 101, 167]. Lemma 3.21 in Section 3.8 proves that there exists a unique λ_{max} satisfying $\lambda_{max} > \lambda_{\text{best}}$ and $\sqrt{\mathbf{D}(\lambda_{max}\partial f(\mathbf{x}_0))} = \sqrt{m-1}$. Similarly, when $m \leq n$, there exists unique $\lambda_{min} < \lambda_{\text{best}}$ satisfying $\sqrt{\mathbf{D}(\lambda_{min}\partial f(\mathbf{x}_0))} = \sqrt{m-1}$. From this, it follows that $\sqrt{m-1} > \sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ if and only if $\lambda \in (\lambda_{min}, \lambda_{max})$. This is exactly the range of values of the regularizer parameter λ for which (3.120) is meaningful; see also Figure 3.5.

The region $(\lambda_{min}, \lambda_{max})$, which our bound characterizes, contains λ_{best}, for which, the bound in (3.120) achieves its minimum value since it is strictly increasing in $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. Note that deriving λ_{best} does not require knowledge of any properties (e.g. variance) of the noise vector. All it requires is knowledge of the particular structure of the unknown signal. For example, in the ℓ_1-case, λ_{best} depends only on the sparsity of $\mathbf{x}_0$, not $\mathbf{x}_0$ itself, and in the nuclear norm case, it only depends on the rank of $\mathbf{x}_0$, not $\mathbf{x}_0$ itself.

3.10.2 Comparison to related work

3.10.2.1 Sparse estimation

Belloni et al. [16] were the first to prove error guarantees for the ℓ_2-lasso (3.4). Their analysis shows that the estimation error is of order $O\left(\sqrt{\frac{k\log(n)}{m}}\right)$, when $m = \Omega(k\log n)$ and $\lambda > \sqrt{2\log(2n)}$[6]. Recalling Table 3.3 for sparsity with $\lambda = \sqrt{2\log(\frac{n}{k})}$ and applying Theorem 3.7 yields the same order-wise error guarantee. Our result is non-asymtpotic and involves explicit coefficients, while the result of [16] is applicable to more general constructions of the measurement matrix $\mathbf{A}$.

[6] [16] also imposes a "growth restriction" on λ, which agrees with the fact that our bound becomes vacuous for $\lambda > \lambda_{\max}$ (see Section 3.10.1).

3.10.2.2 Sharp error bounds

In Section 3.9, we performed a detailed analysis of the regularized lasso problem (3.4) under the additional assumption that the entries of the noise vector $\mathbf{z}$ are distributed $\mathcal{N}(0,\sigma^2)$. In particular, when $\sigma \to 0$ and m is large enough, they prove that with high probability,

$$\|\hat{\mathbf{x}}-\mathbf{x}_0\|_2 \approx \|\mathbf{z}\|_2 \frac{\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}, \tag{3.121}$$

for λ belonging to a particular subset of $(\lambda_{min},\lambda_{max})$. As expected, our bound in Theorem 3.7 is larger than the term in (3.121). However, apart from a factor of 2, it only differs from the quantity in (3.121) in the denominator, where instead of $\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$, we have the smaller $\sqrt{m-1}-\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$. This difference becomes insignificant and indicates that our bound is rather tight when m is large. Although in Theorem 3.2 we conjecture that (3.121) upper bounds the estimation error for arbitrary values of the noise variance σ^2, we will not prove so. In that sense, and to the best of our knowledge, Theorem 3.7 is the first rigorous upper bound on the estimation error of (3.4), which holds for general convex regularizers, is non-asymptotic and requires no assumption on the distribution of $\mathbf{z}$.

3.10.3 Simulation results

Figure 3.6 illustrates the bound of Theorem 3.7, which is given in red for $n=340$, $m=140$, $k=10$ and for $\mathbf{A}$ having $\mathcal{N}(0,\frac{1}{m})$ entries. The upper bound from Section 3.9 is asymptotic in m and only applies to i.i.d Gaussian $\mathbf{z}$, is given in black. In our simulations, we assume $\mathbf{x}_0$ is a random unit norm vector over its support and consider both i.i.d $\mathcal{N}(0,\sigma^2)$, as well as, non-Gaussian noise vectors $\mathbf{z}$. We have plotted the realizations of the normalized error for different values of λ and σ. As noted, the bound in Section 3.9 is occasionally violated since it requires very large m, as well as, i.i.d Gaussian noise. On the other hand, the bound given in (3.120) always holds.

3.11 Proof of Theorem 3.7

It is convenient to rewrite (3.4) in terms of the error vector $\mathbf{w}=\mathbf{x}-\mathbf{x}_0$ as follows:

$$\min_{\mathbf{w}} \|\mathbf{A}\mathbf{w}-\mathbf{z}\|_2+\lambda(f(\mathbf{x}_0+\mathbf{w})-f(\mathbf{x}_0)). \tag{3.122}$$

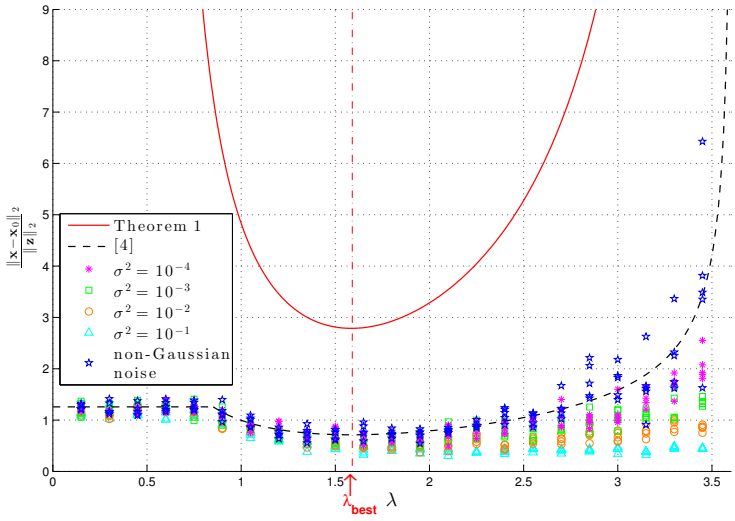


Figure 3.6: The normalized error of (3.4) as a function of λ.

Denote the solution of (3.122) by $\hat{\mathbf{w}}$. Then, $\hat{\mathbf{w}} = \hat{\mathbf{x}} - \mathbf{x}_0$ and (3.120) bounds $\|\hat{\mathbf{w}}\|_2$. To simplify notation, for the rest of the proof, we denote the value of that upper bound as

$$\ell(t) := 2\frac{\|\mathbf{z}\|_2}{\sqrt{m}} \frac{\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} + t}{\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} - t}. \tag{3.123}$$

It is easy to see that the optimal value of the minimization in (3.122) is no greater than $\|\mathbf{z}\|_2$. Observe that $\mathbf{w} = \mathbf{0}$ achieves this value. However, Lemma 3.26 below shows that if we constrain the minimization in (3.122) to be only over vectors $\mathbf{w}$ whose norm is greater than $\ell(t)$, then the resulting optimal value is (with high probability on the measurement matrix $\mathbf{A}$) strictly greater than $\|\mathbf{z}\|_2$. Combining those facts yields the desired result, namely $\|\hat{\mathbf{w}}\|_2 \leq \ell(t)$. Thus, it suffices to prove Lemma 3.26.

Lemma 3.26 *Fix some $\lambda \geq 0$ and $0 < t \leq (\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))})$. Let $\ell(t)$ be defined as in (3.123). Then, with probability $1 - 5\exp(-t^2/32)$, we have,*

$$\min_{\|\mathbf{w}\|_2 \geq \ell(t)} \{\|\mathbf{A}\mathbf{w} - \mathbf{z}\|_2 + \lambda(f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0))\} > \|\mathbf{z}\|_2. \tag{3.124}$$

3.11.1 Proof of Lemma 3.26

Fix λ and t, as in the statement of the lemma. From the convexity of $f(\cdot)$, $f(\mathbf{x}_0+\mathbf{w})-f(\mathbf{x}_0) \geq \max_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}$. Hence, it suffices to prove that w.h.p. over $\mathbf{A}$,

$$\min_{\|\mathbf{w}\|_2\geq\ell(t)} \{\|\mathbf{A}\mathbf{w}-\mathbf{z}\|_2 + \max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}\} > \|\mathbf{z}\|_2.$$

We begin with applying Gordon's Proposition 2.7 to the optimization problem in the expression above. Rewrite $\|\mathbf{A}\mathbf{w}-\mathbf{z}\|_2$ as $\max_{\|\mathbf{a}\|_2=1}\{\mathbf{a}^T\mathbf{A}\mathbf{w}-\mathbf{a}^T\mathbf{z}\}$ and, then, apply Proposition 2.7 with $\mathbf{G}=\mathbf{A}$, $\mathscr{S}=\{\mathbf{w} \mid \|\mathbf{w}\|_2 \geq \ell(t)\}$ and $\psi(\mathbf{w},\mathbf{a}) = -\mathbf{a}^T\mathbf{z} + \max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}$. This leads to the following statement:

$$\mathbb{P}(\text{ (3.124) is true }) \geq 2\cdot\mathbb{P}(\ \mathscr{L}(t;\mathbf{g},\mathbf{h}) > \|\mathbf{z}\|_2\ ) - 1,$$

where, $\mathscr{L}(t;\mathbf{g},\mathbf{h})$ is defined as

$$\min_{\|\mathbf{w}\|_2\geq\ell(t)} \max_{\|\mathbf{a}\|_2=1} \{(\|\mathbf{w}\|_2\mathbf{g}-\mathbf{z})^T\mathbf{a} - \min_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} (\mathbf{h}-\mathbf{s})^T\mathbf{w}\}. \tag{3.125}$$

In the remaining, we analyze the simpler optimization problem defined in (3.125), and prove that $\mathscr{L}(t;\mathbf{g},\mathbf{h}) > \|\mathbf{z}\|_2$ holds with probability $1-\frac{5}{2}\exp(-t^2/32)$. We begin with simplifying the expression for $\mathscr{L}(t;\mathbf{g},\mathbf{h})$, as follows:

$$\begin{aligned}
\mathscr{L}(t;\mathbf{g},\mathbf{h}) &= \min_{\|\mathbf{w}\|_2\geq\ell(t)} \{\|\|\mathbf{w}\|_2\mathbf{g}-\mathbf{z}\|_2 - \min_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} (\mathbf{h}-\mathbf{s})^T\mathbf{w}\} \\
&= \min_{\alpha\geq\ell(t)} \{\|\alpha\mathbf{g}-\mathbf{z}\|_2 - \alpha\text{dist}(\mathbf{h},\lambda\partial f(\mathbf{x}_0))\} \\
&= \min_{\alpha\geq\ell(t)} \{\sqrt{\alpha^2\|\mathbf{g}\|_2^2+\|\mathbf{z}\|_2^2-2\alpha\mathbf{g}^T\mathbf{z}} - \alpha\text{dist}(\mathbf{h},\lambda\partial f(\mathbf{x}_0))\}.
\end{aligned} \tag{3.126}$$

The first equality above follows after performing the trivial maximization over $\mathbf{a}$ in (3.125). The second, uses the fact that $\max_{\|\mathbf{w}\|_2=\alpha} \min_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} (\mathbf{h}-\mathbf{s})^T\mathbf{w} = \min_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \max_{\|\mathbf{w}\|_2=\alpha} (\mathbf{h}-\mathbf{s})^T\mathbf{w} = \alpha\cdot\text{dist}(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$, for all $\alpha\geq 0$. For a proof of this see Section A.4.

Next, we show that $\mathscr{L}(t;\mathbf{g},\mathbf{h})$ is strictly greater than $\|\mathbf{z}\|_2$ with the desired high probability over realiza-

tions of $\mathbf{g}$ and $\mathbf{h}$. Consider the event $\mathscr{E}_t$ of $\mathbf{g}$ and $\mathbf{h}$ satisfying all three conditions listed below,

$$1.\ \|\mathbf{g}\|_2 \geq \gamma_m - t/4, \tag{3.127a}$$

$$2.\ \mathrm{dist}(\mathbf{h}, \lambda\partial f(\mathbf{x}_0)) \leq \sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} + t/4, \tag{3.127b}$$

$$3.\ \mathbf{g}^T\mathbf{z} \leq (t/4)\|\mathbf{z}\|_2. \tag{3.127c}$$

In (3.127a) we have denoted $\gamma_m := \mathbb{E}[\|\mathbf{g}\|_2]$; it is well known that $\gamma_m = \sqrt{2}\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})}$ and $\gamma_m \leq \sqrt{m}$. The conditions in (3.127) hold with high probability. In particular, the first two hold with probability no less than $1-\exp(-t^2/32)$. This is because the ℓ_2-norm and the distance function to a convex set are both 1-Lipschitz functions and, thus, Fact 2.4 applies. The third condition holds with probability at least $1-(1/2)\exp(-t^2/32)$, since $\mathbf{g}^T\mathbf{z}$ is statistically identical to $\mathscr{N}(0,\|\mathbf{z}\|_2^2)$. Union bounding yields,

$$\mathbb{P}(\mathscr{E}_t) \geq 1-(5/2)\exp(-t^2/32). \tag{3.128}$$

Furthermore, Lemma 3.27, below, shows that if $\mathbf{g}$ and $\mathbf{h}$ are such that $\mathscr{E}_t$ is satisfied, then $\mathscr{L}(t;\mathbf{g},\mathbf{h}) > \|\mathbf{z}\|_2$. This, when combined with (3.128) shows that $\mathbb{P}(\mathscr{L}(t;\mathbf{g},\mathbf{h}) > \|\mathbf{z}\|_2) \geq 1-(5/2)\exp(-t^2/32)$, completing the proof of Lemma 3.26.

Lemma 3.27 *Fix any* $0 < t \leq (\sqrt{m-1} - \sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))})$. *Suppose* $\mathbf{g}$ *and* $\mathbf{h}$ *are such that* (3.127) *holds and recall the definition of* $\mathscr{L}(t;\mathbf{g},\mathbf{h})$ *in* (3.126). *Then,* $\mathscr{L}(t;\mathbf{g},\mathbf{h}) > \|\mathbf{z}\|_2$.

Proof: Take any $\alpha \geq \ell(t) > 0$. Following from (3.127), we have that the objective function of the optimization in (3.126) is lower bounded by

$$\phi(\alpha) = \sqrt{\alpha^2(\gamma_m - \frac{t}{4})^2 + \|\mathbf{z}\|_2^2 - \frac{1}{2}\alpha\|\mathbf{z}\|_2 t} - \alpha(\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} + \frac{t}{4}).$$

We will show that $\phi(a) > \|\mathbf{z}\|_2$, for all $\alpha \geq \ell(t)$, and this will complete the proof. Starting with the desired condition $\phi(\alpha) > \|\mathbf{z}\|_2$, using the fact that $\alpha > 0$ and performing some algebra, we have the following

equivalences,

$$\begin{aligned}\phi(a) > \|\mathbf{z}\|_2 \Leftrightarrow \alpha^2(\gamma_m - t/4)^2 + \|\mathbf{z}\|_2^2 - (1/2)\alpha\|\mathbf{z}\|_2 t > \\ (\alpha(\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} + t/4) + \|\mathbf{z}\|_2)^2 \\ \Leftrightarrow \alpha > \frac{2\|\mathbf{z}\|_2(\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} + t/2)}{\gamma_m^2 - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) - \frac{t}{2}(\gamma_m + \sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))})}.\end{aligned} \tag{3.129}$$

Observing that $\gamma_m^2 > \sqrt{m}\sqrt{m-1}$ [86], $\gamma_m \leq \sqrt{m}$ and $\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))} < \sqrt{m}$, it can be shown that $\ell(t)$ is strictly greater than the expression in the right hand side of (3.129). Thus, for all $\alpha \geq \ell(t)$, we have $\phi(\alpha) > \|\mathbf{z}\|_2$, as desired.

■

3.12 ℓ_2^2-LASSO

As we have discussed throughout our main results, one of the critical contributions of this chapter is that, we are able to obtain a formula that predicts the performance of ℓ_2^2-penalized LASSO. We do this by relating ℓ_2-LASSO and ℓ_2^2-LASSO problems. This relation is established by creating a mapping between the penalty parameters λ and τ. While we don't give a theoretical guarantee on ℓ_2^2-LASSO, we give justification based on the predictive power of Gaussian Min-Max Theorem.

3.12.1 Mapping the ℓ_2-penalized to the ℓ_2^2-penalized LASSO problem

Our aim in this section is to provide justification for the mapping function given in (3.20). The following lemma gives a simple condition for ℓ_2-LASSO and ℓ_2^2-LASSO to have the same solution.

Lemma 3.28 *Let $\mathbf{x}_{\ell_2}^*$ be a minimizer of ℓ_2-LASSO program with the penalty parameter λ and assume $\mathbf{y} - \mathbf{A}\mathbf{x}_{\ell_2}^* \neq 0$. Then, $\mathbf{x}_{\ell_2}^*$ is a minimizer of ℓ_2^2-LASSO with penalty parameter $\tau = \lambda \cdot \frac{\|\mathbf{A}\mathbf{x}_{\ell_2}^* - \mathbf{y}\|_2}{\sigma}$.*

Proof: The optimality condition for the ℓ_2^2-LASSO problem (3.4), implies the existence of $\mathbf{s}_{\ell_2} \in \partial f(\mathbf{x}_{\ell_2}^*)$ such that,

$$\lambda\mathbf{s}_{\ell_2} + \frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}_{\ell_2}^* - \mathbf{y})}{\|\mathbf{A}\mathbf{x}_{\ell_2}^* - \mathbf{y}\|_2} = 0 \tag{3.130}$$

On the other hand, from the optimality conditions of (3.5), $\mathbf{x}$ is a minimizer of the ℓ_2^2-LASSO if there exists $\mathbf{s} \in \partial f(\mathbf{x})$ such that,

$$\sigma\tau\mathbf{s} + \mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{y}) = 0. \tag{3.131}$$

Observe that, for $\tau = \lambda \cdot \frac{\|\mathbf{A}\mathbf{x}^*_{\ell_2} - \mathbf{y}\|_2}{\sigma}$, using (3.130), $\mathbf{x}^*_{\ell_2}$ satisfies (3.131) and is thus a minimizer of the ℓ_2^2-LASSO. ■

In order to evaluate the mapping function as proposed in Lemma 3.28, we need to estimate $\|\mathbf{y} - \mathbf{A}\mathbf{x}^*_{\ell_2}\|_2$. We do this relying again on the approximated ℓ_2-LASSO problem in (3.31). Under the first-order approximation, $\mathbf{x}^*_{\ell_2} \approx \mathbf{x}_0 + \hat{\mathbf{w}}^*_{\ell_2} := \hat{\mathbf{x}}_{\ell_2^*}$ and also define, $\hat{f}_p(\mathbf{w}) := \sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}$. Then, from (3.31) and Lemma 3.9,

$$\begin{aligned}\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}^*_{\ell_2}\|_2 &= \hat{\mathscr{F}}^*_{\ell_2}(\mathbf{A}, \mathbf{v}) - \lambda\hat{f}_p(\hat{\mathbf{w}}^*_{\ell_2}) \\ &\approx \sigma\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))} - \lambda\hat{f}_p(\hat{\mathbf{w}}^*_{\ell_2}).\end{aligned} \tag{3.132}$$

Arguing that,

$$\lambda\hat{f}_p(\mathbf{w}^*_{\ell_2}) \approx \sigma\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}, \tag{3.133}$$

and substituting this in (3.132) will result in the desired mapping formula given in (3.20).

In the remaining lines we provide justification supporting our belief that (3.133) is true. Not surprisingly at this point, the core of our argument relies on application of "modified Gordon's Lemma" Proposition 2.7. Following the lines of our discussion in Section 3.5, we use the minimizer $\mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h})$ of the simple optimization (3.13) as a proxy for $\mathbf{w}^*_{\ell_2}$ and expect $\hat{f}_p(\mathbf{w}^*_{\ell_2})$ to concentrate around the same quantity as $\hat{f}_p(\mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h}))$ does. Lemma 3.29 below shows that

$$\begin{aligned}\lambda\hat{f}_p(\mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h})) &= \sigma\frac{\langle\Pi(\mathbf{h}, \lambda\partial f(\mathbf{x}_0)), \mathrm{Proj}(\mathbf{h}, \lambda\partial f(\mathbf{x}_0))\rangle}{\sqrt{\|\mathbf{g}\|_2^2 - \mathrm{dist}(\mathbf{h}, \lambda\partial f(\mathbf{x}_0))^2}} \\ &\approx \sigma\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}},\end{aligned}$$

where the second (approximate) equality follows via standard concentration inequalities.

Lemma 3.29 *Assume* $(1 - \varepsilon_L)m \geq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ *and* m *is sufficiently large. Then, for any constant* $\varepsilon > 0$,

with probability $1-\exp(-\mathscr{O}\left(\min\{m,\frac{m^2}{n}\}\right))$,

$$\left|\lambda\hat{f}_p(\mathbf{w}^*_{low})-\sigma\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}\right|<\varepsilon\sqrt{m}. \tag{3.134}$$

Proof: Recall that $\mathbf{w}^*_{low}(\mathbf{g},\mathbf{h})=\sigma\frac{\Pi(\mathbf{h},\mathscr{C})}{\sqrt{\|\mathbf{g}\|_2^2-\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))}}$ for $\mathscr{C}=\lambda\partial f(\mathbf{x}_0)$. Combining this with Fact 2.2, we obtain,

$$\hat{f}_p(\mathbf{w}_{low})=\max_{\mathbf{s}\in\mathscr{C}}\langle\mathbf{w}_{low},\mathbf{s}\rangle=\frac{\langle\Pi(\mathbf{h},\mathscr{C}),\text{Proj}(\mathbf{h},\mathscr{C})\rangle}{\sqrt{\|\mathbf{g}\|_2^2-\text{dist}(\mathbf{h},\mathscr{C})^2}}.$$

What remains is to show the right hand side concentrates around $\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}$ with the desired probability. Fix a constant $\varepsilon>0$. Consider the denominator. Using Lemma A.5, with probability $1-\exp(-\mathscr{O}(m))$,

$$\left|\frac{\sqrt{\|\mathbf{g}\|_2^2-\text{dist}(\mathbf{h},\mathscr{C})^2}}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}-1\right|<\varepsilon. \tag{3.135}$$

We now apply Lemma A.3 for $\mathbf{C}(\mathscr{C})$ where we choose $t=\frac{m}{\sqrt{\max\{m,n\}}}$ and use the fact that $m>\mathbf{D}(\mathscr{C})$. Then, with probability $1-\exp(-\mathscr{O}\left(\min\{m,\frac{m^2}{n}\}\right))$, we have,

$$|\text{corr}(\mathbf{h},\mathscr{C})-\mathbf{C}(\mathscr{C})|\leq\varepsilon m.$$

Combining this with (3.135) choosing $\varepsilon>0$, sufficiently small (according to ε_L), we find (3.134) with the desired probability. ■

The lemma above shows that, $\lambda\hat{f}_p(\mathbf{w}^*_{low})$ is around $\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}$ with high probability and we obtain the ℓ_2^2 formula by using $\hat{f}_p(\mathbf{w}^*_{low})$ as a proxy for $\lambda\hat{f}_p(\mathbf{w}^*_{\ell_2})$. Can we do further? Possibly yes. To show $\hat{f}_p(\mathbf{w}^*_{\ell_2})$ is indeed around $\hat{f}_p(\mathbf{w}^*_{low})$, we can consider the modified deviation problem $\mathscr{L}^*_{dev}(\mathbf{g},\mathbf{h})=\min_{\mathbf{w}\in S_{dev}}\mathscr{L}(\mathbf{w};\mathbf{g},\mathbf{h})$ where we modify the set S_{dev} to,

$$S_{dev}=\{\mathbf{w}\big|\big|\frac{\lambda\hat{f}_p(\mathbf{w})}{\sigma}-\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}\big|>\varepsilon_{dev}\sqrt{m}\}.$$

We may then repeat the same arguments, i.e., try to argue that the objective restricted to S_{dev} is strictly greater than what we get from the upper bound optimization $\hat{\mathscr{U}}(\mathbf{g},\mathbf{h})$. While this approach may be promising, we believe it is more challenging than our ℓ_2 norm analysis of $\|\mathbf{w}^*_{\ell_2}\|_2$ and it will not be topic of this chapter.

The next section shows that there exists a one-to-one (monotone) mapping of the region $\mathscr{R}_{ON}$ to the entire possible regime of penalty parameters of the ℓ_2^2-LASSO.

3.12.2 Properties of map(λ)

The following result shows that $\mathbf{P}(\lambda\mathscr{C}), \mathbf{D}(\lambda\mathscr{C}), \mathbf{C}(\lambda\mathscr{C})$ (see (3.42)) are Lipschitz continuous and will be useful for the consequent discussion. The proof can be found in Appendix A.1.

Lemma 3.30 *Let $\mathscr{C}$ be a compact and convex set. Given scalar function $g(x)$, define the local Lipschitz constant to be $L_g(x) = \limsup_{x' \to x} \left| \frac{g(x')-g(x)}{x'-x} \right|$. Let $\max_{\mathbf{s}\in\mathscr{C}} \|\mathbf{s}\|_2 = R$. Then, viewing $\mathbf{P}(\lambda\mathscr{C}), \mathbf{D}(\lambda\mathscr{C}), \mathbf{C}(\lambda\mathscr{C})$ as functions of λ, for $\lambda \geq 0$, we have,*

$$\max\{L_{\mathbf{P}}(\lambda), L_{\mathbf{D}}(\lambda), L_{\mathbf{C}}(\lambda)\} \leq 2R(\sqrt{n} + \lambda R).$$

The following proposition is restatement of Theorem 3.3. Recall the definition of $\mathscr{R}_{\text{ON}}$ from Definition 3.4.

Proposition 3.4 *Assume $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{best})$. Recall that $\mathscr{R}_{ON} = (\lambda_{crit}, \lambda_{\max})$. $calib(\lambda) = \frac{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))-\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}$ and $map(\lambda) = \lambda \cdot calib(\lambda)$ have the following properties over $\{\lambda_{crit}\} \cup \mathscr{R}_{ON} \to \{0\} \cup \mathbb{R}^+$.*

- *$calib(\lambda)$ is a nonnegative, increasing and continuous function over $\{\lambda_{crit}\} \cup \mathscr{R}_{ON}$.*
- *$map(\lambda)$ is nonnegative, strictly increasing and continuous at all $\lambda \in \{\lambda_{crit}\} \cup \mathscr{R}_{ON}$.*
- *$map(\lambda_{crit}) = 0$. $\lim_{\lambda \to \lambda_{\max}} map(\lambda) = \infty$. Hence, $map(\lambda) : \{\lambda_{crit}\} \cup \mathscr{R}_{ON} \to \{0\} \cup \mathbb{R}^+$ is bijective.*

Proof: *Proof of the first statement:* Assume $\lambda \in \mathscr{R}_{\text{ON}}$, from Lemma 3.23, $m > \max\{\mathbf{D}(\lambda\partial f(\mathbf{x}_0)), \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{C}(\lambda\partial f(\mathbf{x}_0))\}$ and $\lambda > 0$. Hence, calib(λ) is strictly positive over $\lambda \in \mathscr{R}_{\text{ON}}$. Recall that,

$$\text{calib}(\lambda) = \frac{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) - \mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}} = \sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))} - \frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}.$$

Let $h > 0$. We will investigate the change in calib(λ) by considering calib$(\lambda + h) -$ calib(λ) as $h \to 0^+$. Since $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is differentiable, $\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ is differentiable as well and gives,

$$\frac{\partial\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}{\partial\lambda} = \frac{-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))'}{2\sqrt{m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}. \tag{3.136}$$

For the second term, consider the following,

$$\frac{\mathbf{C}_f(\mathbf{x}_0,\lambda+h)}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda+h)}}-\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}=h[E_1(\lambda,h)+E_2(\lambda,h)],$$

where,

$$E_1(\lambda,h)=\frac{1}{h}[\frac{\mathbf{C}_f(\mathbf{x}_0,\lambda+h)}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda+h)}}-\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda+h)}}],$$
$$E_2(\lambda,h)=\frac{1}{h}[\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda+h)}}-\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}].$$

As $h\to 0^+$, we have,

$$\lim_{h\to 0^+}E_2(\lambda,h)=\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\frac{\partial\frac{1}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}}{\partial\lambda}=\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\mathbf{D}(\lambda\partial f(\mathbf{x}_0))'}{2(m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0)))^{3/2}}\le 0, \tag{3.137}$$

since $\operatorname{sgn}(\mathbf{C}(\lambda\partial f(\mathbf{x}_0)))=-\operatorname{sgn}(\mathbf{D}(\lambda\partial f(\mathbf{x}_0))')$.

Fix arbitrary $\varepsilon_D>0$ and let $R=\sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)}\|\mathbf{s}\|_2$. Using continuity of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and Lemma 3.30, choose h sufficiently small to ensure,

$$|\frac{1}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}-\frac{1}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda+h)}}|<\varepsilon_D,\quad |\mathbf{C}_f(\mathbf{x}_0,\lambda+h)-\mathbf{C}(\lambda\partial f(\mathbf{x}_0))|<3R(\sqrt{n}+\lambda R)h.$$

We then have,

$$E_1(\lambda,h)\le\frac{\mathbf{C}_f(\mathbf{x}_0,\lambda+h)-\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{h}\frac{1}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}+3\varepsilon_D R(\sqrt{n}+\lambda R). \tag{3.138}$$

Denote $\frac{\mathbf{C}_f(\mathbf{x}_0,\lambda+h)-\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{h}$, $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda+h)-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{h}$ by $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{D}}$. Combining (3.137), (3.138) and (3.136), for sufficiently small h, we find,

$$\limsup_{h\to 0^+}\frac{\operatorname{calib}(\lambda+h)-\operatorname{calib}(\lambda)}{h}=\limsup_{h\to 0}[\frac{-\tilde{\mathbf{D}}}{2\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}-\frac{\tilde{\mathbf{C}}}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}$$
$$-\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\mathbf{D}(\lambda\partial f(\mathbf{x}_0))'}{2(m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0)))^{3/2}}+3\varepsilon_D R(\sqrt{n}+\lambda R)].$$

We can let ε_D go to 0 as $h\to 0^+$ and $-\tilde{\mathbf{D}}-2\tilde{\mathbf{C}}$ is always nonnegative as $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ is nondecreasing due to Lemma 3.20. Hence, the right hand side is nonnegative. Observe that the increase is strict for

$\lambda \neq \lambda_{\text{best}}$, as we have $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))\mathbf{D}(\lambda\partial f(\mathbf{x}_0))' > 0$ whenever $\lambda \neq \lambda_{\text{best}}$ due to the fact that $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))'$ (and $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$) is not 0. Since increase is strict around any neighborhood of λ_{best}, this also implies strict increase at $\lambda = \lambda_{\text{best}}$.

Consider the scenario $\lambda = \lambda_{\text{crit}}$. Since $\text{calib}(\lambda)$ is continuous for all $\lambda \in \{\lambda_{\text{crit}}\} \cup \mathscr{R}_{\text{ON}}$ (see next statement) and is strictly increasing at all $\lambda > \lambda_{\text{crit}}$, it is strictly increasing at $\lambda = \lambda_{\text{crit}}$ as well.

To see continuity of $\text{calib}(\lambda)$, observe that, for any $\lambda \in \mathscr{R}_{\text{ON}} \cup \{\lambda_{\text{crit}}\}$, $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) > 0$ and from Lemma 3.30, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)), \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ are continuous functions which ensures continuity of $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0)) - \mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ and $m - \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. Hence, $\text{calib}(\lambda)$ is continuous as well.

Proof of the second statement: Since $\text{calib}(\lambda)$ is strictly increasing on $\mathscr{R}_{\text{ON}}$, $\lambda \cdot \text{calib}(\lambda)$ is strictly increasing over $\mathscr{R}_{\text{ON}}$ as well. Increase at $\lambda = \lambda_{\text{crit}}$ follows from the fact that $\text{map}(\lambda_{\text{crit}}) = 0$ (see next statement). Since $\text{calib}(\lambda)$ is continuous, $\lambda \cdot \text{calib}(\lambda)$ is continuous as well.

Proof of the third statement: From Lemma 3.22, if $\text{calib}(\lambda_{\text{crit}}) > 0$, $\lambda_{\text{crit}} = 0$ hence $\text{map}(\lambda_{\text{crit}}) = 0$. If $\text{calib}(\lambda_{\text{crit}}) = 0$, then $\text{map}(\lambda_{\text{crit}}) = \lambda_{\text{crit}} \cdot \text{calib}(\lambda_{\text{crit}}) = 0$. In any case, $\text{map}(\lambda_{\text{crit}}) = 0$. Similarly, since $\lambda_{\max} > \lambda_{\text{best}}$, $\mathbf{C}_f(\mathbf{x}_0, \lambda_{\max}) < 0$ and as $\lambda \to \lambda_{\max}$ from left side, $\text{calib}(\lambda) \to \infty$. This ensures $\text{map}(\lambda) \to \infty$ as well. Since $\text{map}(\lambda)$ is continuous and strictly increasing and achieves the values 0 and ∞, it maps $\{\lambda_{\text{crit}}\} \cup \mathscr{R}_{\text{ON}}$ to $\{0\} \cup \mathbb{R}^+$ bijectively. ■

3.12.3 On the stability of ℓ_2^2-LASSO

As it has been discussed in Section 3.12.2 in detail, $\text{map}(\cdot)$ takes the interval $[\lambda_{\text{crit}}, \lambda_{\max})$ to $[0, \infty)$ and Theorem 3.2 gives tight stability guarantees for $\lambda \in \mathscr{R}_{\text{ON}}$. Consequently, one would expect ℓ_2^2-LASSO to be stable everywhere as long as the $[\lambda_{\text{crit}}, \lambda_{\max})$ interval exists. λ_{crit} and $\lambda_{\max}$ is well defined for the regime $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})$. Hence, we now expect ℓ_2^2-LASSO to be stable everywhere for $\tau > 0$. The next lemma shows that this is indeed the case under Lipschitzness assumption.

Lemma 3.31 *Consider the ℓ_2^2-LASSO problem (3.5). Assume $f(\cdot)$ is a convex and Lipschitz continuous function and $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$. Let $\mathbf{A}$ have independent standard normal entries and $\sigma\mathbf{v} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_m)$. Assume $(1-\varepsilon_L)m \geq \mathbf{D}(cone(\partial f(\mathbf{x}_0)))$ for a constant $\varepsilon_L > 0$ and m is sufficiently large. Then, there exists a number $C > 0$ independent of σ, such that, with probability $1 - \exp(-\mathscr{O}(m))$,*

$$\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|_2^2}{\sigma^2} \leq C. \tag{3.139}$$

Remark: We are not claiming anything about C except the fact that it is independent of σ. Better results can be given, however, our intention is solely showing that the estimation error is proportional to the noise variance. *Proof:* Consider the widening of the tangent cone defined as,

$$\mathscr{T}_f(\mathbf{x}_0, \varepsilon_0) = \mathrm{Cl}(\{\alpha \cdot \mathbf{w} \big| f(\mathbf{x}_0 + \mathbf{w}) \leq f(\mathbf{x}_0) + \varepsilon_0 \|\mathbf{w}\|_2,\ \alpha \geq 0\}).$$

Appendix A.8 investigates basic properties of this set. In particular, we will make use of Lemma A.14. We can choose sufficiently small numbers $\varepsilon_0, \varepsilon_1 > 0$ (independent of σ) such that,

$$\min_{\mathbf{w} \in \mathscr{T}_f(\mathbf{x}_0, \varepsilon_0), \|\mathbf{w}\|_2 = 1} \|\mathbf{A}\mathbf{w}\|_2 \geq \varepsilon_1, \tag{3.140}$$

with probability $1 - \exp(-\mathscr{O}(m))$ as $\sqrt{m-1} - \sqrt{\mathbf{D}(\mathrm{cone}(\partial f(\mathbf{x}_0)))} \gtrsim (1 - \sqrt{1-\varepsilon_L})\sqrt{m}$. Furthermore, we will make use of the following fact that $\|\mathbf{z}\|_2 \leq 2\sigma\sqrt{m}$ with probability $1 - \exp(-\mathscr{O}(m))$, where we let $\mathbf{z} = \sigma\mathbf{v}$ (see Lemma A.2).

Assuming these hold, we will show the existence of $C > 0$ satisfying (3.139). Define the perturbation function $f_p(\mathbf{w}) = f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0)$. Denote the error vector by $\mathbf{w}^*_{\ell_2^2} = \mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0$. Then, using the optimality of $\mathbf{x}^*_{\ell_2^2}$ we have,

$$\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}^*_{\ell_2^2}\|_2^2 + \sigma\tau f(\mathbf{x}^*_{\ell_2^2}) = \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}) \leq \frac{1}{2}\|\mathbf{z}\|_2^2.$$

On the other hand, expanding the terms,

$$\frac{1}{2}\|\mathbf{z}\|_2^2 \geq \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}) \geq \frac{1}{2}\|\mathbf{z}\|_2^2 - \|\mathbf{z}\|_2\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2 + \frac{1}{2}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}).$$

Using $\|\mathbf{z}\|_2 \leq 2\sigma\sqrt{m}$, this implies,

$$2\sigma\sqrt{m}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2 \geq \|\mathbf{z}\|_2\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2 \geq \frac{1}{2}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}). \tag{3.141}$$

Normalizing by σ,

$$2\sqrt{m}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2 \geq \frac{1}{2\sigma}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 + \tau f_p(\mathbf{w}^*_{\ell_2^2}).$$

The rest of the proof will be split into two cases.

Case 1: Let L be the Lipschitz constant of $f(\cdot)$. If $\mathbf{w}^*_{\ell_2^2} \in \mathscr{T}_f(\mathbf{x}_0, \varepsilon_0)$, using (3.140),

$$2\sqrt{m}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2 \geq \frac{1}{2\sigma}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 - \tau L\|\mathbf{w}^*_{\ell_2^2}\|_2 \geq \frac{1}{2\sigma}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2^2 - \frac{\tau L}{\varepsilon_1}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2.$$

Further simplifying, we find, $2\sigma(2\sqrt{m} + \frac{\tau L}{\varepsilon_1}) \geq \|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|_2 \geq \varepsilon_1\|\mathbf{w}^*_{\ell_2^2}\|_2$. Hence, indeed, $\frac{\|\mathbf{w}^*_{\ell_2^2}\|_2}{\sigma}$ is upper bound by $\frac{4\sqrt{m}}{\varepsilon_1} + \frac{2\tau L}{\varepsilon_1^2}$.

Case 2: Assume $\mathbf{w}^*_{\ell_2^2} \notin \mathscr{T}_f(\mathbf{x}_0, \varepsilon_0)$. Then $f_p(\mathbf{w}^*_{\ell_2^2}) \geq \varepsilon_0\|\mathbf{w}^*_{\ell_2^2}\|_2$. Using this and letting $\hat{\mathbf{w}} = \frac{\mathbf{w}^*_{\ell_2^2}}{\sigma}$, we can rewrite (3.141) without σ as,

$$\frac{1}{2}\|\mathbf{A}\hat{\mathbf{w}}\|_2^2 - 2\sqrt{m}\|\mathbf{A}\hat{\mathbf{w}}\|_2 + 2m + (\tau\varepsilon_0\|\hat{\mathbf{w}}\|_2 - 2m) \leq 0.$$

Finally, observing $\frac{1}{2}\|\mathbf{A}\hat{\mathbf{w}}\|_2^2 - 2\sqrt{m}\|\mathbf{A}\hat{\mathbf{w}}\|_2 + 2m = \frac{1}{2}(\|\mathbf{A}\hat{\mathbf{w}}\|_2 - 2\sqrt{m})^2$, we find,

$$\tau\varepsilon_0\|\hat{\mathbf{w}}\|_2 - 2m \leq 0 \implies \frac{\|\mathbf{w}^*_{\ell_2^2}\|_2}{\sigma} \leq \frac{2m}{\tau\varepsilon_0}.$$

■

3.13 Converse Results

Until now, we have stated the results assuming m is sufficiently large. In particular, we have assumed that $m \geq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ or $m \geq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. It is important to understand the behavior of the problem when m is small. Showing a converse result for $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ or $m < \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ will illustrate the tightness of our analysis. In this section, we focus our attention on the case where $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ and show that the NSE approaches infinity as $\sigma \to 0$. As it has been discussed previously, $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ is the compressed sensing threshold which is the number of measurements required for the success of the noiseless problem (1.4):

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}_0. \tag{3.142}$$

For our analysis, we use Proposition 3.5 below, which is a slight modification of Theorem 1 in [4].

Proposition 3.5 ([4]) *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have independent standard normal entries. Let $\mathbf{y} = \mathbf{A}\mathbf{x}_0$ and assume $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$. Further, for some $t > 0$, assume $m \leq \mathbf{D}(cone(\partial f(\mathbf{x}_0))) - t\sqrt{n}$. Then, $\mathbf{x}_0$ is not*

a minimizer of (3.142) *with probability at least* $1-4\exp(-\frac{t^2}{4})$.

Proposition 3.5 leads to the following useful Corollary.

Corollary 3.7 *Consider the same setting as in Proposition 3.5 and denote* $\mathbf{x}^*$ *the minimizer of* (3.142). *For a given* $t>0$, *there exists an* $\varepsilon>0$ *such that, with probability* $1-8\exp(-\frac{t^2}{4})$, *we have,*

$$f(\mathbf{x}^*) \leq f(\mathbf{x}_0) - \varepsilon$$

Proof: Define the random variable $\chi = f(\mathbf{x}^*) - f(\mathbf{x}_0)$. χ is random since $\mathbf{A}$ is random. Define the events $E = \{\chi < 0\}$ and $E_n = \{\chi \leq -\frac{1}{n}\}$ for positive integers n. From Proposition 3.5, $\mathbb{P}(E) \geq 1 - 4\exp(-\frac{t^2}{4})$. Also, observe that,

$$E = \bigcup_{i=1}^{\infty} E_i \quad \text{and} \quad E_n = \bigcup_{i=1}^{n} E_i,$$

Since E_n is an increasing sequence of events, by continuity property of probability, we have $\mathbb{P}(E) = \lim_{n\to\infty}\mathbb{P}(E_n)$. Thus, we can pick n_0 such that, $\mathbb{P}(E_{n_0}) > 1-8\exp(-\frac{t^2}{4})$. Let $\varepsilon = n_0^{-1}$, to conclude the proof. ■

The results discussed in this section, hold under the following assumption.

Assumption 1 *Let* $m_{lack} := \mathbf{D}(cone(\partial f(\mathbf{x}_0))) - m > 0$. $\mathbf{x}_0$ *is not a minimizer of the convex function* $f(\cdot)$. *For some* $L>0$, $f(\cdot):\mathbb{R}^n \to \mathbb{R}$ *is an L-Lipschitz function, i.e., for all* $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, $|f(\mathbf{x})-f(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|_2$.

3.13.1 Converse Result for C-LASSO

Recall the C-LASSO problem (3.3):

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}_0 + \sigma\mathbf{v} - \mathbf{A}\mathbf{x}\|_2 \quad \text{subject to} \quad f(\mathbf{x}) \leq f(\mathbf{x}_0). \tag{3.143}$$

(3.143) has multiple minimizers, in particular, if $\mathbf{x}^*$ is a minimizer, so is $\mathbf{x}^* + \mathbf{v}$ for any $\mathbf{v} \in \mathcal{N}(\mathbf{A})$. We will argue that when m is small, there exists a feasible minimizer which is far away from $\mathbf{x}_0$. The following theorem is a rigorous statement of this idea.

Theorem 3.8 *Suppose Assumption 1 holds and let* $\mathbf{A},\mathbf{v}$ *have independent standard normal entries. For any given constant* $C_{max}>0$, *there exists* $\sigma_0>0$ *such that, whenever* $\sigma \leq \sigma_0$, *with probability* $1-8\exp(-\frac{m_{lack}^2}{4n})$,

over the generation of $\mathbf{A}, \mathbf{v}$*, there exists a minimizer of* (3.143)*,* $\mathbf{x}_c^*$*, such that,*

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \geq C_{max} \tag{3.144}$$

Proof: From Corollary 3.7, with probability $1 - 8\exp(-\frac{m_{lack}^2}{4n})$, there exists $\varepsilon > 0$ and $\mathbf{x}'$ satisfying $f(\mathbf{x}') \leq f(\mathbf{x}_0) - \varepsilon$ and $\mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x}_0$.Denote $\mathbf{w}' = \mathbf{x}' - \mathbf{x}_0$ and pick a minimizer of (3.143) namely, $\mathbf{x}_0 + \mathbf{w}^*$. Now, let $\mathbf{w}_2^* = \mathbf{w}^* + \mathbf{w}'$. Observe that $\|\sigma\mathbf{v} - \mathbf{A}\mathbf{w}^*\|_2 = \|\sigma\mathbf{v} - \mathbf{A}\mathbf{w}_2^*\|_2$. Hence, $\mathbf{w}_2^* + \mathbf{x}_0$ is a minimizer for C-LASSO if $f(\mathbf{x}_0 + \mathbf{w}_2^*) \leq f(\mathbf{x}_0)$. But,

$$f(\mathbf{x}_0 + \mathbf{w}_2^*) = f(\mathbf{x}' + \mathbf{w}^*) \leq f(\mathbf{x}') + L\|\mathbf{w}^*\|_2,$$

Hence, if $\|\mathbf{w}^*\|_2 \leq \frac{f(\mathbf{x}_0) - f(\mathbf{x}')}{L}$, $\mathbf{w}_2^* + \mathbf{x}_0$ is a minimizer. Let $C_w = \min\{\frac{f(\mathbf{x}_0) - f(\mathbf{x}')}{L}, \frac{1}{2}\|\mathbf{w}'\|_2\}$ and consider,

$$\mathbf{w}_3^* = \begin{cases} \mathbf{w}^* & \text{if} \quad \|\mathbf{w}^*\|_2 \geq C_w, \\ \mathbf{w}_2^* & \text{otherwise.} \end{cases}$$

From the discussion above, $\mathbf{x}_0 + \mathbf{w}_3^*$ is guaranteed to be feasible and minimizer. Now, since $f(\mathbf{x}') \leq f(\mathbf{x}_0) - \varepsilon$ and $f(\cdot)$ is Lipschitz, we have that $\|\mathbf{w}'\|_2 \geq \frac{\varepsilon}{L}$. Consequently, if $\|\mathbf{w}^*\|_2 \geq C_w$, then, we have, $\frac{\|\mathbf{w}_3^*\|_2}{\sigma} \geq \frac{\varepsilon}{2L\sigma}$. Otherwise, $\|\mathbf{w}^*\|_2 \leq \frac{\|\mathbf{w}'\|_2}{2}$, and so,

$$\frac{\|\mathbf{w}_3^*\|_2}{\sigma} = \frac{\|\mathbf{w}_2^*\|_2}{\sigma} \geq \frac{|\|\mathbf{w}'\|_2 - \|\mathbf{w}^*\|_2|}{\sigma} \geq \frac{\|\mathbf{w}'\|_2}{2\sigma} \geq \frac{\varepsilon}{2L\sigma}.$$

In any case, we find that, $\frac{\|\mathbf{w}_3^*\|_2}{\sigma}$ is lower bounded by $\frac{\varepsilon}{2L\sigma}$ with the desired probability. To conclude with (3.144), we can choose σ_0 sufficiently small to ensure $\frac{\varepsilon^2}{4L^2\sigma_0^2} \geq C_{max}$.

■

3.13.2 Converse Results for ℓ_2-LASSO and ℓ_2^2-LASSO

This section follows an argument of similar flavor. We should emphasize that the estimation guarantee provided in Theorem 3.2 was for $m \geq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. However, hereby, the converse guarantee we give is slightly looser, namely, $m \leq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ where $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \leq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ by definition. This is mostly because of the nature of our proof which uses Proposition 3.5 and we believe it is possible to get a converse result for $m \leq \mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ via Proposition 2.7. We leave this to future work. Recall ℓ_2-LASSO

in (3.4):

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}_0 + \sigma\mathbf{v} - \mathbf{A}\mathbf{x}\|_2 + \lambda f(\mathbf{x}) \tag{3.145}$$

The following theorem is a restatement of Theorem 3.4 and summarizes our result on the ℓ_2-LASSO when m is small.

Theorem 3.9 *Suppose Assumption 1 holds and let* $\mathbf{A}, \mathbf{v}$ *have independent standard normal entries. For any given constant* $C_{max} > 0$*, there exists* $\sigma_0 > 0$ *such that, whenever* $\sigma \leq \sigma_0$*, with probability* $1 - 8\exp(-\frac{m_{lack}^2}{4n})$*, over the generation of* $\mathbf{A}, \mathbf{v}$*, the minimizer of* (3.145)*,* $\mathbf{x}^*_{\ell_2}$*, satisfies,*

$$\frac{\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|_2^2}{\sigma^2} \geq C_{max}. \tag{3.146}$$

Proof: From Corollary 3.7, with probability $1 - 8\exp(-\frac{m_{lack}^2}{4n})$, there exists $\varepsilon > 0$ and $\mathbf{x}'$ satisfying $f(\mathbf{x}') \leq f(\mathbf{x}_0) - \varepsilon$ and $\mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x}_0$. Denote $\mathbf{w}' = \mathbf{x}' - \mathbf{x}_0$. Let $\mathbf{w}^* + \mathbf{x}_0$ be a minimizer of (3.145) and let $\mathbf{w}_2^* = \mathbf{w}^* + \mathbf{w}'$. Clearly, $\|\mathbf{A}\mathbf{w}_2^* - \sigma\mathbf{v}\|_2 = \|\mathbf{A}\mathbf{w}^* - \sigma\mathbf{v}\|_2$. Hence, optimality of $\mathbf{w}^*$ implies $f(\mathbf{x}_0 + \mathbf{w}_2^*) \geq f(\mathbf{x}_0 + \mathbf{w}^*)$. Also, using the Lipschitzness of $f(\cdot)$,

$$f(\mathbf{x}_0 + \mathbf{w}_2^*) = f(\mathbf{x}' + \mathbf{w}^*) \leq f(\mathbf{x}') + L\|\mathbf{w}^*\|_2,$$

and

$$f(\mathbf{x}_0 + \mathbf{w}^*) \geq f(\mathbf{x}_0) - L\|\mathbf{w}^*\|_2.$$

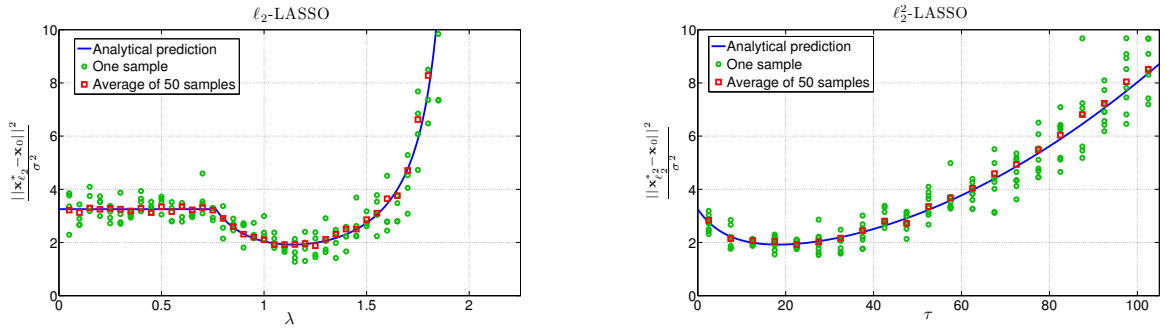
Combining those, we find,

$$f(\mathbf{x}') + L\|\mathbf{w}^*\|_2 \geq f(\mathbf{x}_0 + \mathbf{w}_2^*) \geq f(\mathbf{x}_0 + \mathbf{w}^*) \geq f(\mathbf{x}_0) - L\|\mathbf{w}^*\|_2,$$

which implies, $\|\mathbf{w}^*\|_2 \geq \frac{f(\mathbf{x}_0) - f(\mathbf{x}')}{2L} \geq \frac{\varepsilon}{2L}$, and gives the desired result (3.146) when $\sigma_0 \leq \frac{\varepsilon}{4L\sqrt{C_{max}}}$. ■

For the ℓ_2^2-LASSO result, let us rewrite (3.5) as,

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}_0 + \sigma\mathbf{v} - \mathbf{A}\mathbf{x}\|_2^2 + \sigma\tau f(\mathbf{x}) \tag{3.147}$$



(a) (b)

Figure 3.7: Sparse signal estimation with $n = 1500, m = 750, k = 150$. a) ℓ_1-penalized ℓ_2-LASSO NSE. b) ℓ_1-penalized ℓ_2^2-LASSO NSE. Observe that the minimum achievable NSE is same for both (around 1.92).

The next theorem shows that ℓ_2^2-LASSO does not recover $\mathbf{x}_0$ stably when $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. Its proof is identical to the proof of Theorem 3.9.

Theorem 3.10 *Suppose Assumption 1 holds and let* $\mathbf{A}, \mathbf{v}$ *have independent standard normal entries. For any given constant* $C_{max} > 0$*, there exists* $\sigma_0 > 0$ *such that, whenever* $\sigma \leq \sigma_0$*, with probability* $1 - 8\exp(-\frac{m_{lack}^2}{4n})$*, over the generation of* $\mathbf{A}, \mathbf{v}$*, the minimizer of* (3.147)*,* $\mathbf{x}^*_{\ell_2^2}$*, satisfies,*

$$\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|_2^2}{\sigma^2} \geq C_{max}.$$

3.14 Numerical Results

Simulation results presented in this section support our analytical predictions. We consider two standard estimation problems, namely sparse signal estimation and low rank matrix recovery from linear observations.

3.14.1 Sparse Signal Estimation

First, consider the sparse signal recovery problem, where $\mathbf{x}_0$ is a k sparse vector in $\mathbb{R}^n$ and $f(\cdot)$ is the ℓ_1 norm. We wish to verify our predictions in the small noise regime.

We fix $n = 1500$, $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$. Observe that, these particular choice of ratios has also been used in the Figures 3.3 and 3.4. $\mathbf{x}_0 \in \mathbb{R}^n$ is generated to be k sparse with standard normal nonzero entries and then normalized to satisfy $\|\mathbf{x}_0\| = 1$. To investigate the small σ regime, the noise variance is set to be $\sigma^2 = 10^{-5}$. We observe $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}(0, \sigma \mathbf{I}_m)$ and solve the ℓ_2-LASSO and the ℓ_2^2-LASSO problems with

ℓ_1 penalization. To obtain clearer results, each data point (red square markers) is obtained by averaging over 50 iterations of independently generated $\mathbf{A}, \mathbf{z}, \mathbf{x}_0$. The effect of averaging on the NSE is illustrated in Figure 3.7.

ℓ_2-LASSO: λ is varied from 0 to 2. The analytical predictions are calculated via the formulas given in Appendix A.7 for the regime $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$. We have investigated three properties.

- **NSE:** In Figure 3.7(a), we plot the simulation results with the small σ NSE formulas. Based on Theorem 3.2 and Section 3.9, over $\mathscr{R}_{\text{ON}}$, we plotted $\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ and over $\mathscr{R}_{\text{OFF}}$, we used $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{crit}})}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{crit}})}$ for analytical prediction. We observe that NSE formula indeed matches with simulations. On the left hand side, observe that NSE is flat and on the right hand side, it starts increasing as λ gets closer to $\lambda_{\max}$.
- **Normalized cost:** We plotted the cost of ℓ_2-LASSO normalized by σ in Figure 3.8(a). The exact function is $\frac{1}{\sigma}(\|\mathbf{y}-\mathbf{A}\mathbf{x}^*_{\ell_2}\| + \lambda(f(\mathbf{x}^*_{\ell_2}) - f(\mathbf{x}_0)))$. In $\mathscr{R}_{\text{ON}}$, this should be around $\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ due to Theorem 3.9. In $\mathscr{R}_{\text{OFF}}$, we expect cost to be linear in λ, in particular $\frac{\lambda}{\lambda_{\text{crit}}}\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{crit}})}$.
- **Normalized fit:** In Figure 3.8(b), we plotted $\frac{\|\mathbf{y}-\mathbf{A}\mathbf{x}^*_{\ell_2}\|}{\sigma}$, which is significant as it corresponds to the calibration function calib(λ) as described in Section 3.12. In $\mathscr{R}_{\text{ON}}$, we analytically expect this to be $\frac{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))-\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{\sqrt{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}}$. In $\mathscr{R}_{\text{OFF}}$, as discussed in Section 3.9.2, the problem behaves as (1.4) and we have $\mathbf{y} = \mathbf{A}\mathbf{x}_{\ell_2}$. Numerical results for small variance verify our expectations.

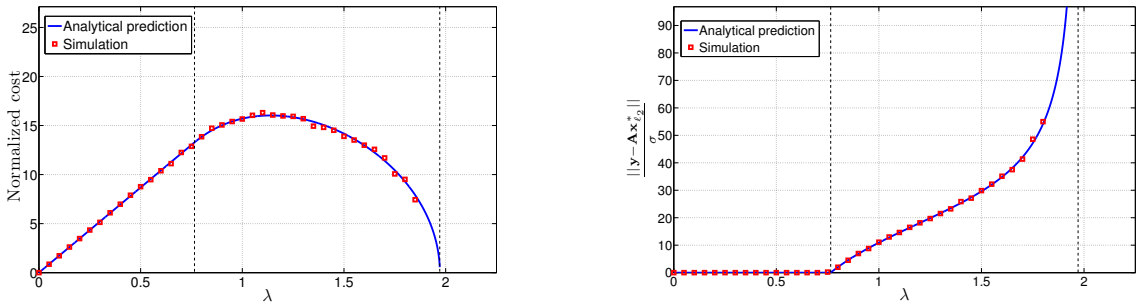
ℓ_2^2-LASSO: We consider the exact same setup and solve ℓ_2^2-LASSO. We vary τ from 0 to 100 and test the accuracy of Formula 1 in Figure 3.7(b). We find that, ℓ_2^2-LASSO is robust everywhere as expected and the minimum achievable NSE is same as ℓ_2-LASSO and around 1.92 as we estimate $\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{best}})$ to be around 330.

3.14.2 Low-Rank Matrix Estimation

For low rank estimation, we choose the nuclear norm $\|\cdot\|_\star$ as a surrogate for rank [178]. Nuclear norm is the sum of singular values of a matrix and basically takes the role of ℓ_1 minimization.

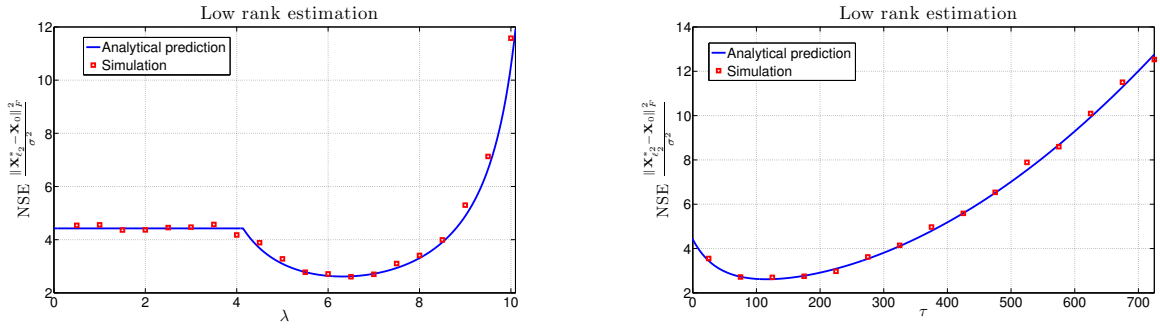
Since we will deal with matrices, we will use a slightly different notation and consider a low rank matrix $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$. Then, $\mathbf{x}_0 = \text{vec}(\mathbf{X}_0)$ will be the vector representation of $\mathbf{X}_0$, $n = d\times d$ and $\mathbf{A}$ will effectively be a Gaussian linear map $\mathbb{R}^{d\times d} \to \mathbb{R}^m$. Hence, for ℓ_2-LASSO, we solve,

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times d}} \|\mathbf{y}-\mathbf{A}\cdot\text{vec}(\mathbf{X})\| + \lambda\|\mathbf{X}\|_\star.$$



(a) (b)

Figure 3.8: ℓ_2-LASSO with $n = 1500$, $m = 750$, $k = 150$. a) Normalized cost of the optimization. b) How well the LASSO estimate fits the observations $\mathbf{y}$. This also corresponds to the calib(λ) function on $\mathscr{R}_{\text{ON}}$. In $\mathscr{R}_{\text{OFF}}$, ($\lambda \leq \lambda_{\text{crit}} \approx 0.76$) observe that $\mathbf{y} = \mathbf{A}\mathbf{x}^*_{\ell_2}$ indeed holds.



(a) (b)

Figure 3.9: $d = 45$, $m = 0.6d^2$, $r = 6$. We estimate $\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx 880$. a) ℓ_2-LASSO NSE as a function of the penalization parameter. b) ℓ_2^2-LASSO NSE as a function of the penalization parameter.

where $\mathbf{y} = \mathbf{A} \cdot \text{vec}(\mathbf{X}_0) + \mathbf{z}$.

Setup: We fixed $d = 45$, $\text{rank}(\mathbf{X}_0) = 6$ and $m = 0.6d^2 = 1215$. To generate $\mathbf{X}_0$, we picked i.i.d. standard normal matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times r}$ and set $\mathbf{X}_0 = \frac{\mathbf{U}\mathbf{V}^T}{\|\mathbf{U}\mathbf{V}^T\|_F}$ which ensures $\mathbf{X}_0$ is unit norm and rank r. We kept $\sigma^2 = 10^{-5}$. The results for ℓ_2 and ℓ_2^2-LASSO are provided in Figures 3.9(b) and 3.9(a) respectively. Each simulation point is obtained by averaging NSE's of 50 simulations over $\mathbf{A}, \mathbf{z}, \mathbf{X}_0$.

To find the analytical predictions, based on Appendix A.7, we estimated $\mathbf{D}(\lambda \partial f(\mathbf{x}_0)), \mathbf{C}(\lambda \partial f(\mathbf{x}_0))$ in the asymptotic regime: $n \to \infty$, $\frac{r}{d} = 0.133$ and $\frac{m}{n} = 0.6$. In particular, we estimate $\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx 880$ and best case NSE $\frac{\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})}{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})} \approx 2.63$. Even for such arguably small values of d and r, the simulation results are quite consistent with our analytical predictions.

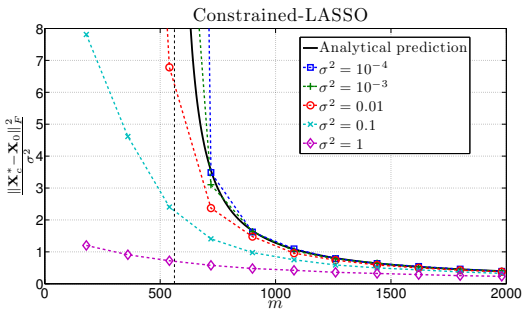


Figure 3.10: $\mathbf{X}_0$ is a 40×40 matrix with rank 4. As σ decreases, NSE increases. The vertical dashed lines marks the estimated $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ where we expect a transition in stability.

3.14.3 C-LASSO with varying σ

Consider the low rank estimation problem as in Section 3.14.2, but use the C-LASSO as an estimator:

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times d}} \|\mathbf{y} - \mathbf{A}\cdot\text{vec}(\mathbf{X})\| \quad \text{subject to} \quad \|\mathbf{X}\|_\star \leq \|\mathbf{X}_0\|_\star.$$

This time, we generate $\mathbf{A}$ with i.i.d. Bernoulli entries where each entry is either 1 or -1, with equal probability. The noise vector$\mathbf{z}$, the signal of interest $\mathbf{X}_0$ and the simulation points are generated in the same way as in Section 3.14.2. Here, we used $d = 40, r = 4$ and varied m from 0 to 2000 and σ^2 from 1 to 10^{-4}. The resulting curve is given in Figure 3.10. We observe that as the noise variance increases, the NSE decreases. The worst case NSE is achieved as $\sigma \to 0$, as Theorem 3.1 predicts. Our formula for the small σ regime $\frac{\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}{m-\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))}$ indeed provides a good estimate of NSE for $\sigma^2 = 10^{-4}$ and upper bounds the remaining ones. In particular, we estimate $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ to be around 560. Based on Theorems 3.4 and 3.1, as m moves from $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$ to $m > \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$, we expect a change in robustness. Observe that, for larger noise variances (such as $\sigma^2 = 1$) this change is not that apparent and the NSE is still relatively small. For $\sigma^2 \leq 10^{-2}$, the NSE becomes noticeably high for the regime $m < \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$.

3.15 Future Directions

This chapter sets up the fundamentals for a number of possible extensions. We enlist here some of those promising directions to be explored in future work.

- **ℓ_2^2-LASSO formula:** While Section 3.12 provides justification behind Formula 1, a rigorous proof is arguably the most important point missing in this chapter. Such a proof would close the gap in this

chapter and will extend results of [13, 14] to arbitrary convex functions.

- **Error formulas for arbitrary σ:** Another issue that hasn't been fully explored in this chapter is the regime where σ is not small. For C-LASSO, we have shown that the NSE for arbitrary values of σ is upper bounded by the NSE at $\sigma \to 0$. For ℓ_2-LASSO, our results in Section 3.10 are slightly weaker than the $\sigma \to 0$ bound given by Theorem 3.2; however, the empirical observations suggest that our prediction (that $\sigma \to 0$ is the worst case) is correct both for the ℓ_2 and ℓ_2^2-LASSO. Proving that this is the case is one open issue. What might be even more interesting, is computing exact error formulae for the arbitrary σ regime. As we have discussed previously, we expect such formulae to not only depend on the subdifferential of the function.
- **A better understanding:** Part of our discussion consists of repeated applications of Proposition 2.7 to the lasso problem and its dual to tightly sandwich the cost. We believe a more concise treatment to the lasso objective may be possible by carrying the duality arguments into Proposition 2.7. Towards this direction, our recent work shows that, under additional compactness and convexity assumptions, the comparison inequality of Proposition 2.7 can be shown to be tight [199]. This can help us obtain the upper and lower bounds on the objective in a single application. For the sake of completeness, we provide the result of [199].

Theorem 3.11 *Let $\mathbf{G} \in \mathbb{R}^{m \times n}$, $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0,1)$ entries that are independent of each other. Also, let $\Phi_1 \subset \mathbb{R}^n$, $\Phi_2 \subset \mathbb{R}^m$ be nonempty convex and compact sets and $\psi(\cdot,\cdot)$ be a continuous and convex-concave function on $\Phi_1 \times \Phi_2$. Finally, define*

$$\mathscr{G}(\mathbf{G}) := \min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{y}\in\Phi_2} \mathbf{y}^T\mathbf{G}\mathbf{x} + \psi(\mathbf{x},\mathbf{y}),$$
$$\mathscr{L}(\mathbf{g},\mathbf{h}) := \min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{y}\in\Phi_2} \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|_2\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y}).$$

Then, for any $c_- \in \mathbb{R}$ and $c_+ \in \mathbb{R}$:

$$\mathbb{P}\left(\mathscr{G}(\mathbf{G}) < c_-\right) \leq 2\mathbb{P}\left(\mathscr{L}(\mathbf{g},\mathbf{h}) \leq c_-\right), \tag{3.148}$$
$$\mathbb{P}\left(\mathscr{G}(\mathbf{G}) > c_+\right) \leq 2\mathbb{P}\left(\mathscr{L}(\mathbf{g},\mathbf{h}) \geq c_+\right). \tag{3.149}$$

Observe that (3.148) is basically identical to Proposition 2.7, while (3.149) is new.

- **Extension to multiple structures:** Throughout this chapter, we have focused on the recovery of a single signal $\mathbf{x}_0$. In general, one may consider a scenario, where we observe mixtures of multiple structures. A classic example used to motivate such problems includes estimation of matrices that can be represented as sum of a low rank and a sparse component [36, 51, 144, 221]. Another example, which is closer to our framework, is when the measurements $\mathbf{A}\mathbf{x}_0$ experience not only additive i.i.d. noise $\mathbf{z}$, but also sparse corruptions $\mathbf{s}_0$ [101, 133]. In this setup, we observe $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{s}_0 + \mathbf{z}$ and we wish to estimate $\mathbf{x}_0$ from $\mathbf{y}$. The authors in [101, 145] provide sharp recovery guarantees for the noiseless problem, but do not address the precise noise analysis. We believe, our framework can be extended to the exact noise analysis of the following constrained problem:

$$\min_{\mathbf{x},\mathbf{s}} \|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{s}\|_2 \quad \text{subject to} \quad g(\mathbf{s}) \leq g(\mathbf{s}_0) \ \text{ and } \ f(\mathbf{x}) \leq f(\mathbf{x}_0).$$

  where $g(\cdot)$ is typically the ℓ_1 norm.
- **Application specific results:** In this chapter, we focused on a generic signal-function pair $\mathbf{x}_0, f$ and stated our results in terms of the convex geometry of the problem. We also provided numerical experiments on NSE of sparse and low rank recovery and showed that, theory and simulations are consistent. On the other hand, it would be useful to derive case-specific guarantees other than NSE. For example, for sparse signals, we might be interested in the sparsity of the LASSO estimate, which has been considered by Bayati and Montanari [13, 14]. Similarly, in low rank matrix estimation, we might care about the rank and nuclear norm of the LASSO estimate. On the other hand, our generic results may be useful to obtain NSE results for a growing set of specific problems with little effort, [91, 144, 160, 168, 181, 221]. In particular, one can find an NSE upper bound to a LASSO problem as long as he has an upper bound to $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ or $\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$.
- **Mean-Squared-Error (MSE) Analysis:** In this chapter, we focused on the ℓ_2-norm square of the LASSO error and provided high probability guarantees. It is of interest to give guarantees in terms of mean-squared-error where we consider the expected NSE. Naturally, we expect our formulae to still hold true for the MSE, possibly requiring some more assumptions.

Chapter 4

Elementary equivalences in compressed sensing

This chapter consists of several results that provide useful and concise equivalences for the linear inverse problems. We will first illustrate that the sample complexity of i.i.d Bernoulli measurements can be related to that of Gaussian measurement ensemble, by establishing a relation between the two matrix ensembles. In the next section, our focus will be establishing a similarity between the signal structures. We will show that the strong recovery conditions for the sparse and low-rank approximation problems are inherently related, and hence, one can "translate" the RIP constants and the associated recovery guarantees from sparse vectors to low-rank matrices with no effort.

4.1 A comparison between the Bernoulli and Gaussian ensembles

Recall that, the Gaussian measurement ensemble has particular importance in low-dimensional representation problems

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{A}\mathbf{x}_0 = \mathbf{A}\mathbf{x}, \tag{4.1}$$

as one can carry out sharp and nonasymptotic analysis of the performance of BP when $\mathbf{A}$ has independent $\mathscr{N}(0,1)$ entries. Our interest in this section is to lay out a framework to obtain results for i.i.d nongaussian measurements by constructing a proper similarity to the Gaussian measurements. We will restrict our attention to symmetric Bernoulli (i.e. Rademacher) measurements, which are equally likely to be $+1$ and -1. However, the proposed framework can be trivially extended from Bernoulli to, first, other discrete distributions and then to continuous distributions with more effort. Focusing on Bernoulli measurements will make our results cleaner and arguably more elegant. Bernoulli measurement ensemble is interesting in its

own right as it is advantageous both from computation and storage points of view [176, 232].

We show that Bernoulli measurements can be used for linear inverse problems in a similar manner to Gaussian's by paying a price of constant multiplier in front of the sample complexity. This is along the lines of [206], which provide similar guarantees for the subgaussian ensemble up to unknown constants. We present a novel strategy that allows us to measure the similarity between two distributions, which yields explicit constants.

To give an initial intuition, we start with a basic comparison between a matrix with independent $\mathcal{N}(0,1)$ entries and one with symmetric Bernoulli entries.

Proposition 4.1 *Let $\mathcal{S}$ be a closed subset of the unit sphere in $\mathbb{R}^n$. Let $\mathbf{G} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ be matrices with independent $\mathcal{N}(0,1)$ and symmetric Bernoulli entries respectively. Suppose, for some $\varepsilon > 0$, we have*

$$(1-\varepsilon)m \leq \mathbb{E}\min_{\mathbf{v}\in\mathcal{S}} \|\mathbf{G}\mathbf{v}\|_2^2 \leq \mathbb{E}\max_{\mathbf{v}\in\mathcal{S}} \|\mathbf{G}\mathbf{v}\|_2^2 \leq (1+\varepsilon)m. \tag{4.2}$$

Then, we also have

$$(1-\frac{\pi}{2}\varepsilon)m \leq \mathbb{E}\min_{\mathbf{v}\in\mathcal{S}} \|\mathbf{B}\mathbf{v}\|_2^2 \leq \mathbb{E}\max_{\mathbf{v}\in\mathcal{S}} \|\mathbf{B}\mathbf{v}\|_2^2 \leq (1+\frac{\pi}{2}\varepsilon)m.$$

Observe that, for a fixed unit length vector $\mathbf{v}$, we have that

$$\mathbb{E}\,\|\mathbf{G}\mathbf{v}\|_2^2 = \mathbb{E}\,\|\mathbf{B}\mathbf{v}\|_2^2 = m.$$

On the other hand, when $\mathcal{S}$ is not a singleton, it is less nontrivial to estimate $\min_{\mathbf{v}\in\mathcal{S}} \|\mathbf{A}\mathbf{v}\|_2$ and $\max_{\mathbf{v}\in\mathcal{S}} \|\mathbf{A}\mathbf{v}\|_2$. From the discussion in Chapter 1, we know that, these quantities are important for the analysis of (4.1) and has been the subject of interest recently. A standard example is when we let $\mathcal{S}$ to be the set of at most k sparse (and normalized) vectors, i.e.

$$\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^n \big| \|\mathbf{v}\|_0 \leq k,\ \|\mathbf{v}\|_2 = 1\}.$$

In this case, the smallest possible ε in (4.2) effectively corresponds to the k-Restricted Isometry Constant δ_k of the Gaussian matrix. Hence, Proposition 4.1 relates the $\delta_k(\mathbf{G})$ and $\delta_k(\mathbf{B})$, namely, $\delta_k(\mathbf{B}) \leq \frac{\pi}{2}\delta_k(\mathbf{G})$. For the following discussion, probability density functions (p.d.f) will be denoted by lower case letters and the cumulative density functions (c.d.f) will be denoted by the corresponding upper case letters. Given a p.d.f $f(\cdot)$, mean(f) and var(f) will correspond to the mean and variance of the associated random variable. Also,

let us recall the definition of restricted singular value.

Definition 4.1 (Restricted Singular Value) *Given a closed cone $\mathscr{C}$ and a matrix $\mathbf{A}$, the restricted minimum and maximum singular values of $\mathbf{A}$ at $\mathscr{C}$ are defined as,*

$$\sigma_{\mathscr{C}}(\mathbf{A}) = \min_{\mathbf{v}\in\mathscr{C},\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2,\ \Sigma_{\mathscr{C}}(\mathbf{A}) = \max_{\mathbf{v}\in\mathscr{C},\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2.$$

We will now describe how to establish a similarity between symmetric Bernoulli and standard normal distribution.

4.1.1 Proportional Mean Decomposition

Given a piecewise continuous density function f_C and discrete distribution f_D, we propose the following partitioning of the continuous distribution in terms of the discrete one.

Definition 4.2 (Proportional mean decomposition (PMD)) *Let f_C be a zero-mean probability distribution and f_D be a zero-mean discrete distribution with alphabet size of K, given by,*

$$f_D(x) = \sum_{i=1}^{K} p_i \delta(x - a_i)$$

where $\sum_{i=1}^{K} p_i = 1$ and $\{a_i\}_{i=1}^{n}$'s are ordered increasingly and $\delta(\cdot)$ is the Dirac delta function. We say $\{f_i\}_{i=1}^{K}$ is a proportional mean decomposition of f_C with respect to f_D with the similarity constant c_S, if, there exists probability distributions $\{f_i\}_{i=1}^{K}$ satisfying,

$$f_C = \sum_{i=1}^{K} p_i f_i$$
$$mean(f_i) = c_S a_i$$

Additionally, let $\sigma_i = \sqrt{var(f_i)}$ and $\sigma_{\max} = \max_{1\le i\le K} \sigma_i$.

4.1.1.1 Examples

To provide a better intuition, we provide two examples on PMD when $f_C \sim \mathscr{N}(0,1)$.

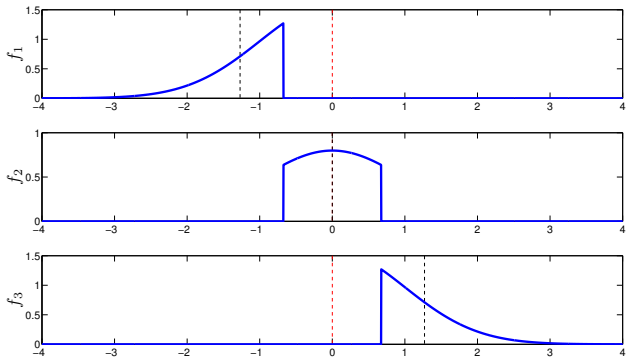


Figure 4.1: Dashed black lines correspond to mean(f_i).

- Suppose f_D is symmetric Bernoulli ± 1. Let

$$f_1(x) = \sqrt{\frac{2}{\pi}}\exp(-\frac{x^2}{2}) \quad \text{for} \quad x \geq 0,$$
$$f_2(x) = \sqrt{\frac{2}{\pi}}\exp(-\frac{x^2}{2}) \quad \text{for} \quad x < 0.$$

  Then, $c_S^2 = \frac{2}{\pi}$ and $\sigma_{\max}^2 = \sigma_1^2 = \sigma_2^2 = 1 - \frac{2}{\pi}$.

- Suppose f_D is the ternary distribution,

$$f_D = \frac{1}{4}\delta(x+\sqrt{2}) + \frac{1}{2}\delta(x) + \frac{1}{4}\delta(x-\sqrt{2}).$$

  Let Q be the tail function of $\mathcal{N}(0,1)$. Then, let

$$f_1(x) = \sqrt{\frac{2}{\pi}}\exp(-\frac{x^2}{2}) \quad \text{for} \quad x \geq Q^{-1}(1/4),$$
$$f_2(x) = \sqrt{\frac{2}{\pi}}\exp(-\frac{x^2}{2}) \quad \text{for} \quad |x| < Q^{-1}(1/4),$$
$$f_3(x) = f_1(-x) \quad \text{for} \quad x \leq -Q^{-1}(1/4).$$

  as described in Figure 4.1. In this case, $\sigma_{\max}^2 = \sigma_1^2 = \sigma_3^2 \approx 0.242$, $\sigma_2^2 \approx 0.143$ and $c_S^2 \approx 0.808$.

4.1.1.2 Properties of PMD

PMD satisfies the following properties.

Lemma 4.1 *Consider the setup in Definition 4.2.*

- *The set of achievable similarity constants c_S is convex and contains* 0.
- $\sum_{k=1}^{K} p_i \sigma_i^2 = 1 - c_S^2$.
- *Suppose $x_D \in \mathbb{R}$ is distributed with f_D (i.e. $x_D \sim f_D$). Define x_C conditioned on x_D as follows,*

$$x_C \sim f_i \quad \textit{iff} \quad x_D = a_i \quad 1 \leq i \leq K.$$

Then, $x_C \sim f_C$. Furthermore, $x_C - c_S x_D$ has variance $1 - c_S^2$ and conditioned on x_D, $x_C - c_S x_D$ is zero-mean.

Proof: **First statement:** $c_S = 0$ can be achieved by choosing $f_i = f_C$ for $1 \leq i \leq K$ as f_C is zero-mean. Let c_1, c_2 be two similarity constants with the corresponding p.d.f's $\{f_{1,i}\}_{i=1}^{K}, \{f_{2,i}\}_{i=1}^{K}$. One can achieve $c_S = \alpha c_1 + (1-\alpha)c_2$ for $0 \leq \alpha \leq 1$, by choosing,

$$f_{\alpha,i} = \alpha f_{1,i} + (1-\alpha) f_{2,i}$$

and using the linearity of expectation.

Second statement: For each i, we have that, $\text{var}(f_i) + \text{mean}(f_i)^2 = \int_{-\infty}^{\infty} x^2 f_i(x) dx$, and,

$$1 = \text{var}(f_C) = \int_{-\infty}^{\infty} x^2 \sum_{i=1}^{K} p_i f_i(x) dx = \sum_{i=1}^{K} p_i \int_{-\infty}^{\infty} x^2 f_i(x) dx$$

Recalling $\text{mean}(f_i) = c_S a_i$, and $\text{var}(f_D) = \sum_{i=1}^{K} p_i a_i^2 = 1$, we find,

$$1 = c_S^2 \text{var}(f_D) + \sum_{i=1}^{K} p_i \sigma_i^2 = c_S^2 + \sum_{i=1}^{K} p_i \sigma_i^2$$

Third statement: Focusing on the c.d.f of x_C,

$$\begin{aligned} \mathbb{P}(x_C \leq \alpha) &= \sum_{k=1}^{K} \mathbb{P}(x_C \leq \alpha \big| x_D = a_k) \mathbb{P}(x_D = a_k) \\ &= \sum_{k=1}^{K} p_k F_k(\alpha) = F_C(\alpha). \end{aligned}$$

Hence, p.d.f of x_C is indeed f_C. Similarly, conditioning over x_D,

$$\mathbb{E}[x_C - c_S x_D | x_D = a_i] = \mathbb{E}[x_C | x_D = a_i] - c_S a_i = 0.$$

Using the mean is equal to zero,

$$\text{var}(x_C - c_S x_D) = \mathbb{E}[(x_C - c_S x_D)^2] = \sum_{i=1}^{K} \mathbb{E}[(x_C - c_S x_D)^2 | x_D = a_i] p_i = \sum_{i=1}^{K} [\int_{\mathbb{R}} t^2 f_i(t) dt] p_i - \sum_{i=1}^{K} c_S^2 a_i^2 p_i = 1 - c_S^2.$$

■

4.1.1.3 From scalars to i.i.d matrices

Our next aim is to use PMD to obtain results on random matrices.

Definition 4.3 (Sensing matrices) *Consider Definition 4.2. Let* $\mathbf{D} \in \mathbb{R}^{m \times n}$ *be a matrix with i.i.d entries distributed as* f_D*. Let* $\mathbf{C}$ *be a matrix satisfying,*

$$\mathbf{C}_{i,j} \sim f_k \text{ if } \mathbf{D}_{i,j} = a_k, \quad \forall\, 1 \leq k \leq K,\ 1 \leq i \leq m,\ 1 \leq j \leq n.$$

Finally, define the residual matrix to be $\mathbf{R} := \mathbf{R}(f_C, f_D) = \mathbf{C} - c_S \mathbf{D}$.

The following proposition provides an initial motivation on PMD.

Proposition 4.2 (Bound in Expectation) *Suppose* $\mathbf{D}$ *and* $\mathbf{C}$ *are as defined above. Let* $\sigma_{\min} = \min_{1 \leq k \leq K} \sigma_i$*. Then, for any closed cone* $\mathscr{C} \in \mathbb{R}^n$*,*

$$\mathbb{E}[\sigma_{\mathscr{C}}(\mathbf{D})^2] \geq \frac{\mathbb{E}[\sigma_{\mathscr{C}}(\mathbf{C})^2] - \sigma_{\max}^2 m}{c_S^2},\ \mathbb{E}[\Sigma_{\mathscr{C}}(\mathbf{D})^2] \leq \frac{\mathbb{E}[\Sigma_{\mathscr{C}}(\mathbf{C})^2] - \sigma_{\min}^2 m}{c_S^2}.$$

Proof: To prove the first statement, given $\mathbf{D}$ and $\mathscr{C}$, let,

$$\hat{\mathbf{v}} = \arg \min_{\mathbf{v} \in \mathscr{C} \cap \mathscr{S}^{n-1}} \|\mathbf{D}\mathbf{v}\|_2.$$

where $\mathscr{S}^{n-1}$ is the unit ℓ_2 sphere. Conditioned on $\mathbf{D}$, $\hat{\mathbf{v}}$ is fixed and $\mathbf{C} - c_S \mathbf{D}$ has independent, zero-mean entries. Hence,

$$\mathbb{E}_{\mathbf{C}|\mathbf{D}}[\|\mathbf{C}\hat{\mathbf{v}}\|_2^2] = \|c_S \mathbf{D}\hat{\mathbf{v}}\|_2^2 + \mathbb{E}_{\mathbf{C}|\mathbf{D}}[\|(\mathbf{C} - c_S \mathbf{D})\hat{\mathbf{v}}\|_2^2] \tag{4.3}$$

Since $\mathbf{v}$ has unit length and the entries of $\mathbf{C}-c_S\mathbf{D}$ has variance at most $\sigma_{\max}^2$, $\mathbb{E}_{\mathbf{C}|\mathbf{D}}[\|(\mathbf{C}-c_S\mathbf{D})\hat{\mathbf{v}}\|_2^2] \leq \sigma_{\max}^2 m$. Hence, taking the expectation over $\mathbf{D}$, we find,

$$\mathbb{E}[\sigma_{\mathscr{C}}^2(\mathbf{C})] \leq \mathbb{E}[\|\mathbf{C}\hat{\mathbf{v}}\|_2^2] \leq \mathbb{E}[\|c_S\mathbf{D}\hat{\mathbf{v}}\|_2^2] + \sigma_{\max}^2 m = c_S^2\mathbb{E}[\sigma_{\mathscr{C}}(\mathbf{D})^2] + \sigma_{\max}^2 m \tag{4.4}$$

To prove the second statement, let $\hat{\mathbf{v}} = \arg\max_{\mathbf{v}\in\mathscr{C}\cap\mathscr{S}^{n-1}} \|\mathbf{D}\mathbf{v}\|_2$ and observe that $\mathbb{E}_{\mathbf{C}|\mathbf{D}}[\|(\mathbf{C}-c_S\mathbf{D})\hat{\mathbf{v}}\|_2^2] \geq \sigma_{\min}^2 m$. Instead of (4.4), we use,

$$\mathbb{E}[\Sigma_{\mathscr{C}}^2(\mathbf{C})] \geq \mathbb{E}[\|\mathbf{C}\hat{\mathbf{v}}\|_2^2] \geq \mathbb{E}[\|c_S\mathbf{D}\hat{\mathbf{v}}\|_2^2] + \sigma_{\min}^2 m = c_S^2\mathbb{E}[\Sigma_{\mathscr{C}}(\mathbf{D})^2] + \sigma_{\min}^2 m.$$

■

4.1.1.4 Proof of Proposition 4.1

We are in a position to prove Proposition 4.1; which is essentially a corollary of Proposition 4.2. For $f_C \sim \mathscr{N}(0,1)$ and f_D is symmetric Bernoulli, we have $c_S^2 = \frac{2}{\pi}$, $\sigma_{\max}^2 = \sigma_{\min}^2 = 1-\frac{2}{\pi}$. Hence, if $\mathbb{E}[\sigma_{\mathscr{C}}^2(\mathbf{C})] \geq (1-\varepsilon)m$,

$$\mathbb{E}[\sigma_{\mathscr{C}}(\mathbf{D})^2] \geq \frac{(1-\varepsilon)m - (1-\frac{2}{\pi})m}{\frac{2}{\pi}} = (1-\frac{\pi}{2}\varepsilon)m.$$

Similarly, using $\mathbb{E}[\Sigma_{\mathscr{C}}^2(\mathbf{C})] \leq (1+\varepsilon)m$,

$$\mathbb{E}[\Sigma_{\mathscr{C}}(\mathbf{D})^2] \leq \frac{(1+\varepsilon)m - (1-\frac{2}{\pi})m}{\frac{2}{\pi}} = (1+\frac{\pi}{2}\varepsilon)m.$$

Proposition 4.2 considers the crude bounds involving $\sigma_{\min}^2$ and $\sigma_{\max}^2$ in the statements. In fact, one can always replace them with $1-c_S^2$ by moving from a deterministic statement to a probabilistic one. This can be done by arguing that, with high probability (for large m), each a_i occurs at most $(1+\varepsilon')mp_i$ times at each column of $\mathbf{D}$. For such $\mathbf{D}$'s, the expected energy of each column of $\mathbf{C}-c_S\mathbf{D}$ can be upper bounded by $(1+\varepsilon')m(1-c_S^2)$. One can similarly obtain lower bounds on the column sizes with $(1-\varepsilon')$ multiplicity and then repeat the argument in Proposition 4.2 to get results that hold with high probability over $\mathbf{D}$.

4.1.2 On the sample complexity of Bernoulli ensemble

We will obtain results for Bernoulli matrices by using Gordon's Comparison Theorem. Recall that, for success of (4.1), we need, $\sigma_{\mathscr{T}_f(\mathbf{x}_0)}(\mathbf{A}) > 0$ where $\mathscr{T}_f(\mathbf{x}_0)$ is the tangent cone of f at $\mathbf{x}_0$. Recall that $\delta(\mathscr{C})$ is the phase transition point for (4.1) when the measurements are Gaussian. The next result, obtains a probabilistic success result for symmetric Bernoulli's in terms of $\delta(\mathscr{C})$.

Theorem 4.1 (Sample complexity bound for Bernoulli) *Let $f:\mathbb{R}^n \to \mathbb{R}$ be a convex and continuous function. Suppose $\mathbf{B} \in \mathbb{R}^{m\times n}$ is a matrix with independent entries that are ± 1 equally likely. Fix a tolerance level ρ. Then, $\mathbf{x}_0$ is the unique solution of (4.1), with probability $1-\exp(-\frac{c^2}{2})$, whenever*

$$\sqrt{m} \geq 2.6(\omega(\mathscr{T}_f(\mathbf{x}_0)\cap \mathscr{B}^{n-1}) + c + 4).$$

Proof: Let $\mathscr{C} = \mathscr{T}_f(\mathbf{x}_0)$. Define $\mathbf{G} \in \mathbb{R}^{m\times n}$ with i.i.d $f_G \sim \mathscr{N}(0,1)$ entries based on $\mathbf{B}$ as in Definition 4.3. From Proposition 4.2 (in particular (4.3)), we know that

$$\mathbb{E}[\sigma^2_{\mathscr{C}}(\mathbf{G})|\mathbf{B}] = \sigma^2_{\mathscr{C}}(\mathbf{B}) + (1-\frac{2}{\pi})m. \tag{4.5}$$

In order to estimate the left hand side, we will use a probabilistic argument. Combining 1-Lipschitzness of RSV and Proposition 1.3, for $t < \gamma_m - \omega(\mathscr{C}\cap\mathscr{B}^{n-1})$, with probability $1-\exp(-\frac{t^2}{2})$,

$$\sigma^2_{\mathscr{C}}(\mathbf{G}) \geq (\gamma_m - \omega(\mathscr{C}\cap\mathscr{B}^{n-1}) - t)^2.$$

Now, let let E be the event $\{\sigma_{\mathscr{C}}(\mathbf{G}) \leq \gamma_m - \omega(\mathscr{C}\cap\mathscr{B}^{n-1}) - t\}$ and given $\rho := \exp(-\frac{c^2}{2}) > 0$, $S \subset \{1,-1\}^{m\times n}$ be the set of $\mathbf{B}$ such that $\mathbb{P}(E\big|\mathbf{B}) \geq \rho^{-1}\exp(-\frac{t^2}{2})$. We have,

$$\begin{aligned}
\exp(-\frac{t^2}{2}) \geq \mathbb{P}(E) &= \sum_{\mathbf{A}\in\{1,-1\}^{m\times n}} \mathbb{P}(E\big|\mathbf{B}=\mathbf{A})\mathbb{P}(\mathbf{B}=\mathbf{A}) \\
&\geq \mathbb{P}(E\big|\mathbf{B}\in S)\mathbb{P}(\mathbf{B}\in S) \\
&\geq \rho^{-1}\exp(-\frac{t^2}{2})\mathbb{P}(\mathbf{B}\in S).
\end{aligned}$$

It follows that, $\mathbb{P}(\mathbf{B} \in S) \leq \rho$. Hence, with probability $1-\rho$ over $\mathbf{B}$,

$$\mathbb{P}(\sigma_{\mathscr{C}}(\mathbf{G}) \geq \gamma_m - \omega(\mathscr{C} \cap \mathscr{B}^{n-1}) - t \big| \mathbf{B}) \geq 1 - \rho^{-1} \exp(-\frac{t^2}{2}).$$

Hence, with the same probability,

$$\mathbb{E}[\sigma_{\mathscr{C}}(\mathbf{G})^2] \geq (\gamma_m - \omega(\mathscr{C} \cap \mathscr{B}^{n-1}) - t)^2 (1 - \rho^{-1} \exp(-\frac{t^2}{2})).$$

Choose $t = \sqrt{2 \log \rho^{-1}} + 3$. This ensures, $(1 - \rho^{-1} \exp(-\frac{t^2}{2})) \geq 0.98$. Using $\gamma_m \geq \sqrt{m} - 1$, to show $\sigma_{\mathscr{C}}(\mathbf{B}) > 0$, from (4.5), we wish to guarantee,

$$\sqrt{m} - (\omega(\mathscr{C} \cap \mathscr{B}^{n-1}) + t + 1) \geq \frac{1}{0.98} \sqrt{1 - \frac{2}{\pi}} \sqrt{m}.$$

Hence, we simply need $\sqrt{m} \geq 2.6(\omega(\mathscr{C} \cap \mathscr{B}^{n-1}) + t + 1)$. ■

4.1.3 Concluding remarks

We have introduced the "proportional mean decomposition" as a way to capture similarity of one distribution to another and discussed how it can be useful in compressed sensing, especially when the measurement matrix has i.i.d Bernoulli entries. While we are able to obtain small explicit constants in Proposition 4.1 and Theorem 4.1, our basic approach fails to capture the universality phenomenon, which is the common belief that, the sample complexity for i.i.d Bernoulli (and more generally i.i.d subgaussian) and i.i.d Gaussian ensembles are asymptotically equal. This remains as an important open question, which is partially answered by Montanari et al. in the case of ℓ_1 minimization [12].

4.2 An equivalence between the recovery conditions for sparse signals and low-rank matrices

The Restricted Isometry Property (RIP) was introduced by Candès and Tao in [32, 45] and has played a major role in proving recoverability of sparse signals from compressed measurements. The first recovery algorithm that was analyzed using RIP was ℓ_1 minimization in [32,45]. Since then, many algorithms including Reweighed ℓ_1 [159], GraDes [104] have been analyzed using RIP. Analogous to the vector case, RIP

has also been used in the analysis of algorithms for *low rank matrix recovery*, for example Nuclear Norm Minimization [178], Reweighted Trace Minimization [156], and SVP [150]. Other recovery conditions have also been proposed for recovery of both sparse vectors and low-rank matrices including the Null Space Property [163,233] and the Spherical Section Property [87,233] (also known as the 'almost Euclidean' property) for the nullspace. The first matrix RIP result was due to Recht et. al. [178] where it was shown that the RIP is sufficient for low rank recovery using nuclear norm minimization, and that it holds with high probability as long as number of measurements are sufficiently large. This analysis was improved in [39] to require a minimal order of measurements. Recently, [156] improved the RIP constants with a stronger analysis similar to [27].

In this section, we show that if a set of conditions are sufficient for the robust recovery of sparse vectors with sparsity at most k, then *"extension"* (defined later) of the same set of conditions are sufficient for the robust recovery of low rank matrices up to rank k. In particular, we show RIP for matrices implies extension of RIP for vectors hence one can easily translate the best known RIP conditions for vector recovery to low rank matrices. While the recovery analysis in [156] and [39] (Theorem 2.4) is complicated and lengthy, our results (see "Main Theorem") are easily derived due to the use of a key singular value inequality (Lemma 4.2). The best known bounds on the RIP constants δ_k and δ_{2k} for sparse vector recovery using ℓ_1 minimization are 0.309 and 0.5 respectively [27,130]. A simple consequence of this section is the following: $\delta_k < 0.309$ or $\delta_{2k} < 0.5$ are sufficient for robust recovery of matrices with rank at most k improving the previous conditions of $\delta_{2k} < 0.307$ in [156]. Improving the RIP conditions is not the focus of this section, although such improvements have been of independent mathematical interest (e.g., [27,28]).

Our results also apply another recovery condition known as the Nullspace Spherical Section Property (SSP) hence it easily follows from our main theorem that the spherical section constant $\Delta > 4r$ is sufficient for the recovery of matrices up to rank r as in the vector case [233]. This approach not only simplifies the analysis in [87], but also gives a better condition (as compared to $\Delta > 6r$ in [87]). Our final contribution is to give nullspace based conditions for recovery of low-rank matrices using Schatten-p quasi-norm minimization, which is analogous to ℓ_p minimization with $0 < p < 1$ for vectors and has motivated algorithms such as IRLS [65,155] that have been shown to empirically improve on the recovery performance of nuclear norm minimization.

4.2.1 Section specific notation

$\bar{\mathbf{x}}$ denotes the vector obtained by decreasingly sorting the absolute values of the entries of $\mathbf{x}$, and $\mathbf{x}^k$ denotes the vector obtained by restricting $\mathbf{x}$ to its k largest elements (in absolute value). Let $\text{diag}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^n$ return the vector of diagonal entries of a matrix, and $\text{diag}(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ return a diagonal matrix with the entries of the input vector on the diagonal. Let $n_1 \leq n_2$. Denote the ith largest singular value of a matrix $\mathbf{X}$ by $\sigma_i(\mathbf{X})$. Let $\Sigma(\mathbf{X}) = [\sigma_1(\mathbf{X}), \ldots, \sigma_{n_1}(\mathbf{X})]^T$ be vector of decreasingly sorted singular values of $\mathbf{X}$. $\mathbf{X}^k$ denotes the matrix obtained by taking the first k terms in the singular value decomposition of $\mathbf{X}$. Let $\Sigma_{\mathbf{X}} = \text{diag}(\Sigma(\mathbf{X})) \in \mathbb{R}^{n_1\times n_1}$. We call $(\mathbf{U},\mathbf{V})$ a unitary pair if $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{V}^T\mathbf{V} = I$. Then, for the rest of the section, we'll use the following singular value decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\Sigma_{\mathbf{X}}\mathbf{V}^T$ where $(\mathbf{U},\mathbf{V})$ is some unitary pair. Obviously $\mathbf{U} \in \mathbb{R}^{n_1\times n_1}, \mathbf{V} \in \mathbb{R}^{n_2\times n_1}$. Notice that the set of matrices $\mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{D}$ is diagonal, form an n_1 dimensional subspace. Denote this space by $S(\mathbf{U},\mathbf{V})$. Let $\mathscr{A}(\cdot) : R^{n_1\times n_2} \to \mathbb{R}^m$ be a linear operator. $\mathscr{A}_{\mathbf{U},\mathbf{V}}(\cdot) : \mathbb{R}^{n_1} \to \mathbb{R}^m$ is called the restriction of $\mathscr{A}$ to unitary pair $(\mathbf{U},\mathbf{V})$ if we have $\mathscr{A}_{\mathbf{U},\mathbf{V}}(\mathbf{x}) = \mathscr{A}(\mathbf{U}\text{diag}(\mathbf{x})\mathbf{V}^T)$ for all $\mathbf{x} \in \mathbb{R}^{n_1}$. In particular, $\mathscr{A}_{\mathbf{U},\mathbf{V}}(\cdot)$ can be represented by a matrix $\mathbf{A}_{\mathbf{U},\mathbf{V}}$.

Consider the problem of recovering the desired vector $\mathbf{x}_0 \in \mathbb{R}^n$ from corrupted measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$, with $\|\mathbf{z}\|_2 \leq \varepsilon$ where ε denotes the noise energy, and $\mathbf{A} \in \mathbb{R}^{m\times n}$ denotes the measurement matrix. It is known that sparse recovery can be achieved under certain conditions by solving the following convex problem,

$$\begin{aligned} &\text{minimize} && \|\mathbf{x}\|_1 \\ &\text{subject to} && \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \varepsilon, \end{aligned} \tag{4.6}$$

where recovery is known to be robust to noise as well as imperfect sparsity. We say $\mathbf{x}^*$ is *as good as* $\mathbf{x}_0$ if $\|\mathbf{A}\mathbf{x}^* - \mathbf{y}\|_2 \leq \varepsilon$ and $\|\mathbf{x}^*\|_1 \leq \|\mathbf{x}_0\|_1$. In particular, the optimal solution of the problem 4.6 is as good as $\mathbf{x}_0$.

With a slight abuse of notation, let $\mathbf{X}_0 \in \mathbb{R}^{n_1\times n_2}$ with $n = n_1 \leq n_2$ and let $\mathscr{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ be the measurement operator. We observe corrupted measurements $\mathbf{y} = \mathscr{A}(\mathbf{X}_0) + \mathbf{z}$ with $\|\mathbf{z}\|_2 \leq \varepsilon$. For low rank recovery we solve the following convex problem,

$$\begin{aligned} &\text{minimize} && \|\mathbf{X}\|_\star \\ &\text{subject to} && \|\mathscr{A}(\mathbf{X}) - \mathbf{y}\|_2 \leq \varepsilon. \end{aligned} \tag{4.7}$$

Similar to the vector case, we say that $\mathbf{X}^*$ is *as good as* $\mathbf{X}_0$ if $\|\mathscr{A}(\mathbf{X}^*) - \mathbf{y}\|_2 \leq \varepsilon$ and $\|\mathbf{X}^*\|_\star \leq \|\mathbf{X}_0\|_\star$. We now give the definitions for certain recovery conditions on the measurement map, the Restricted Isometry

Property and the Spherical Section Property.

Definition 4.4 (Restricted Isometry Constant) *δ_k of a matrix $\mathbf{A}$ is the smallest constant for which*

$$(1-\delta_k)\|\mathbf{x}\|_2^2 < \|\mathbf{A}\mathbf{x}\|_2^2 < (1+\delta_k)\|\mathbf{x}\|_2^2$$

holds for all vectors $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq k$.

RIP constant δ_k for matrices is defined similarly, with $\mathbf{X}$ instead of $\mathbf{x}$, $\mathscr{A}$ instead of $\mathbf{A}$ and $\text{rank}(\mathbf{X}) \leq k$ instead of $\|\mathbf{x}\|_0 \leq k$.

Definition 4.5 (Restricted Orthogonality Constant) *$\theta_{k,k'}$ of a matrix $\mathbf{A}$ is the smallest constant for which*

$$|\left\langle <\mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}'> \right\rangle| \leq \theta_{k,k'}\|\mathbf{x}\|_2\|\mathbf{x}'\|_2$$

holds for all vectors $\mathbf{x}, \mathbf{x}'$ with disjoint supports and $\|\mathbf{x}\|_0 \leq k$ and $\|\mathbf{x}'\|_0 \leq k'$.

Our definition of matrix ROC will be slightly looser than the one given in [156]. For an operator $\mathscr{A}$, $\theta_{k,k'}$ is the smallest constant for which

$$|\left\langle <\mathscr{A}(\mathbf{X}), \mathscr{A}(\mathbf{X}')> \right\rangle| \leq \theta_{k,k'}\|\mathbf{X}\|_F\|\mathbf{X}'\|_F$$

holds for all matrices $\mathbf{X}, \mathbf{X}'$ such that $\text{rank}(\mathbf{X}) \leq k$, $\text{rank}(\mathbf{X}') \leq k'$ and both column and row spaces of $\mathbf{X}, \mathbf{X}'$ are orthogonal, i.e., in a suitable basis we can write $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & 0 \\ 0 & 0 \end{bmatrix}$ and $\mathbf{X}' = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{X}'_{22} \end{bmatrix}$.

As it will be clear in the subsequent sections, *Restricted Isometry Property (RIP)* is basically a set of conditions on restricted isometry constants of the measurement operator.

Definition 4.6 (Spherical Section Constant) *Spherical section constant $\Delta(\mathscr{A})$ is defined as follows:*

$$\Delta(\mathscr{A}) = \left(\min_{Z \in Null(\mathscr{A}) \setminus \{0\}} \frac{\|Z\|_*}{\|Z\|_F} \right)^2.$$

Furthermore, We say $\mathscr{A}$ satisfies Δ Spherical Section Property (SSP), if $\Delta(\mathscr{A}) \geq \Delta$.

The definition of SSP for a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ for analyzing recovery of sparse vectors is analogous to the above definition. Another way to describe this property is to note that a large Δ implies the nullspace is

an *almost Euclidean* subspace [233], where the ratio of ℓ_1 to ℓ_2 norms cannot be small and therefore the subspace cannot be aligned with the coordinate planes.

4.2.2 Key Observations

Throughout this note, many of the proofs involving matrices are repeated applications of the following useful Lemma which enables us to "vectorize" matrices when dealing with matrix norm inequalities.

Lemma 4.2 *([119])* **(Key Lemma)** *Let* $\mathbf{Z} = \mathbf{X} - \mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$. *Then we have the following inequality:*

$$\sum_{i=1}^{n_1} |\sigma_i(\mathbf{X}) - \sigma_i(\mathbf{Y})| \leq \|\mathbf{Z}\|_\star. \tag{4.8}$$

We now give a useful application of Lemma 4.2.

Lemma 4.3 *Given* $\mathbf{W}$ *with singular value decomposition* $\mathbf{U}\Sigma_{\mathbf{W}}\mathbf{V}^T$, *if there is an* $\mathbf{X}_0$ *for which* $\|\mathbf{X}_0 + \mathbf{W}\|_\star \leq \|\mathbf{X}_0\|_\star$ *then there exists* $\mathbf{X}_1 \in S(\mathbf{U}, \mathbf{V})$ *with* $\Sigma(\mathbf{X}_1) = \Sigma(\mathbf{X}_0)$ *such that* $\|\mathbf{X}_1 + \mathbf{W}\|_\star \leq \|\mathbf{X}_1\|_\star$. *In particular this is true for* $\mathbf{X}_1 = -\mathbf{U}\Sigma_{\mathbf{X}_0}\mathbf{V}^T$.

Proof: From Lemma 4.2 we have

$$\|\mathbf{X}_0 + \mathbf{W}\|_\star \geq \sum_i |\sigma_i(\mathbf{X}_0) - \sigma_i(\mathbf{W})|. \tag{4.9}$$

On the other hand, for $\mathbf{X}_1 = -\mathbf{U}\Sigma_{\mathbf{X}_0}\mathbf{V}^T, \mathbf{W}$ we have

$$\|\mathbf{X}_1 + \mathbf{W}\|_\star = \| - \Sigma_{\mathbf{X}_0} + \Sigma_{\mathbf{W}}\|_\star = \sum_i |\sigma_i(\mathbf{X}_0) - \sigma_i(\mathbf{W})|. \tag{4.10}$$

Then from (4.9) and (4.10) it follows

$$\|\mathbf{X}_1 + \mathbf{W}\|_\star \leq \|\mathbf{X}_0 + \mathbf{W}\|_\star \leq \|\mathbf{X}_0\|_\star = \|\mathbf{X}_1\|_\star.$$

■

Although, Lemma 4.3 is trivial to prove its implications are important. It suggests that if there exists a "bad" $\mathbf{X}_0$ for a particular perturbation $\mathbf{W}$, then there is also a "bad" $\mathbf{X}_1$, which is "similar" to $\mathbf{X}_0$, but lies on the same restricted subspace $S(\mathbf{U}, \mathbf{V})$ as $\mathbf{W}$. On the other hand, as will be clear in a moment, if we have

a guarantee that none of such subspaces contains a bad $(\mathbf{X}_1, \mathbf{W})$ pair, then we can also guarantee that there won't be a bad $(\mathbf{X}_0, \mathbf{W})$ pair even if we consider a union of subspaces.

To further illustrate the similarity between sparse and low rank recovery, we state the null space conditions for noiseless recovery.

Lemma 4.4 *([96])* **Null space condition for sparse recovery**

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ *be a measurement matrix. Assume* $\varepsilon = 0$ *then one can perfectly recover all vectors* $\mathbf{x}_0$ *with* $\|\mathbf{x}_0\|_0 \leq k$ *via program 4.6 if and only if for any* $\mathbf{w} \in Null(\mathbf{A})$ *we have:*

$$\sum_{i=1}^{k} \bar{w}_i < \sum_{i=k+1}^{n} \bar{w}_i, \tag{4.11}$$

where $\bar{w}_i$ *is the ith entry of* $\bar{\mathbf{w}}$ *defined previously.*

We now state a critical result, that provides an "if and only if" condition for the strong recovery of low-rank matrices via NNM. Our condition is strikingly similar to Lemma 4.4.

Proposition 4.3 (also see [163]) Null space condition for low-rank recovery

Let $\mathscr{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ *be a linear measurement operator. Assume* $\varepsilon = 0$ *then one can recover all matrices* $\mathbf{X}_0$ *with* $\text{rank}(\mathbf{X}_0) \leq k$ *via program 4.7 if and only if for any* $\mathbf{W} \in Null(\mathscr{A})$ *we have:*

$$\sum_{i=1}^{k} \sigma_i(\mathbf{W}) < \sum_{i=k+1}^{n_1} \sigma_i(\mathbf{W}). \tag{4.12}$$

Proof: Suppose (4.12) holds for all $\mathbf{W} \in \text{Null}(\mathscr{A})$. Then, for any feasible perturbation $\mathbf{W}$, using Lemma 4.2,

$$\sum_{i=1}^{n_1} |\sigma_i(\mathbf{X}_0) - \sigma_i(\mathbf{W})| \leq \|\mathbf{X}_0 + \mathbf{W}\|_\star.$$

Now, to conclude with $\|\mathbf{X}_0 + \mathbf{W}\|_\star > \|\mathbf{X}_0\|_\star$, observe that $\sigma_i(\mathbf{X}_0) = 0$ for $i > k$ and,

$$\sum_{i=1}^{n_1} |\sigma_i(\mathbf{X}_0) - \sigma_i(\mathbf{W})| \geq \sum_{i=1}^{k} (\sigma_i(\mathbf{X}_0) - \sigma_i(\mathbf{W})) + \sum_{i=k+1}^{n_1} \sigma_i(\mathbf{W}) > \|\mathbf{X}_0\|_\star$$

To show the failure result, assume there exists a $\mathbf{W} \in \text{Null}(\mathscr{A})$ for which (4.12) does not hold. Assume $\mathbf{W}$ has SVD $\mathbf{U}\,\text{diag}\,(\mathbf{w})\,\mathbf{V}^T$ and choose $\mathbf{X}_0 = \mathbf{U}\,\text{diag}\,(\mathbf{d})\,\mathbf{V}^T$ where $\mathbf{d}_i = -\mathbf{w}_i$ for $1 \leq i \leq k$ and 0 else. Clearly,

$\text{rank}(\mathbf{X}_0) \leq k$ and

$$\|\mathbf{X}_0 + \mathbf{W}\|_\star = \|\mathbf{d} + \mathbf{w}\|_1 = \sum_{i=k+1}^{n_1} \sigma_i(\mathbf{W}) \leq \sum_{i=1}^{k} \sigma_i(\mathbf{W}) = \|\mathbf{w}\|_1 = \|\mathbf{X}_0\|_\star$$

hence, $\mathbf{X}_0$ is not the unique minimizer. ■

4.2.3 Main Result

Definition 4.7 (Extension) *Let* $\mathbf{P}$ *be a property defined for matrices in* $\mathbb{R}^n \to \mathbb{R}^m$. *We denote* extension *of* $\mathbf{P}$ *by* $\mathbf{P}^e$ *which is a property of linear operators* $\mathscr{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ *as follows:*
$\mathscr{A}(\cdot)$ *has property* $\mathbf{P}^e$ *if all its restrictions,* $\mathbf{A}_{\mathbf{U},\mathbf{V}}$, *have property* $\mathbf{P}$.

In this section, we state our main result which enables us to translate vector recovery to matrix recovery via extension. In particular, as we discuss later, standard RIP and SSP based conditions for matrices implies extensions of RIP and SSP based conditions for vectors, thus enabling translation of results from vectors to matrices as an application of the main theorem. Let $\|\cdot\|_v$ be an arbitrary norm on $\mathbb{R}^n$ with $\|\mathbf{x}\|_v = \|\bar{\mathbf{x}}\|_v$ for all $\mathbf{x}$. Let $\|\cdot\|_m$ be the corresponding unitarily invariant matrix norm on $\mathbb{R}^{n_1 \times n_2}$ such that $\|\mathbf{X}\|_m = \|\Sigma(\mathbf{X})\|_v$. For the sake of clarity, we use the following shorthand notation for statements regarding recovery of vectors, in the main theorem:

- V_1: A matrix $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ satisfies a property $\mathbf{P}$.
- V_2: In program 4.6, for any $\mathbf{x}_0$, $\|\mathbf{z}\|_2 \leq \varepsilon$, $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + z$ and any $\mathbf{x}^*$ as good as $\mathbf{x}_0$ we have,

$$\|\mathbf{x}^* - \mathbf{x}_0\|_v \leq h(\mathbf{x}_0, \varepsilon).$$

- V_3: For any $\mathbf{w} \in \text{Null}(\mathbf{A})$, $\mathbf{w}$ satisfies a certain property $\mathbf{Q}$.

We also use the following shorthand for statements regarding recovery of matrices, in the main theorem:

- M_1: A linear operator $\mathscr{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ satisfies the extension property $\mathbf{P}^e$.
- M_2: In program 4.7, for any $\mathbf{X}_0$, $\|\mathbf{z}\|_2 \leq \varepsilon$, $\mathbf{y} = \mathscr{A}(\mathbf{X}_0) + \mathbf{z}$ and any $\mathbf{X}^*$ as good as $\mathbf{X}_0$ we have,

$$\|\mathbf{X}^* - \mathbf{X}_0\|_m \leq h(\Sigma(\mathbf{X}_0), \varepsilon).$$

- M_3: For any $\mathbf{W} \in \text{Null}(\mathscr{A})$, $\Sigma(\mathbf{W})$ satisfies property $\mathbf{Q}$.

Theorem 4.2 (Main Theorem) *For a given* $\mathbf{P}$*, the following implications hold true:*

$$(V_1 \implies V_2) \implies (M_1 \implies M_2). \tag{4.13}$$

$$(V_1 \implies V_3) \implies (M_1 \implies M_3). \tag{4.14}$$

Proof: $(V_1 \implies V_2) \implies (M_1 \implies M_2)$.

Assume $V_1 \implies V_2$ and $\mathscr{A}$ satisfies $\mathbf{P}^e$. Consider program 4.7 where measurements are $\mathbf{y}_0 = \mathscr{A}(\mathbf{X}_0) + \mathbf{z}_0$ with $\|\mathbf{z}_0\|_2 \leq \varepsilon$. Also let $\mathbf{X}^* = \mathbf{X}_0 + \mathbf{W}$ be as good as $\mathbf{X}_0$. This implies that $\|\mathbf{X}_0 + \mathbf{W}\|_\star \leq \|\mathbf{X}_0\|_\star$ and $\|\mathscr{A}(\mathbf{X}_0 + \mathbf{W}) - \mathbf{y}_0\|_2 \leq \varepsilon$. Then, from Lemma 4.3 for $\mathbf{X}_1 = -\mathbf{U}\Sigma_{\mathbf{X}_0}\mathbf{V}^T$ (where $\mathbf{W}$ has SVD $\mathbf{U}\Sigma_{\mathbf{W}}\mathbf{V}^T$) we have $\|\mathbf{X}_1 + \mathbf{W}\|_\star \leq \|\mathbf{X}_1\|_\star$. Now, let $\mathbf{y}_1 = \mathscr{A}(\mathbf{X}_1) + \mathbf{z}_0$. Clearly

$$\|\mathscr{A}(\mathbf{X}_1 + \mathbf{W}) - \mathbf{y}_1\|_2 = \|\mathscr{A}(\mathbf{X}_0 + \mathbf{W}) - \mathbf{y}_0\|_2 \leq \varepsilon$$

hence $\mathbf{X}_1 + \mathbf{W}$ is as good as $\mathbf{X}_1$. Now consider program 1 with $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ as measurement matrix, $\mathbf{y}_1$ as measurements, $\mathbf{x}_1 = -\Sigma(\mathbf{X}_0)$ as unknown vector and $\mathbf{w} = \Sigma(\mathbf{W})$ to be perturbation. Notice that $\mathbf{x}_1 + \mathbf{w}$ is as good as $\mathbf{x}_1$. Also since $\mathscr{A}$ has $\mathbf{P}^e$, $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ has $\mathbf{P}$. Using $V_1 \implies V_2$ we conclude

$$\|\mathbf{W}\|_m = \|\mathbf{w}\|_v \leq h(\bar{\mathbf{x}}_1, \varepsilon) = h(\Sigma(\mathbf{X}_0), \varepsilon)$$

Hence, we found: $M_1 \implies M_2$.

Using similar arguments, now we show that $(V_1 \implies V_3) \implies (M_1 \implies M_3)$.

Assume $V_1 \implies V_3$ and $\mathscr{A}$ has $\mathbf{P}^e$. Consider any $\mathbf{W} \in \text{Null}(\mathscr{A})$ with SVD of $\mathbf{W} = \mathbf{U}\Sigma_{\mathbf{W}}\mathbf{V}^T$. Then, $\mathscr{A}(\mathbf{W}) = \mathbf{A}_{\mathbf{U},\mathbf{V}}\Sigma(\mathbf{W}) = 0$. Also $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ satisfies $\mathbf{P}$. Using $V_1 \implies V_3$, we find $\Sigma(\mathbf{W})$ satisfies $\mathbf{Q}$. Hence $M_1 \implies M_3$. ■

As it can be seen from Main Theorem, throughout this section, we are actually dealing with a *strong* notion of recovery. By strong we mean, $\mathbf{P}$ guarantees a recovery result for all $\mathbf{x}_0$ ($\mathbf{X}_0$) with sparsity (rank) at most k instead of a particular $\mathbf{x}_0$ ($\mathbf{X}_0$). For example, matrix completion results in the literature don't have *strong* recovery. On the other hand it is known that (good) RIP or SSP conditions guarantee recoverability for all vectors hence they result in a *strong* recovery.

To apply the main theorem, we require linear maps on matrices to satisfy the extensions of the properties of linear maps on vectors. Below, we apply our results to RIP and SSP.

4.2.3.1 Applications of Main Theorem to RIP based recovery

We first show that RIP for matrices implies extension of RIP for vectors thus Theorem 4.2 is applicable for RIP. We say $f(\delta_{i_1},\ldots,\delta_{i_m},\theta_{j_1,j'_1},\ldots,\theta_{j_n,j'_n}) \leq c$ is an RIP inequality where $c \geq 0$ is a constant and $f(\cdot)$ is an increasing function of its parameters (RIC and ROC) and $f(0,\ldots,0) = 0$. Let $\mathbf{F}$ be a set of RIP inequalities namely $f_1,\ldots,f_N$ where k'th inequality is of the form:

$$f_k(\delta_{i_{k,1}},\ldots,\delta_{i_{k,m_k}},\theta_{j_{k,1},j'_{k,1}},\ldots,\theta_{j_{k,n_k},j'_{k,n_k}}) \leq c_k.$$

Lemma 4.5 *If $\mathscr{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ satisfies a set of (matrix) RIP and ROC inequalities $\mathbf{F}$, then for all unitary pairs $(\mathbf{U},\mathbf{V})$, $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ will satisfy the same inequalities, thus RIP for matrices implies extension of RIP for vectors.*

Proof: Let $\delta'_k, \theta'_{k,k'}$ denote RIC and ROC of of $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ and $\delta_k, \theta_{k,k'}$ denote RIC and ROC of $\mathscr{A}(\cdot)$. Then we claim: $\delta'_k \leq \delta_k$ and $\theta'_{k,k'} \leq \theta_{k,k'}$. For any $\mathbf{x}$ of $\|\mathbf{x}\|_0 \leq k$, let $\mathbf{X} = \mathbf{U}\text{diag}(\mathbf{x})\mathbf{V}^T$. Using $\|\mathbf{x}\|_2 = \|\mathbf{X}\|_F$ and $\mathbf{A}_{\mathbf{U},\mathbf{V}}\mathbf{x} = \mathscr{A}(\mathbf{X})$ we have:

$$(1-\delta_k)\|\mathbf{x}\|_2^2 < \|\mathscr{A}(\mathbf{X})\|_F^2 = \|\mathbf{A}_{\mathbf{U},\mathbf{V}}\mathbf{x}\|_2^2 < (1+\delta_k)\|\mathbf{x}\|_2^2$$

Hence $\delta'_k \leq \delta_k$. Similarly let $\mathbf{x},\mathbf{x}'$ have disjoint supports with sparsity at most k,k' respectively. Then obviously $\mathbf{X} = \mathbf{U}\text{diag}(\mathbf{x})\mathbf{V}^T$ and $\mathbf{X}' = \mathbf{U}\text{diag}(\mathbf{x}')\mathbf{V}^T$ satisfies the condition in ROC definition. Hence:

$$\begin{aligned} |\left\langle <\mathbf{A}_{\mathbf{U},\mathbf{V}}\mathbf{x}, \mathbf{A}_{\mathbf{U},\mathbf{V}}\mathbf{x}'\right\rangle > | &= |\left\langle <\mathscr{A}(\mathbf{X}), \mathscr{A}(\mathbf{X}')\right\rangle > | \\ &\leq \theta_{k,k'}\|\mathbf{X}\|_F\|\mathbf{X}'\|_F = \theta_{k,k'}\|\mathbf{x}\|_2\|\mathbf{x}'\|_2 \end{aligned}$$

Hence $\theta'_{k,k'} \leq \theta_{k,k'}$. Thus, $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ satisfies the set of inequalities $\mathbf{F}$ as $f_i(\cdot)$'s are increasing function of δ_k's and $\theta_{k,k'}$'s. ■

By using this observation and the main theorem, we can smoothly translate any implication of RIP for vectors to corresponding implication for matrices. In particular, some typical RIP implications are as follows.

Proposition 4.4 (RIP implications for k-sparse recovery ([32])) *Suppose $A : \mathbb{R}^n \to \mathbb{R}^m$ satisfies, a set of RIP inequalities $\mathbf{F}$. Then for all $\mathbf{x}_0$, $\|\mathbf{z}\|_2 \leq \varepsilon$ and $\mathbf{x}^*$ as good as $\mathbf{x}_0$ we have the following ℓ_2 and ℓ_1 robustness results,*

$$\|\mathbf{x}_0-\mathbf{x}^*\|_2 \leq \frac{C_1}{\sqrt{k}}\|\mathbf{x}_0-\mathbf{x}_0^k\|_1+C_2\varepsilon \tag{4.15}$$
$$\|\mathbf{x}_0-\mathbf{x}^*\|_1 \leq C_3\|\mathbf{x}_0-\mathbf{x}_0^k\|_1$$

For some constants $C_1, C_2, C_3 > 0$.

Now, using Theorem 4.2, we translate these implications to matrices in the following lemma.

Lemma 4.6 *Suppose* $\mathscr{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ *satisfies the same inequalities* $\mathbf{F}$ *as in (4.15). Then for all* $\mathbf{X}_0$, $\|\mathbf{z}\|_2 \leq \varepsilon$ *and* $\mathbf{X}^*$ *as good as* $\mathbf{X}_0$ *we have the following Frobenius norm and nuclear norm robustness results,*

$$\|\mathbf{X}_0-\mathbf{X}^*\|_F \leq \frac{C_1}{\sqrt{k}}\|\mathbf{X}_0-\mathbf{X}_0^k\|_\star+C_2\varepsilon$$
$$\|\mathbf{X}_0-\mathbf{X}^*\|_\star \leq C_3\|\mathbf{X}_0-\mathbf{X}_0^k\|_\star$$

4.2.3.2 Application of Main Theorem to SSP based recovery

The following lemma suggests that SSP for linear operators on matrices implies extension of SSP for linear operators on vectors.

Lemma 4.7 *Let* $\Delta > 0$. *If* $\mathscr{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ *satisfies* Δ*-SSP, then for all unitary pairs* $(\mathbf{U},\mathbf{V})$, $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ *satisfies* Δ*-SSP.*

Proof: Consider any $\mathbf{w} \in \text{Null}(\mathscr{A}_{\mathbf{U},\mathbf{V}})$. Then, $\mathscr{A}(\mathbf{U}\text{diag}(\mathbf{w})\mathbf{V}^T) = 0$ and therefore $\mathbf{W} = \mathbf{U}\text{diag}(\mathbf{w})\mathbf{V}^T \in \text{Null}(\mathscr{A})$. Since $\mathscr{A}$ satisfies Δ-SSP, we have $\frac{\|\mathbf{W}\|_\star}{\|\mathbf{W}\|_F} = \frac{\|\mathbf{w}\|_1}{\|\mathbf{w}\|_2} \geq \sqrt{\Delta(\mathscr{A})} \geq \sqrt{\Delta}$. Thus $\mathbf{A}_{\mathbf{U},\mathbf{V}}$ satisfies Δ-SSP. ■

Now we give the following SSP based result for matrices as an application of main theorem.

Theorem 4.3 *Consider program 4.7 with* $\mathbf{z} = 0$, $\mathbf{y} = \mathscr{A}(\mathbf{X}_0)$. *Let* $\mathbf{X}^*$ *be as good as* $\mathbf{X}_0$. *Then if* $\mathscr{A}$ *satisfies* Δ*-SSP with* $\Delta > 4r$, *it holds that*

$$\|\mathbf{X}^*-\mathbf{X}_0\|_\star\| \leq C\|\mathbf{X}_0-\mathbf{X}_0^r\|_\star$$

where $C = \frac{2}{1-2\sqrt{r/\Delta}}$.

Note that the use of main theorem and *Key Lemma* simplifies the recovery analysis in [87] and also improves the sufficient condition of $r < \frac{\Delta}{6}$ in [87] to $r < \frac{\Delta}{4}$. This improved sufficient condition matches the sufficient condition given in [233] for the sparse vector recovery problem.

4.2.4 Simplified Robustness Conditions

We show that various robustness conditions are equivalent to simple conditions on the measurement operator. The case of noiseless and perfectly sparse signals, is already given in Lemmas 4.4 and 4.3. Such simple conditions might be useful for analysis of nuclear norm minimization in later works. We state the conditions for matrices only; however, vector and matrix conditions will be identical (similar to Lemmas 4.4, 4.3) as one can expect from Theorem 4.2. The proofs follow from simple algebraic manipulations with the help of Lemma 4.2.

Lemma 4.8 (Nuclear Norm Robustness for Matrices)
Assume $\varepsilon = 0$ (no noise). Let $C > 1$ be constant. Then for any $\mathbf{X}_0$ and any $\mathbf{X}^*$ as good as $\mathbf{X}_0$ we are guaranteed to have:

$$\|\mathbf{X}_0 - \mathbf{X}^*\|_\star < 2C\|\mathbf{X}_0 - \mathbf{X}_0^k\|_\star$$

if and only if for all $\mathbf{W} \in Null(\mathscr{A})$ we have:

$$\|\mathbf{W}^k\|_\star < \frac{C-1}{C+1}\|\mathbf{W} - \mathbf{W}^k\|_\star$$

Lemma 4.9 (Frobenius Norm Robustness for Matrices)
Let $\varepsilon = 0$. Then for any $\mathbf{X}_0$ and $\mathbf{X}^*$ as good as $\mathbf{X}_0$,

$$\|\mathbf{X}_0 - \mathbf{X}^*\|_F < \frac{C}{\sqrt{k}}\|\mathbf{X}_0 - \mathbf{X}_0^k\|_\star,$$

if and only if for all $\mathbf{W} \in Null(\mathscr{A})$,

$$\|\mathbf{W} - \mathbf{W}^k\|_\star - \|\mathbf{W}^k\|_\star > \frac{2\sqrt{k}}{C}\|\mathbf{W}\|_F$$

Lemma 4.10 (Matrix Noise Robustness) *For any $\mathbf{X}_0$ with $\mathrm{rank}(\mathbf{X}_0) \leq k$, any $\|\mathbf{z}\|_2 \leq \varepsilon$ and any $\mathbf{X}^*$ as good as $\mathbf{X}_0$,*

$$\|\mathbf{X}_0 - \mathbf{X}^*\|_F < C\varepsilon,$$

if and only if for any $\mathbf{W}$ with $\|\mathbf{W}^k\|_\star \geq \|\mathbf{W} - \mathbf{W}^k\|_\star$,

$$\|\mathbf{W}\|_F < \frac{C}{2}\|\mathscr{A}(\mathbf{W})\|_2$$

4.2.5 Null space based recovery result for Schatten-p quasi-norm minimization

In the previous sections, we stated the main theorem and considered its applications on RIP and SSP based conditions to show that results for recovery of sparse vectors can be analogously extended to recovery of low-rank matrices without making the recovery conditions stronger. In this section, we consider extending results from vectors to matrices using an algorithm different from ℓ_1 minimization or nuclear norm minimization.

The ℓ_p quasi-norm (with $0 < p < 1$) is given by $\|\mathbf{x}\|_p^p = \sum_{i=1}^n |\mathbf{x}_i|^p$. Note that for $p = 0$, this is nothing but the cardinality function. Thus it is natural to consider the minimization of the ℓ_p quasi-norm (as a surrogate for minimizing the cardinality function). Indeed, ℓ_p minimization has been a starting point for algorithms including Iterative Reweighted Least Squares [65] and Iterative Reweighted ℓ_1 minimization [29, 100]. Note that although ℓ_1 minimization is convex, ℓ_p minimization with $0 < p < 1$ is *non-convex*. However empirically, ℓ_p minimization based algorithms with $0 < p < 1$ have a better recovery performance as compared to ℓ_1 minimization (see e.g. [54], [100]). The recovery analysis of these algorithms has mostly been based on RIP. However Null space based recovery conditions analogous to those for ℓ_1 minimization have been given for ℓ_p minimization (see e.g. [219]).

Let $\operatorname{Tr}|\mathbf{A}|^p = \operatorname{Tr}(\mathbf{A}^T\mathbf{A})^{\frac{p}{2}} = \sum_{i=1}^n \sigma_i^p(\mathbf{A})$ denote the *Schatten-p quasi norm* with $0 < p < 1$. Analogous to the vector case, one can consider the minimization of the Schatten-p quasi-norm for the recovery of low-rank matrices,

$$\begin{aligned} &\text{minimize} && \operatorname{Tr}|\mathbf{X}|^p \\ &\text{subject to} && \mathscr{A}(\mathbf{X}) = y \end{aligned} \tag{4.16}$$

where $y = \mathscr{A}(\mathbf{X}_0)$ with $\mathbf{X}_0$ being the low-rank solution we wish to recover. IRLS-p has been proposed as an algorithm to find a local minimum to (4.16) in [155]. However no null-space based recovery condition has been given for the recovery analysis of Schatten-p quasi norm minimization. We give such a condition below, after mentioning a few useful inequalities.

Lemma 4.11 ([210]) *For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ it holds that $\sum_{i=1}^k (\sigma_i^p(\mathbf{A}) - \sigma_i^p(\mathbf{B})) \leq \sum_{i=1}^k \sigma_i^p(\mathbf{A} - \mathbf{B})$ for all $k = 1, 2, \ldots, n$.*

Note that the p quasi-norm of a vector satisfies the triangle inequality ($x, y \in \mathbb{R}^n$, $\sum_{i=1}^n |x_i + y_i|^p \leq \sum_{i=1}^n |x_i|^p + \sum_{i=1}^n |y_i|^p$). Lemma 4.11 generalizes this result to matrices.

Lemma 4.12 ([118]) *For any two matrices P, Q it holds that*

$$\sigma_{t+s-1}(P+Q) \leq \sigma_t(P) + \sigma_s(Q)$$

where $t+s-1 \leq n$ and $t, s \geq 0$.

The following lemma easily follows as a consequence.

Lemma 4.13 *For any two matrices, $\mathbf{A}, \mathbf{B}$ with $\mathbf{B}$ of rank r and any $p > 0$,*

$$\sum_{i=r+1}^{n-r} \sigma_i^p(\mathbf{A}-\mathbf{B}) \geq \sum_{i=2r+1}^{n} \sigma_i^p(\mathbf{A}).$$

Theorem 4.4 *Let $\mathrm{tr}((\mathbf{X}_0)) = r$ and let $\bar{\mathbf{X}}$ denote the global minimizer of (4.16). A sufficient condition for $\bar{\mathbf{X}} = \mathbf{X}_0$ is that $\sum_{i=1}^{2r} \sigma_i^p(\mathbf{W}) \leq \sum_{i=2r+1}^{n} \sigma_i^p(\mathbf{W})$ for all $\mathbf{W} \in Null(\mathscr{A})$. A necessary condition for $\bar{\mathbf{X}} = \mathbf{X}_0$ is that $\sum_{i=1}^{r} \sigma_i^p(\mathbf{W}) \leq \sum_{i=r+1}^{n} \sigma_i^p(\mathbf{W})$ for all $\mathbf{W} \in Null(\mathscr{A})$.*

Proof: ($\Rightarrow$) For any $\mathbf{W} \in \mathrm{Null}(\mathscr{A})\backslash\{0\}$,

$$\begin{aligned}
\mathrm{Tr}\,|\mathbf{X}_0+\mathbf{W}|^p &= \sum_{i=1}^{r} \sigma_i^p(\mathbf{X}_0+\mathbf{W}) + \sum_{i=r+1}^{n} \sigma_i^p(\mathbf{X}_0+\mathbf{W}) \\
&\geq \sum_{i=1}^{r} (\sigma_i^p(\mathbf{X}_0) - \sigma_i^p(\mathbf{W})) + \sum_{i=2r+1}^{n} \sigma_i^p(\mathbf{W}) \\
&\geq \mathrm{Tr}\,|\mathbf{X}_0|^p
\end{aligned}$$

where the first inequality follows from Lemma 4.11 and Lemma 4.13. The necessary condition is easy to show analgous to the results for nuclear norm minimization. ■

Note that there is a gap between the necessary and sufficient conditions. We observe through numerical experiments that a better inequality such as $\sum_{i=1}^{n} \sigma_i^p(\mathbf{A}-\mathbf{B}) \geq \sum_{i=1}^{n} |\sigma_i^p(\mathbf{A}) - \sigma_i^p(\mathbf{B})|$ seems to hold for any two matrices $\mathbf{A}, \mathbf{B}$. Note that this is in particular true for $p = 1$ (Lemma 4.2) and $p = 0$. If this inequality is proven true for all $0 < p < 1$, then we could bridge the gap between necessity and sufficiency in Theorem 4.4. Thus we have that singular value inequalities including those in Lemma 4.2 and Lemma 4.13, 4.11 play a fundamental role in extending recovery results from vectors to matrices. Although, our condition is not tight, we can still use the Theorems 4.2 and 4.4 to conclude the following:

Lemma 4.14 *Assume property* **S** *on matrices* $\mathbb{R}^n \to \mathbb{R}^m$ *implies perfect recovery of all vectors with sparsity at most* $2k$ *via* ℓ_p *quasi-norm minimization where* $0 < p < 1$. *Then* $\mathbf{S}^e$ *implies perfect recovery of all matrices with rank at most* k *via Schatten-*p *quasi-norm minimization.*

Proof Idea: Since **S** implies perfect recovery of all vectors with sparsity at most $2k$, it necessarily implies a *null space property* (call it **Q**) similar to the one given in Lemma 4.4, (4.11) but with p in the exponent (see e.g. [219]) and k replaced by $2k$. Now the main theorem, (4.14), combined with Theorem 4.4 implies that $\mathbf{S}^e$ is sufficient for perfect revovery of rank k matrices.

In particular, using Lemma 4.14, we can conclude, any set of RIP conditions that are sufficient for recovery of vectors of sparsity up to $2k$ via ℓ_p minimization, are also sufficient for recovery of matrices of rank up to k via Schatten-p minimization. As an immediate consequence it follows that the results in [53, 100] can be easily extended.

4.2.6 Conclusions

This section presented a general result stating that *extension* of any sufficient condition for the recovery of sparse vectors using ℓ_1 minimization is also sufficient for the recovery of low-rank matrices using nuclear norm minimization. As an immediate consequence of this result, we have that the best known RIP-based recovery conditions of $\delta_k < 0.309$, $\delta_{2k} < 0.5$ for sparse vector recovery is also sufficient for low-rank matrix recovery. We also show that a Null-space based sufficient condition (Spherical Section Property) given in [233] easily extends to the matrix case, tightening the existing conditions for low-rank matrix recovery [87]. Finally, we gave null-space based conditions for recovery using Schatten-p quasi-norm minimization and showed that RIP based conditions for ℓ_p minimization extend to the matrix case. We note that all of these results rely on the ability to "vectorize" matrices through the use of key singular value inequalities including Lemma 4.2, 4.11.

Chapter 5

Simultaneously Structured Models

There are many applications where the model of interest is known to have *several* structures at the same time (Section 5.0.8). We then seek a signal that lies in the intersection of several sets defining the individual structures (in a sense that we will make precise later). The most common convex regularizer (penalty) used to promote all structures together is a linear combination of well-known regularizers for each structure. However, there is currently no general analysis and understanding of how well such regularization performs in terms of the number of observations required for successful recovery of the desired model. This chapter addresses this ubiquitous yet unexplored problem; i.e., the recovery of *simultaneously structured models*.

An example of a simultaneously structured model is a matrix that is *simultaneously sparse and low-rank*. One would like to come up with algorithms that exploit both types of structures to minimize the number of measurements required for recovery. An $n \times n$ matrix with rank $r \ll n$ can be described by $\mathcal{O}(rn)$ parameters, and can be recovered using $\mathcal{O}(rn)$ generic measurements via nuclear norm minimization [39, 178]. On the other hand, a block-sparse matrix with a $k \times k$ nonzero block where $k \ll n$ can be described by k^2 parameters and can be recovered with $\mathcal{O}\left(k^2 \log \frac{n}{k}\right)$ generic measurements using ℓ_1 minimization. However, a matrix that is *both* rank r and block-sparse can be described by $\mathcal{O}(rk)$ parameters. The question is whether we can exploit this joint structure to efficiently recover such a matrix with $\mathcal{O}(rk)$ measurements.

In this chapter we give a negative answer to this question in the following sense: if we use multi-objective optimization with the ℓ_1 and nuclear norms (used for sparse signals and low rank matrices, respectively), then the number of measurements required is lower bounded by $\mathcal{O}\left(\min\{k^2, rn\}\right)$. In other words, we need at least this number of observations for the desired signal to lie on the Pareto optimal front traced by the ℓ_1 norm and the nuclear norm. This means we can do *no better than an algorithm that exploits only one of the two structures*.

We introduce a framework to express general simultaneous structures, and as our main result, we prove that the same phenomenon happens for a general set of structures. We are able to analyze a wide range of measurement ensembles, including subsampled standard basis (i.e. matrix completion), Gaussian and subgaussian measurements, and quadratic measurements. Table 5.1 summarizes known results on recovery of some common structured models, along with a result of this chapter specialized to the problem of low-rank and sparse matrix recovery. The first column gives the number of parameters needed to describe the model (often referred to as its 'degrees of freedom'), the second and third columns show how many generic measurements are needed for successful recovery. In 'nonconvex recovery', we assume we are able to find the global minimum of a nonconvex problem. This is clearly intractable in general, and not a practical recovery method—we consider it as a benchmark for theoretical comparison with the (tractable) convex relaxation in order to determine how powerful the relaxation is.

The first and second rows are the results on k sparse vectors in $\mathbb{R}^n$ and rank r matrices in $\mathbb{R}^{n\times n}$ respectively, [43, 45]. The third row considers the recovery of "low-rank plus sparse" matrices. Consider a matrix $\mathbf{X}\in\mathbb{R}^{n\times n}$ that can be decomposed as $\mathbf{X}=\mathbf{X}_L+\mathbf{X}_S$ where $\mathbf{X}_L$ is a rank r matrix and $\mathbf{X}_S$ is a matrix with only k nonzero entries. The degrees of freedom of $\mathbf{X}$ is $\mathscr{O}(rn+k)$. Minimizing the infimal convolution of ℓ_1 norm and nuclear norm, i.e., $f(\mathbf{X})=\min_{\mathbf{Y}}\|\mathbf{Y}\|_\star+\lambda\|\mathbf{X}-\mathbf{Y}\|_1$ subject to random Gaussian measurements on $\mathbf{X}$, gives a convex approach for recovering $\mathbf{X}$. It has been shown that under reasonable incoherence assumptions, $\mathbf{X}$ can be recovered from $\mathscr{O}\left((rn+k)\log^2 n\right)$ measurements which is suboptimal only by a logarithmic factor [222]. Finally, the last row in Table 5.1 shows one of the results in this chapter. Let $\mathbf{X}\in\mathbb{R}^{n\times n}$ be a rank r matrix whose entries are zero outside a $k_1\times k_2$ submatrix. The degrees of freedom of $\mathbf{X}$ is $\mathscr{O}((k_1+k_2)r)$. We consider both convex and non-convex programs for the recovery of this type of matrices. The nonconvex method involves minimizing the number of nonzero rows, columns and rank of the matrix jointly, as discussed in Section 5.2.2. As shown later, $\mathscr{O}((k_1+k_2)r\log n)$ measurements suffices for this program to successfully recover the original matrix. The convex method minimizes any convex combination of the individual structure-inducing norms, namely the nuclear norm and the $\ell_{1,2}$ norm of the matrix, which encourage low-rank and column/row-sparse solutions respectively. We show that with high probability this program cannot recover the original matrix with fewer than $\Omega(rn)$ measurements. In summary, while nonconvex method is slightly suboptimal, the convex method performs poorly as the number of measurements scales with n rather than k_1+k_2.

Model	Degrees of Freedom	Nonconvex recovery	Convex recovery
Sparse vectors	k	$\mathcal{O}(k)$	$\mathcal{O}\left(k \log \frac{n}{k}\right)$
Low rank matrices	$r(2n-r)$	$\mathcal{O}(rn)$	$\mathcal{O}(rn)$
Low rank plus sparse	$\mathcal{O}(rn+k)$	not analyzed	$\mathcal{O}\left((rn+k)\log^2 n\right)$
Low rank and sparse	$\mathcal{O}(r(k_1+k_2))$	$\mathcal{O}(r(k_1+k_2)\log n)$	$\Omega(rn)$

Table 5.1: Summary of results in recovery of structured signals. This chapter shows a gap between the performance of convex and nonconvex recovery programs for simultaneously structured matrices (last row).

5.0.7 Contributions

This chapter describes a general framework for analyzing the recovery of models that have more than one structure, by combining penalty functions corresponding to each structure. The framework proposed includes special cases that are of interest in their own right, e.g., sparse and low-rank matrix recovery and low-rank tensor completion [103, 114]. Our contributions can be summarized as follows.

Poor performance of convex relaxations. We consider a model with several structures and associated structure-inducing norms. For recovery, we consider a multi-objective optimization problem to minimize the individual norms simultaneously. Using Pareto optimality, we know that minimizing a weighted sum of the norms and varying the weights traces out all points of the Pareto-optimal front (i.e., the trade-off surface, Section 5.1). We obtain a lower bound on the number of measurements for any convex function combining the individual norms. A sketch of our main result is as follows.

> *Given a model* $\mathbf{x}_0$ *with* τ *simultaneous structures, the number of measurements required for recovery with high probability using any linear combination of the individual norms satisfies the lower bound*
>
> $$m \geq c\, m_{\min} = c \min_{i=1,\ldots,\tau} m_i$$
>
> *where* m_i *is an intrinsic lower bound on the required number of measurements when minimizing the* i*th norm only. The term* c *depends on the measurement ensemble we are dealing with.*

For the norms of interest, m_i will be approximately proportional to the degrees of freedom of the ith model, as well as the sample complexity of the associated norm. With $m_{\min}$ as the bottleneck, this result indicates that the combination of norms performs no better than using only one of the norms, even though the target model has a very small degree of freedom.

Different measurement ensembles. Our characterization of recovery failure is easy to interpret and deterministic in nature. We show that it can be used to obtain probabilistic failure results for various random measurement ensembles. In particular, our results hold for measurement matrices with i.i.d subgaussian rows, quadratic measurements and matrix completion type measurements.

Understanding the effect of weighting. We characterize the sample complexity of the multi-objective function as a function of the weights associated with the individual norms. Our upper and lower bounds reveal that the sample complexity of the multi-objective function is related to a certain convex combination of the sample complexities associated with the individual norms. We give formulas for this combination as a function of the weights.

Incorporating general cone constraints. In addition, we can incorporate side information on $\mathbf{x}_0$, expressed as convex cone constraints. This additional information helps in recovery; however, quantifying how much the cone constraints help is not trivial. Our analysis explicitly determines the role of the cone constraint: Geometric properties of the cone such as its Gaussian width determines the constant factors in the bound on the number of measurements.

Sparse and Low-rank matrix recovery: illustrating a gap. As a special case, we consider the recovery of simultaneously sparse and low-rank matrices and prove that there is a significant gap between the performance of convex and non-convex recovery programs. This gap is surprising when one considers similar results in low-dimensional model recovery discussed above in Table 5.1.

5.0.8 Applications

We survey several applications where simultaneous structures arise, as well as existing results specific to these applications. These applications all involve models with simultaneous structures, but the measurement model and the norms that matter differ among applications.

Sparse signal recovery from quadratic measurements. Sparsity has long been exploited in signal processing, applied mathematics, statistics and computer science for tasks such as compression, denoising, model selection, image processing and more. Despite the great interest in exploiting sparsity in various applications, most of the work to date has focused on recovering sparse or low rank data from linear mea-

surements. Recently, the basic sparse recovery problem has been generalized to the case in which the measurements are given by nonlinear transforms of the unknown input, [15]. A special case of this more general setting is quadratic compressed sensing [189] in which the goal is to recover a sparse vector $\mathbf{x}$ from quadratic measurements $b_i = \mathbf{x}^T \mathbf{A}_i \mathbf{x}$. This problem can be linearized by *lifting*, where we wish to recover a "low rank and sparse" matrix $\mathbf{X} = \mathbf{x}\mathbf{x}^T$ subject to measurements $b_i = \langle \mathbf{A}_i, \mathbf{X} \rangle$.

Sparse recovery problems from quadratic measurements arise in a variety of problems in optics. One example is sub-wavelength optical imaging [189, 197] in which the goal is to recover a sparse image from its far-field measurements, where due to the laws of physics the relationship between the (clean) measurement and the unknown image is quadratic. In [189] the quadratic relationship is a result of using partially-incoherent light. The quadratic behavior of the measurements in [197] arises from coherent diffractive imaging in which the image is recovered from its intensity pattern. Under an appropriate experimental setup, this problem amounts to reconstruction of a sparse signal from the magnitude of its Fourier transform.

A related and notable problem involving sparse and low-rank matrices is Sparse Principal Component Analysis (SPCA), mentioned in Section 5.8.

Sparse phase retrieval. Quadratic measurements appear in phase retrieval problems, in which a signal is to be recovered from the magnitude of its measurements $b_i = |\mathbf{a}_i^T \mathbf{x}|$, where each measurement is a linear transform of the input $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{a}_i$'s are arbitrary, possibly complex-valued measurement vectors. An important case is when $\mathbf{a}_i^T \mathbf{x}$ is the Fourier Transform and b_i^2 is the power spectral density. Phase retrieval is of great interest in many applications such as optical imaging [154, 218], crystallography [115], and more [97, 105, 120].

The problem becomes linear when $\mathbf{x}$ is *lifted* and we consider the recovery of $\mathbf{X} = \mathbf{x}\mathbf{x}^T$ where each measurement takes the form $b_i^2 = \langle \mathbf{a}_i \mathbf{a}_i^T, \mathbf{X} \rangle$. In [189], an algorithm was developed to treat phase retrieval problems with sparse $\mathbf{x}$ based on a semidefinite relaxation, and low-rank matrix recovery combined with a row-sparsity constraint on the resulting matrix. More recent works also proposed the use of semidefinite relaxation together with sparsity constraints for phase retrieval [121, 134, 139, 162]. An alternative algorithm was recently designed in [188] based on a greedy search. In [121], the authors also consider sparse signal recovery based on combinatorial and probabilistic approaches and give uniqueness results under certain conditions. Stable uniqueness in phase retrieval problems is studied in [92]. The results of [35, 44] applies to general (non-sparse) signals where in some cases *masked* versions of the signal are required.

Fused lasso. Suppose the signal of interest is sparse and its entries vary slowly, i.e., the signal can be approximated by a piecewise constant function. To encourage sparsity, one can use the ℓ_1 norm, and to encourage the piece-wise constant structure, discrete total variation can be used, defined as

$$\|\mathbf{x}\|_{TV} = \sum_{i=1}^{n-1} |\mathbf{x}_{i+1} - \mathbf{x}_i|.$$

$\|\cdot\|_{TV}$ is basically the ℓ_1 norm of the gradient of the vector; and is approximately sparse. The resulting optimization problem is known as the fused-lasso [202], and is given as

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \lambda \|\mathbf{x}\|_{TV} \quad \text{s.t.} \quad \mathscr{A}(\mathbf{x}) = \mathscr{A}(\mathbf{x}_0). \tag{5.1}$$

To the best of our knowledge, the sample complexity of fused lasso has not been analyzed from a compressed sensing point of view. However, there is a series of recent work on the total variation minimization, which may lead to analysis of (5.1) in the future [160].

We remark that TV regularization is also used together with the nuclear norm to encourage a low-rank and smooth (i.e., slowly varying entries) solution. This regularization finds applications in imaging and physics [110, 187].

Low-rank tensors. Tensors with small Tucker rank can be seen as a generalization of low-rank matrices [209]. In this setup, the signal of interest is the tensor $\mathbf{X}_0 \in \mathbb{R}^{n_1 \times \cdots \times n_t}$, and $\mathbf{X}_0$ is low-rank along its unfoldings which are obtained by reshaping $\mathbf{X}_0$ as a matrix with size $n_i \times \frac{n}{n_i}$, where $n = \prod_{i=1}^{t} n_i$. Denoting the i'th unfolding by $\mathscr{U}_i(\mathbf{X}_0)$, a standard approach to estimate $\mathbf{X}_0$ from $\mathbf{y} = \mathscr{A}(\mathbf{X}_0)$ is minimizing the weighted nuclear norms of the unfoldings,

$$\min_{\mathbf{X}} \sum_{i=1}^{t} \lambda_i \|\mathscr{U}_i(\mathbf{X})\|_\star \quad \text{subject to} \quad \mathbf{y} = \mathscr{A}(\mathbf{X}_0) \tag{5.2}$$

Low-rank tensors have applications in machine learning, physics, computational finance and high dimensional PDE's [114]. (5.2) has been investigated by several papers [103, 137]. Closer to us, [158] recently showed that the convex relaxation (5.2) performs poorly compared to information theoretically optimal bounds for Gaussian measurements. Our results can extend those to the more applicable tensor completion setup, where we observe the entries of the tensor.

Other applications of simultaneously structured signals include Collaborative Hierarchical Sparse Mod-

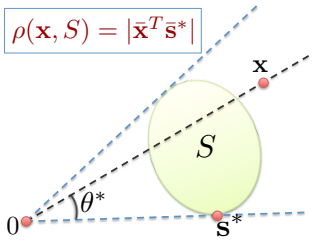


Figure 5.1: Depiction of the correlation between a vector $\mathbf{x}$ and a set S. $\mathbf{s}^*$ achieves the largest angle with $\mathbf{x}$, hence $\mathbf{s}^*$ has the minimum correlation with $\mathbf{x}$.

eling [190] where sparsity is considered within the non-zero blocks in a block-sparse vector, and the recovery of hyperspectral images where we aim to recover a simultaneously block sparse and low rank matrix from compressed observations [109].

5.0.9 Outline of the chapter

The chapter is structured as follows. Background and definitions are given in Section 5.1. An overview of the main results is provided in Section 5.2. Section 5.3 discusses some measurement ensembles for which our results apply. Section 5.4 provides upper bounds for the convex relaxations for the Gaussian measurement ensemble. The proofs of the general results are presented in Section 5.5. The proofs for the special case of simultaneously sparse and low-rank matrices are given in Section 5.6, where we compare corollaries of the general results with the results on non-convex recovery approaches, and illustrate a gap. Numerical simulations in Section 5.7 empirically support the results on sparse and low-rank matrices. Future directions of research and discussion of results are in Section 5.8.

5.1 Problem Setup

We begin by recalling some basic notation. In this chapter, the $\ell_{1,2}$ norm will be the sum of the ℓ_2 norms of the columns of a matrix. With this definition, minimizing the $\ell_{1,2}$ norm will encourage a column-sparse solution, [175, 204]; see section 5.5.4 for more detailed discussion of these norms and their subdifferentials. Overlines denote normalization, i.e., for a vector $\mathbf{x}$ and a matrix $\mathbf{X}$, $\bar{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ and $\bar{\mathbf{X}} = \frac{\mathbf{X}}{\|\mathbf{X}\|_F}$. The set of $n \times n$ positive semidefinite (PSD) and symmetric matrices are denoted by $\mathbb{S}^n_+$ and $\mathbb{S}^n$ respectively. $\mathscr{A}(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is a linear measurement operator if $\mathscr{A}(\mathbf{x})$ is equivalent to the matrix multiplication $\mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$. If $\mathbf{x}$ is a matrix, $\mathscr{A}(\mathbf{x})$ will be a matrix multiplication with a suitably vectorized $\mathbf{x}$.

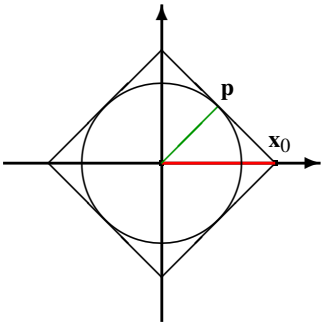


Figure 5.2: Consider the scaled norm ball passing through $\mathbf{x}_0$, then $\kappa = \frac{\|\mathbf{p}\|_2}{\|\mathbf{x}_0\|_2}$, where $\mathbf{p}$ is any of the closest points on the scaled norm ball to the origin.

Definition 5.1.1 (Correlation) *Given a nonzero vector* $\mathbf{x}$ *and a set* S, $\rho(\mathbf{x}, S)$ *is defined as*

$$\rho(\mathbf{x}, S) := \inf_{0 \neq \mathbf{s} \in S} \frac{|\mathbf{x}^T \mathbf{s}|}{\|\mathbf{x}\|_2 \|\mathbf{s}\|_2}.$$

$\rho(\mathbf{x}, S)$ corresponds to the minimum absolute-valued correlation between the vector $\mathbf{x}$ and elements of S. Let $\bar{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. The correlation between $\mathbf{x}$ and the associated subdifferential has a simple form.

$$\rho(\mathbf{x}, \partial\|\mathbf{x}\|) = \inf_{\mathbf{g} \in \partial\|\mathbf{x}\|} \frac{\bar{\mathbf{x}}^T \mathbf{g}}{\|\mathbf{g}\|_2} = \frac{\|\bar{\mathbf{x}}\|}{\sup_{\mathbf{g} \in \partial\|\mathbf{x}\|} \|\mathbf{g}\|_2}.$$

Here, we used the fact that, for norms, subgradients $\mathbf{g} \in \partial\|\mathbf{x}\|$ satisfy $\mathbf{x}^T \mathbf{g} = \|\mathbf{x}\|$, [220]. The denominator of the right hand side is the local Lipschitz constant of $\|\cdot\|$ at $\mathbf{x}$ and is upper bounded by the Lipschitz constant L of $\|\cdot\|$. Consequently, $\rho(\mathbf{x}, \partial\|\mathbf{x}\|) \geq \frac{\|\bar{\mathbf{x}}\|}{L}$. We will denote $\frac{\|\bar{\mathbf{x}}\|}{L}$ by κ. Recently, this quantity has been studied by Mu et al. to analyze the simultaneously structured signals in a similar spirit to us for Gaussian measurements [158][1]. Similar calculations as above gives an alternative interpretation for κ which is illustrated in Figure 5.2.

κ is a measure of alignment between the vector $\mathbf{x}$ and the subdifferential. For the norms of interest, it is associated with the model complexity. For instance, for a k-sparse vector $\mathbf{x}$, $\|\bar{\mathbf{x}}\|_1$ lies between 1 and $\sqrt{k}$ depending on how spiky nonzero entries are. Also $L = \sqrt{n}$. When nonzero entries are ± 1, we find $\kappa^2 = \frac{k}{n}$. Similarly, given a $d \times d$, rank r matrix $\mathbf{X}$, $\|\bar{\mathbf{X}}\|_\star$ lies between 1 and $\sqrt{r}$. If the singular values are spread (i.e.

[1]This chapter is based on the author's work [168]. The work [158] is submitted after initial submission of [168]; which was projecting the subdifferential onto a carefully chosen subspace to obtain bounds on the sample complexity (see Proposition 5.5.1). Inspired from [158], projection onto $\mathbf{x}_0$ and the use of κ led to the simplification of the notation and improvement of the results in [168], in particular, Section 5.3.

± 1), we find $\kappa^2 = \frac{r}{d} = \frac{rd}{d^2}$. In these cases, κ^2 is proportional to the model complexity normalized by the ambient dimension.

Simultaneously structured models. We consider a signal $\mathbf{x}_0$ which has several low-dimensional structures $S_1, S_2, \ldots, S_\tau$ (e.g., sparsity, group sparsity, low-rank). Suppose each structure i corresponds to a norm denoted by $\|\cdot\|_{(i)}$ which promotes that structure (e.g., ℓ_1, $\ell_{1,2}$, nuclear norm). We refer to such an $\mathbf{x}_0$ as a *simultaneously structured model*.

5.1.1 Convex recovery program

We investigate the recovery of the simultaneously structured $\mathbf{x}_0$ from its linear measurements $\mathscr{A}(\mathbf{x}_0)$. To recover $\mathbf{x}_0$, we would like to simultaneously minimize the norms $\|\cdot\|_{(i)}$, $i = 1, \ldots, \tau$, which leads to a multi-objective (vector-valued) optimization problem. For all feasible points $\mathbf{x}$ satisfying $\mathscr{A}(\mathbf{x}) = \mathscr{A}(\mathbf{x}_0)$ and side information $\mathbf{x} \in \mathscr{C}$, consider the set of achievable norms $\{\|\mathbf{x}\|_{(i)}\}_{i=1}^{\tau}$ denoted as points in $\mathbb{R}^\tau$. The minimal points of this set with respect to the positive orthant $\mathbb{R}_+^\tau$ form the *Pareto-optimal* front, as illustrated in Figure 5.3. Since the problem is convex, one can alternatively consider the set

$$\{\mathbf{v} \in \mathbb{R}^\tau : \exists \mathbf{x} \in \mathbb{R}^n \text{ s.t. } \mathbf{x} \in \mathscr{C}, \mathscr{A}(\mathbf{x}) = \mathscr{A}(\mathbf{x}_0), v_i \geq \|\mathbf{x}\|_{(i)}, \text{ for } i = 1, \ldots, \tau\},$$

which is convex and has the same Pareto optimal points as the original set (see, e.g., [23, Chapter 4]).

Definition 5.1.2 (Recoverability) *We call $\mathbf{x}_0$ recoverable if it is a Pareto optimal point; i.e., there does not exist a feasible $\mathbf{x}' \neq \mathbf{x}$ satisfying $\mathscr{A}(\mathbf{x}') = \mathscr{A}(\mathbf{x}_0)$ and $\mathbf{x}' \in \mathscr{C}$, with $\|\mathbf{x}'\|_{(i)} \leq \|\mathbf{x}_0\|_{(i)}$ for $i = 1, \ldots, \tau$.*

The vector-valued convex recovery program can be turned into a scalar optimization problem as

$$\begin{aligned} \underset{\mathbf{x} \in \mathscr{C}}{\text{minimize}} \quad & f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)}) \\ \text{subject to} \quad & \mathscr{A}(\mathbf{x}) = \mathscr{A}(\mathbf{x}_0), \end{aligned} \tag{5.3}$$

where $h : \mathbb{R}_+^\tau \to \mathbb{R}_+$ is convex and non-decreasing in each argument (i.e., non-decreasing and strictly increasing in at least one coordinate). For convex problems with strong duality, it is known that we can recover all of the Pareto optimal points by optimizing weighted sums $f(\mathbf{x}) = \sum_{i=1}^{\tau} \lambda_i \|\mathbf{x}\|_{(i)}$, with positive weights λ_i, among all possible functions $f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)})$. For each $\mathbf{x}_0$ on the Pareto, the coefficients of such recovering function are given by the hyperplane supporting the Pareto at $\mathbf{x}_0$ [23, Chapter 4].

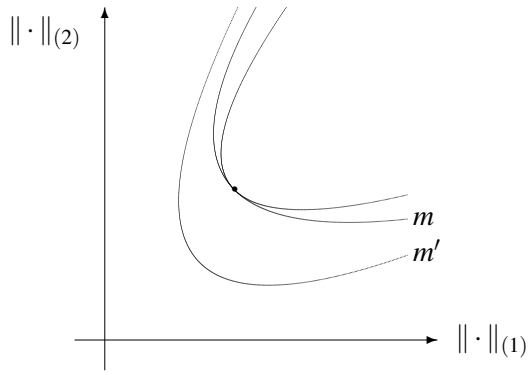


Figure 5.3: Suppose $\mathbf{x}_0$ corresponds to the point shown with a dot. We need at least m measurements for $\mathbf{x}_0$ to be recoverable since for any $m' < m$ this point is not on the Pareto optimal front.

In Figure 5.3, consider the smallest m that makes $\mathbf{x}_0$ recoverable. Then one can choose a function h and recover $\mathbf{x}_0$ by (5.3) using the m measurements. If the number of measurements is any less, then *no* function can recover $\mathbf{x}_0$. Our goal is to provide lower bounds on m.

Note that in [50], Chandrasekaran et al. propose a general theory for constructing a suitable penalty, called an *atomic norm*, given a single set of atoms that describes the structure of the target object. In the case of simultaneous structures, this construction requires defining new atoms, and then ensuring the resulting atomic norm can be minimized in a computationally tractable way, which is nontrivial and often intractable. We briefly discuss such constructions as a future research direction in Section 5.8.

5.2 Main Results: Theorem Statements

In this section, we state our main theorems that aim to characterize the number of measurements needed to recover a simultaneously structured signal by convex or nonconvex programs. We first present our general results, followed by results for simultaneously sparse and low-rank matrices as a specific but important instance of the general case. The proofs are given in Sections 5.5 and 5.6. All of our statements will implicitly assume $\mathbf{x}_0 \neq 0$. This will ensure that $\mathbf{x}_0$ is not a trivial minimizer and 0 is not in the subdifferentials.

5.2.1 General simultaneously structured signals

This section deals with the recovery of a signal $\mathbf{x}_0$ that is simultaneously structured with $S_1, S_2, \ldots, S_\tau$ as described in Section 5.1. We give a lower bound on the required number of measurements, using the geometric properties of the individual norms.

Theorem 5.2.1 (Deterministic failure) *Suppose* $\mathscr{C} = \mathbb{R}^n$ *and,*

$$\rho(\mathbf{x}_0, \partial f(\mathbf{x}_0)) := \inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} |\bar{\mathbf{g}}^T \bar{\mathbf{x}}_0| > \frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2}{\sigma_{\min}(\mathbf{A}^T)}. \tag{5.4}$$

Then, $\mathbf{x}_0$ *is not a minimizer of* (5.3).

Theorem 5.2.1 is deterministic in nature. However, it can be easily specialized to specific random measurement ensembles. The left hand side of (5.4) depends only on the vector $\mathbf{x}_0$ and the subdifferential $\partial f(\mathbf{x}_0)$, hence it is independent of the measurement matrix $\mathbf{A}$. For simultaneously structured models, we will argue that, the left hand side cannot be made too small, as the subgradients are *aligned* with the signal. On the other hand, the right hand side depends only on $\mathbf{A}$ and $\mathbf{x}_0$ and is independent of the subdifferential. In linear inverse problems, $\mathbf{A}$ is often assumed to be random. For large class of random matrices, we will argue that, the right hand side is approximately $\sim \sqrt{\frac{m}{n}}$ which will yield a lower bound on the number of required measurements.

Typical measurement ensembles include the following,

- **Sampling entries:** In low-rank matrix and tensor completion problems, we observe the entries of $\mathbf{x}_0$ uniformly at random. In this case, rows of $\mathbf{A}$ are chosen from the standard basis in $\mathbb{R}^n$. We should remark that, instead of the standard basis, one can consider other orthonormal bases such as the Fourier basis.
- **Matrices with i.i.d. rows:** $\mathbf{A}$ has independent and identically distributed rows with certain moment conditions. This is a widely used setup in compressed sensing as each measurement we make is associated with the corresponding row of $\mathbf{A}$ [38].
- **Quadratic measurements:** Arises in the phase retrieval problem as discussed in Section 5.0.8.

In Section 5.3, we find upper bounds on the right hand side of (5.4) for these ensembles. As it will be discussed in Section 5.3, we can do modifications in the rows of $\mathbf{A}$ to get better bounds as long as it does not affect its null space. For instance, one can discard the identical rows to improve conditioning. However, as m increases and $\mathbf{A}$ has more linearly independent rows, $\sigma_{\min}(\mathbf{A}^T)$ will naturally decrease and (5.4) will no longer hold after a certain point. In particular, (5.4) cannot hold beyond $m \geq n$ as $\sigma_{\min}(\mathbf{A}^T) = 0$. This is indeed natural as the system becomes overdetermined.

The following proposition lower bounds the left hand side of (5.4) in an interpretable manner. In particular, the correlation $\rho(\mathbf{x}_0, \partial f(\mathbf{x}_0))$ can be lower bounded by the smallest individual correlation.

Proposition 5.2.1 *Let L_i be the Lipschitz constant of the i'th norm and $\kappa_i = \frac{\|\bar{\mathbf{x}}_0\|_{(i)}}{L_i}$ for $1 \leq i \leq \tau$. Set $\kappa_{\min} = \min\{\kappa_i : i = 1, \ldots, \tau\}$. We have the following,*

- *All functions $f(\cdot)$ in (5.3) satisfy, $\rho(\mathbf{x}_0, \partial f(\mathbf{x}_0)) \geq \kappa_{\min}$[2].*
- *Suppose $f(\cdot)$ is a weighted linear combination $f(\mathbf{x}) = \sum_{i=1}^{\tau} \lambda_i \|\mathbf{x}\|_{(i)}$ for nonnegative $\{\lambda_i\}_{i=1}^{\tau}$. Let $\bar{\lambda}_i = \frac{\lambda_i L_i}{\sum_{i=1}^{\tau} \lambda_i L_i}$ for $1 \leq i \leq \tau$. Then, $\rho(\mathbf{x}_0, \partial f(\mathbf{x}_0)) \geq \sum_{i=1}^{\tau} \bar{\lambda}_i \kappa_i$.*

Proof: From Lemma 5.5.3, any subgradient of $f(\cdot)$ can be written as, $\mathbf{g} = \sum_{i=1}^{\tau} w_i \mathbf{g}_i$ for some nonnegative w_i's. On the other hand, from [220], $\langle \bar{\mathbf{x}}_0, \mathbf{g}_i \rangle = \|\bar{\mathbf{x}}_0\|_{(i)}$. Combining, we find,

$$\mathbf{g}^T \bar{\mathbf{x}}_0 = \sum_{i=1}^{\tau} w_i \|\bar{\mathbf{x}}_0\|_{(i)}.$$

From triangle inequality, $\|\mathbf{g}\|_2 \leq \sum_{i=1}^{\tau} w_i L_i$. To conclude, we use,

$$\frac{\sum_{i=1}^{\tau} w_i \|\bar{\mathbf{x}}_0\|_{(i)}}{\sum_{i=1}^{\tau} w_i L_i} \geq \min_{1 \leq i \leq \tau} \frac{w_i \|\bar{\mathbf{x}}_0\|_{(i)}}{w_i L_i} = \kappa_{\min}. \tag{5.5}$$

For the second part, we use the fact that for the weighted sums of norms, $w_i = \lambda_i$ and subgradients has the form $\mathbf{g} = \sum_{i=1}^{\tau} \lambda_i \mathbf{g}_i$, [23]. Then, substitute $\bar{\lambda}_i$ for λ_i on the left hand side of (5.5).

■

Before stating the next result, let us recall the Gaussian distance definition from Chapter 2, which will be useful throughout.

Definition 5.2.1 (Gaussian squared-distance) *Let $\mathscr{M}$ be a closed convex set in $\mathbb{R}^n$ and let $\mathbf{h} \in \mathbb{R}^n$ be a vector with independent standard normal entries. Then, the Gaussian distance of $\mathscr{M}$ is defined as*

$$\mathbf{D}(\mathscr{M}) = \mathbb{E}[\inf_{\mathbf{v} \in \mathscr{M}} \|\mathbf{h} - \mathbf{v}\|_2^2]$$

For notational simplicity, let the normalized distance be $\bar{\mathbf{D}}(\mathscr{M}) = \frac{\mathbf{D}(\mathscr{M})}{n}$.

We will now state our result for Gaussian measurements; which can additionally include cone constraints for the lower bound. One can obtain results for the other ensembles by referring to Section 5.3.

Theorem 5.2.2 (Gaussian lower bound) *Suppose $\mathbf{A}$ has independent $\mathscr{N}(0,1)$ entries. Whenever $m \leq m_{low}$, $\mathbf{x}_0$ will not be a minimizer of any of the recovery programs in (5.3) with probability at least $1 -$*

[2]The lower bound $\kappa_{\min}$ is directly comparable to Theorem 5 of [158]. Indeed, our lower bounds on the sample complexity will have the form $\mathscr{O}\left(\kappa_{\min}^2 n\right)$.

$10\exp(-\frac{1}{16}\min\{m_{low}, (1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})})^2 n\})$, *where*

$$m_{low} \triangleq \frac{(1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})})n\kappa_{\min}^2}{100}.$$

Remark: When $\mathscr{C} = \mathbb{R}^n$, $\bar{\mathbf{D}}(\mathscr{C}) = 0$ hence, the lower bound simplifies to $m_{low} = \frac{n\kappa_{\min}^2}{100}$.

Here $\bar{\mathbf{D}}(\mathscr{C})$ depends only on $\mathscr{C}$ and can be viewed as a constant. For instance, for the positive semidefinite cone, we show $\bar{\mathbf{D}}(\mathbb{S}_+^n) < \frac{3}{4}$. Observe that for a smaller cone $\mathscr{C}$, it is reasonable to expect a smaller lower bound to the required number of measurements. Indeed, as $\mathscr{C}$ gets smaller, $\mathbf{D}(\mathscr{C})$ gets larger.

As discussed before, there are various options for the scalarizing function in (5.3), with one choice being the weighted sum of norms. In fact, for a recoverable point $\mathbf{x}_0$ there always exists a weighted sum of norms which recovers it. This function is also often the choice in applications, where the space of positive weights is searched for a good combination. Thus, we can state the following theorem as a general result.

Corollary 5.2.1 (Weighted lower bound) *Suppose* $\mathbf{A}$ *has i.i.d* $\mathscr{N}(0,1)$ *entries and* $f(\mathbf{x}) = \sum_{i=1}^{\tau} \lambda_i \|\mathbf{x}\|_{(i)}$ *for nonnegative weights* $\{\lambda_i\}_{i=1}^{\tau}$. *Whenever* $m \leq m'_{low}$, $\mathbf{x}_0$ *will not be a minimizer of the recovery program* (5.3) *with probability at least* $1 - 10\exp(-\frac{1}{16}\min\{m'_{low}, (1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})})^2 n\})$, *where*

$$m'_{low} \triangleq \frac{n(1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})})(\sum_{i=1}^{\tau} \bar{\lambda}_i \kappa_i)^2}{100},$$

and $\bar{\lambda}_i = \frac{\lambda_i L_i}{\sum_{i=1}^{\tau} \lambda_i L_i}$.

Observe that Theorem 5.2.2 is stronger than stating "a particular function $h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)})$ will not work". Instead, our result states that with high probability none of the programs in the class (5.3) can return $\mathbf{x}_0$ as the optimal unless the number of measurements are sufficiently large.

To understand the result better, note that the required number of measurements is proportional to $\kappa_{min}^2 n$ which is often proportional to the sample complexity of the best individual norm. As we have argued in Section 5.1, $\kappa_i^2 n$ corresponds to how structured the signal is. For sparse signals it is equal to the sparsity, and for a rank r matrix, it is equal to the degrees of freedom of the set of rank r matrices. Consequently, Theorem 5.2.2 suggests that even if the signal satisfies multiple structures, the required number of measurements is effectively determined by only one dominant structure.

Intuitively, the degrees of freedom of a simultaneously structured signal can be much lower, which is provable for the S&L matrices. Hence, there is a considerable gap between the expected measurements

Model	$f(\cdot)$	L	$\|\bar{\mathbf{x}}_0\| \leq$	$n\kappa^2 \leq$
k sparse vector	$\|\cdot\|_1$	$\sqrt{n}$	$\sqrt{k}$	k
k column-sparse matrix	$\|\cdot\|_{1,2}$	$\sqrt{d}$	$\sqrt{k}$	kd
Rank r matrix	$\|\cdot\|_\star$	$\sqrt{d}$	$\sqrt{r}$	rd
S&L (k,k,r) matrix	$h(\|\cdot\|_\star, \|\cdot\|_1)$	–	–	$\min\{k^2, rd\}$

Table 5.2: Summary of the parameters that are discussed in this section. The last three lines is for a $d \times d$ S&L (k,k,r) matrix where $n = d^2$. In the fourth column, the corresponding entry for S&L is $\kappa_{\min} = \min\{\kappa_{\ell_1}, \kappa_\star\}$.

based on model complexity and the number of measurements needed for recovery via (5.3) ($\kappa_{\min}^2 n$).

5.2.2 Simultaneously Sparse and Low-rank Matrices

We now focus on a special case, namely simultaneously sparse and low-rank (S&L) matrices. We consider matrices with nonzero entries contained in a small submatrix where the submatrix itself is low rank. Here, norms of interest are $\|\cdot\|_{1,2}$, $\|\cdot\|_1$ and $\|\cdot\|_\star$ and the cone of interest is the PSD cone. We also consider nonconvex approaches and contrast the results with convex approaches. For the nonconvex problem, we replace the norms $\|\cdot\|_1, \|\cdot\|_{1,2}, \|\cdot\|_\star$ with the functions $\|\cdot\|_0, \|\cdot\|_{0,2}, \text{rank}(\cdot)$ which give the number of nonzero entries, the number of nonzero columns and rank of a matrix respectively and use the same cone constraint as the convex method. We show that convex methods perform poorly as predicted by the general result in Theorem 5.2.2, while nonconvex methods require optimal number of measurements (up to a logarithmic factor). Proofs are given in Section 5.6.

Definition 5.2.2 *We say $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ is an S&L matrix with (k_1, k_2, r) if the smallest submatrix that contains nonzero entries of $\mathbf{X}_0$ has size $k_1 \times k_2$ and $\text{rank}(\mathbf{X}_0) = r$. When $\mathbf{X}_0$ is symmetric, let $d = d_1 = d_2$ and $k = k_1 = k_2$. We consider the following cases.*

(a) General: $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ *is S&L with* (k_1, k_2, r).

(b) PSD model: $\mathbf{X}_0 \in \mathbb{R}^{n \times n}$ *is PSD and S&L with* (k, k, r).

We are interested in S&L matrices with $k_1 \ll d_1, k_2 \ll d_2$ so that the matrix is sparse, and $r \ll \min\{k_1, k_2\}$ so that the submatrix containing the nonzero entries is low rank. Recall from Section 5.1.1 that our goal is to recover $\mathbf{X}_0$ from linear observations $\mathscr{A}(\mathbf{X}_0)$ via convex or nonconvex optimization programs. The measurements can be equivalently written as $\mathbf{A}\,\text{vec}(\mathbf{X}_0)$, where $\mathbf{A} \in \mathbb{R}^{m \times d_1 d_2}$ and $\text{vec}(\mathbf{X}_0) \in \mathbb{R}^{d_1 d_2}$ denotes the vector obtained by stacking the columns of $\mathbf{X}_0$.

Based on the results in Section 5.2.1, we obtain lower bounds on the number of measurements for convex recovery. We additionally show that significantly fewer measurements are sufficient for non-convex programs to uniquely recover $\mathbf{X}_0$; thus proving a performance gap between convex and nonconvex approaches. The following theorem summarizes the results.

Theorem 5.2.3 (Performance of S&L matrix recovery) *Suppose $\mathscr{A}(\cdot)$ is an i.i.d Gaussian map and consider recovering $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ via*

$$\underset{\mathbf{X} \in \mathscr{C}}{\text{minimize}} \; f(\mathbf{X}) \quad \text{subject to} \quad \mathscr{A}(\mathbf{X}) = \mathscr{A}(\mathbf{X}_0). \tag{5.6}$$

For the cases given in Definition 5.2.2, the following convex and nonconvex recovery results hold for some positive constants c_1, c_2.

(a) *General model:*

(a1) *Let $f(\mathbf{X}) = \|\mathbf{X}\|_{1,2} + \lambda_1 \|\mathbf{X}^T\|_{1,2} + \lambda_2 \|\mathbf{X}\|_\star$ where $\lambda_1, \lambda_2 \geq 0$ and $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$. Then, (5.6) will fail to recover $\mathbf{X}_0$ with probability $1 - \exp(-c_1 m_0)$ whenever $m \leq c_2 m_0$ where $m_0 = \min\{d_1 k_2, d_2 k_1, (d_1 + d_2) r\}$.*

(a2) *Let $f(\mathbf{X}) = \frac{1}{k_2}\|\mathbf{X}\|_{0,2} + \frac{1}{k_1}\|\mathbf{X}^T\|_{0,2} + \frac{1}{r}\operatorname{rank}(\mathbf{X})$ and $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$. Then, (5.6) will uniquely recover $\mathbf{X}_0$ with probability $1 - \exp(-c_1 m)$ whenever $m \geq c_2 \max\{(k_1 + k_2) r, k_1 \log \frac{d_1}{k_1}, k_2 \log \frac{d_2}{k_2}\}$.*

(b) *PSD with $\ell_{1,2}$:*

(b1) *Let $f(\mathbf{X}) = \|\mathbf{X}\|_{1,2} + \lambda \|\mathbf{X}\|_\star$ where $\lambda \geq 0$ and $\mathscr{C} = \mathbb{S}^d_+$. Then, (5.6) will fail to recover $\mathbf{X}_0$ with probability $1 - \exp(-c_1 r d)$ whenever $m \leq c_2 r d$.*

(b2) *Let $f(\mathbf{X}) = \frac{2}{k}\|\mathbf{X}\|_{0,2} + \frac{1}{r}\operatorname{rank}(\mathbf{X})$ and $\mathscr{C} = \mathbb{S}^d$. Then, (5.6) will uniquely recover $\mathbf{X}_0$ with probability $1 - \exp(-c_1 m)$ whenever $m \geq c_2 \max\{rk, k \log \frac{d}{k}\}$.*

(c) *PSD with ℓ_1:*

(c1) *Let $f(\mathbf{X}) = \|\mathbf{X}\|_1 + \lambda \|\mathbf{X}\|_\star$ and $\mathscr{C} = \mathbb{S}^d_+$. Then, (5.6) will fail to recover $\mathbf{X}_0$ with probability $1 - \exp(-c_1 m_0)$ for all possible $\lambda \geq 0$ whenever $m \leq c_2 m_0$ where $m_0 = \min\{\|\bar{\mathbf{X}}_0\|_1^2, \|\bar{\mathbf{X}}_0\|_\star^2 d\}$.*

(c2) *Suppose $\operatorname{rank}(\mathbf{X}_0) = 1$. Let $f(\mathbf{X}) = \frac{1}{k^2}\|\mathbf{X}\|_0 + \operatorname{rank}(\mathbf{X})$ and $\mathscr{C} = \mathbb{S}^d$. Then, (5.6) will uniquely recover $\mathbf{X}_0$ with probability $1 - \exp(-c_1 m)$ whenever $m \geq c_2 k \log \frac{d}{k}$.*

Remark on "PSD with ℓ_1": In the special case, $\mathbf{X}_0 = \mathbf{a}\mathbf{a}^T$ for a k-sparse vector $\mathbf{a}$, we have $m_0 = \min\{\|\bar{\mathbf{a}}\|_1^4, d\}$. When nonzero entries of $\mathbf{a}$ are ± 1, we have $m_0 = \min\{k^2, d\}$.

Setting	Nonconvex sufficient m	Convex required m
General model	$\mathcal{O}\left(\max\{rk, k\log\frac{d}{k}\}\right)$	$\Omega(rd)$
PSD with $\ell_{1,2}$	$\mathcal{O}\left(\max\{rk, k\log\frac{d}{k}\}\right)$	$\Omega(rd)$
PSD with ℓ_1	$\mathcal{O}\left(k\log\frac{d}{k}\right)$	$\Omega(\min\{k^2, rd\})$

Table 5.3: Summary of recovery results for models in Definition 5.2.2, assuming $d_1 = d_2 = d$ and $k_1 = k_2 = k$. For the PSD with ℓ_1 case, we assume $\frac{\|\bar{\mathbf{x}}_0\|_1}{k}$ and $\frac{\|\bar{\mathbf{x}}_0\|_\star}{\sqrt{r}}$ to be approximately constants for the sake of simplicity. Nonconvex approaches are optimal up to a logarithmic factor, while convex approaches perform poorly.

The nonconvex programs require almost the same number of measurements as the degrees of freedom (or number of parameters) of the underlying model. For instance, it is known that the degrees of freedom of a rank r matrix of size $k_1 \times k_2$ is simply $r(k_1 + k_2 - r)$ which is $\mathcal{O}\left((k_1 + k_2)r\right)$. Hence, the nonconvex results are optimal up to a logarithmic factor. On the other hand, our results on the convex programs that follow from Theorem 5.2.2 show that the required number of measurements are significantly larger. Table 5.3 provides a quick comparison of the results on S&L.

For the S&L (k,k,r) model, from standard results one can easily deduce that [43, 178, 195],

- ℓ_1 penalty only: requires at least k^2,
- $\ell_{1,2}$ penalty only: requires at least kd,
- Nuclear norm penalty only: requires at least rd measurements.

These follow from the model complexity of the sparse, column-sparse and low-rank matrices. Theorem 5.2.2 shows that, combination of norms require at least as much as the best individual norm. For instance, combination of ℓ_1 and the nuclear norm penalization yields the lower bound $\mathcal{O}\left(\min\{k^2, rd\}\right)$ for S&L matrices whose singular values and nonzero entries are spread. This is indeed what we would expect from the interpretation that $\kappa^2 n$ is often proportional to the sample complexity of the corresponding norm and, the lower bound $\kappa_{\min}^2 n$ is proportional to that of the best individual norm.

As we saw in Section 5.2.1, adding a cone constraint to the recovery program does not help in reducing the lower bound by more than a constant factor. In particular, we discuss the positive semidefiniteness assumption that is beneficial in the sparse phase retrieval problem,, and show that the number of measurements remain high even when we include this extra information. On the other hand, the nonconvex recovery programs performs well even without the PSD constraint.

We remark that, we could have stated Theorem 5.2.3 for more general measurements given in Section 5.3 without the cone constraint. For instance, the following result holds for the weighted linear combination of individual norms and for the subgaussian ensemble.

Corollary 5.2.2 *Suppose $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$ obeys the general model with $k_1 = k_2 = k$ and $\mathscr{A}$ is a linear sub-gaussian map as described in Proposition 5.3.1. Choose $f(\mathbf{X}) = \lambda_{\ell_1}\|\mathbf{X}\|_1 + \lambda_\star\|\mathbf{X}\|_\star$, where $\lambda_{\ell_1} = \beta$, $\lambda_\star = (1-\beta)\sqrt{d}$ and $0 \le \beta \le 1$. Then, whenever, $m \le \min\{m_{low}, c_1 n\}$, where,*

$$m_{low} = \frac{(\beta\|\bar{\mathbf{X}}_0\|_1 + (1-\beta)\|\bar{\mathbf{X}}_0\|_\star\sqrt{d})^2}{2},$$

(5.6) fails with probability $1 - 4\exp(-c_2 m_{low})$. Here $c_1, c_2 > 0$ are constants as described in Proposition 5.3.1.

Remark: Choosing $\mathbf{X}_0 = \mathbf{a}\mathbf{a}^T$ where nonzero entries of $\mathbf{a}$ are ± 1 yields $\frac{1}{2}(\beta k + (1-\beta)\sqrt{d})^2$ on the right hand side. An explicit construction of an S&L matrix with maximal $\|\bar{\mathbf{X}}\|_1, \|\bar{\mathbf{X}}\|_\star$ is provided in Section 5.6.3.

This corollary compares well with the upper bound obtained in Corollary 5.4.1 of Section 5.4. In particular, both the bounds and the penalty parameters match up to logarithmic factors. Hence, together, they sandwich the sample complexity of the combined cost $f(\mathbf{X})$.

5.3 Measurement ensembles

This section will make use of standard results on sub-gaussian random variables and random matrix theory to obtain probabilistic statements. We will explain how one can analyze the right hand side of (5.4) for,

- Matrices with sub-gaussian rows,
- Subsampled standard basis (in matrix completion),
- Quadratic measurements arising in phase retrieval.

5.3.1 Sub-gaussian measurements

We first consider the measurement maps with sub-gaussian entries. The following definitions are borrowed from [213].

Definition 5.3.1 (Sub-gaussian random variable) *A random variable x is sub-gaussian if there exists a constant $K > 0$ such that for all $p \ge 1$,*

$$(\mathbb{E}\,|x|^p)^{1/p} \le K\sqrt{p}.$$

The smallest such K is called the sub-gaussian norm of x and is denoted by $\|x\|_{\Psi_2}$. A sub-exponential random variable y is one for which there exists a constant K' such that, $\mathbb{P}(|y| > t) \le \exp(1 - \frac{t}{K'})$. x is sub-gaussian

if and only if x^2 is sub-exponential.

Definition 5.3.2 (Isotropic sub-gaussian vector) *A random vector $\mathbf{x} \in \mathbb{R}^n$ is sub-gaussian if the one dimensional marginals $\mathbf{x}^T\mathbf{v}$ are sub-gaussian random variables for all $\mathbf{v} \in \mathbb{R}^n$. The sub-gaussian norm of $\mathbf{x}$ is defined as,*

$$\|\mathbf{x}\|_{\Psi_2} = \sup_{\|\mathbf{v}\|=1} \|\mathbf{x}^T\mathbf{v}\|_{\Psi_2}$$

$\mathbf{x}$ is also isotropic, if its covariance is equal to identity, i.e. $\mathbb{E}\,\mathbf{x}\mathbf{x}^T = \mathbf{I}_n$.

Proposition 5.3.1 (Sub-gaussian measurements) *Suppose $\mathbf{A}$ has i.i.d rows in either of the following forms,*

- *a copy of a zero-mean isotropic sub-gaussian vector $\mathbf{a} \in \mathbb{R}^n$, where $\|\mathbf{a}\|_2 = \sqrt{n}$ almost surely.*
- *have independent zero-mean unit variance sub-gaussian entries.*

Then, there exists constants c_1, c_2 depending only on the sub-gaussian norm of the rows, such that, whenever $m \leq c_1 n$, with probability $1 - 4\exp(-c_2 m)$, we have,

$$\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}^2(\mathbf{A}^T)} \leq \frac{2m}{n}$$

Proof: Using Theorem 5.58 of [213], there exists constants c, C depending only on the sub-gaussian norm of $\mathbf{a}$ such that for any $t \geq 0$, with probability $1 - 2\exp(-ct^2)$

$$\sigma_{\min}(\mathbf{A}^T) \geq \sqrt{n} - C\sqrt{m} - t$$

Choosing $t = C\sqrt{m}$ and $m \leq \frac{n}{100C^2}$ would ensure $\sigma_{\min}(\mathbf{A}^T) \geq \frac{4\sqrt{n}}{5}$.

Next, we shall estimate $\|\mathbf{A}\bar{\mathbf{x}}_0\|_2$. $\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2$ is sum of i.i.d. sub-exponential random variables identical to $|\mathbf{a}^T\bar{\mathbf{x}}_0|^2$. Also, $\mathbb{E}[|\mathbf{a}^T\bar{\mathbf{x}}_0|^2] = 1$. Hence, Proposition 5.16 of [213] gives,

$$\mathbb{P}(\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2 \geq m + t) \leq 2\exp(-c' \min\{\frac{t^2}{m}, t\})$$

Choosing $t = \frac{7m}{25}$, we find that $\mathbb{P}(\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2 \geq \frac{32m}{25}) \leq 2\exp(-c''m)$. Combining the two, we obtain,

$$\mathbb{P}(\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}^2(\mathbf{A}^T)} \leq \frac{2m}{n}) \geq 1 - 4\exp(-c'''m)$$

The second statement can be proved in the exact same manner by using Theorem 5.39 of [213] instead of Theorem 5.58. ■

Remark: While Proposition 5.3.1 assumes $\mathbf{a}$ has fixed ℓ_2 norm, this can be ensured by properly normalizing rows of $\mathbf{A}$ (assuming they stay sub-gaussian). For instance, if the ℓ_2 norm of the rows are larger than $c\sqrt{n}$ for a positive constant c, normalization will not affect sub-gaussianity. Note that, scaling rows of a matrix do not change its null space.

5.3.2 Randomly sampling entries

We now consider the scenario where each row of $\mathbf{A}$ is chosen from the standard basis uniformly at random. Note that, when m is comparable to n, there is a nonnegligible probability that $\mathbf{A}$ will have duplicate rows. Theorem 5.2.1 does not take this situation into account which would make $\sigma_{\min}(\mathbf{A}^T) = 0$. In this case, one can discard the copies as they don't affect the recoverability of $\mathbf{x}_0$. This would get rid of the ill-conditioning, as the new matrix is well-conditioned with the exact same null space as the original, and would correspond to a "sampling without replacement" scheme where we ensure each row is different.

Similar to achievability results in matrix completion [40], the following failure result requires true signal to be incoherent with the standard basis, where incoherence is characterized by $\|\bar{\mathbf{x}}_0\|_\infty$, which lies between $\frac{1}{\sqrt{n}}$ and 1.

Proposition 5.3.2 (Sampling entries) *Let $\{\mathbf{e}_i\}_{i=1}^n$ be the standard basis in $\mathbb{R}^n$ and suppose each row of $\mathbf{A}$ is chosen from $\{\mathbf{e}_i\}_{i=1}^n$ uniformly at random. Let $\hat{\mathbf{A}}$ be the matrix obtained by removing the duplicate rows in $\mathbf{A}$. Then, with probability $1-\exp(-\frac{m}{4n\|\bar{\mathbf{x}}_0\|_\infty^2})$, we have,*

$$\frac{\|\hat{\mathbf{A}}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}^2(\hat{\mathbf{A}})} \leq \frac{2m}{n}$$

Proof: Let $\hat{\mathbf{A}}$ be the matrix obtained by discarding the rows of $\mathbf{A}$ that occur multiple times except one of them. Clearly $\text{Null}(\hat{\mathbf{A}}) = \text{Null}(\mathbf{A})$ hence they are equivalent for the purpose of recovering $\mathbf{x}_0$. Furthermore, $\sigma_{\min}(\hat{\mathbf{A}}) = 1$. Hence, we are interested in upper bounding $\|\hat{\mathbf{A}}\bar{\mathbf{x}}_0\|_2$.

Clearly $\|\hat{\mathbf{A}}\bar{\mathbf{x}}_0\|_2 \leq \|\mathbf{A}\bar{\mathbf{x}}_0\|_2$. Hence, we will bound $\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2$ probabilistically. Let $\mathbf{a}$ be the first row of $\mathbf{A}$. $|\mathbf{a}^T\bar{\mathbf{x}}_0|^2$ is a random variable, with mean $\frac{1}{n}$ and is upper bounded by $\|\bar{\mathbf{x}}_0\|_\infty^2$. Hence, applying the Chernoff Bound would yield,

$$\mathbb{P}(\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2 \geq \frac{m}{n}(1+\delta)) \leq \exp(-\frac{m\delta^2}{2(1+\delta)n\|\bar{\mathbf{x}}_0\|_\infty^2})$$

Setting $\delta = 1$, we find that, with probability $1 - \exp(-\frac{m}{4n\|\bar{\mathbf{x}}_0\|_\infty^2})$, we have,

$$\frac{\|\hat{\mathbf{A}}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}(\hat{\mathbf{A}})^2} \leq \frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}(\hat{\mathbf{A}})^2} \leq \frac{2m}{n}$$

■

A significant application of this result would be for the low-rank tensor completion problem, where we randomly observe some entries of a low-rank tensor and try to reconstruct it. A promising approach for this problem is using the weighted linear combinations of nuclear norms of the unfoldings of the tensor to induce the low-rank tensor structure described in (5.2), [103, 114]. Related work [158] shows the poor performance of (5.2) for the special case of Gaussian measurements. Combination of Theorem 5.2.1 and Proposition 5.3.2 will immediately extend the results of [158] to the more applicable tensor completion setup (under proper incoherence conditions that bound $\|\bar{\mathbf{x}}_0\|_\infty$).

Remark: In Propositions 5.3.1 and 5.3.2, we can make the upper bound for the ratio $\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{\sigma_{\min}(\mathbf{A})^2}$ arbitrarily close to $\frac{m}{n}$ by changing the proof parameters. Combined with Proposition 5.2.1, this would suggest that, failure happens, when $m < n\kappa_{\min}$.

5.3.3 Quadratic measurements

As mentioned in the phase retrieval problem, quadratic measurements $|\mathbf{v}^T\mathbf{a}|^2$ of the vector $\mathbf{a} \in \mathbb{R}^d$ can be linearized by the change of variable $\mathbf{a} \to \mathbf{X}_0 = \mathbf{a}\mathbf{a}^T$ and using $\mathbf{V} = \mathbf{v}\mathbf{v}^T$. The following proposition can be used to obtain a lower bound for such ensembles when combined with Theorem 5.2.1.

Proposition 5.3.3 *Suppose we observe quadratic measurements $\mathscr{A}(\mathbf{X}_0) \in \mathbb{R}^m$ of a matrix $\mathbf{X}_0 = \mathbf{a}\mathbf{a}^T \in \mathbb{R}^{d\times d}$. Here, assume that i'th entry of $\mathscr{A}(\mathbf{X}_0)$ is equal to $|\mathbf{v}_i^T\mathbf{a}|^2$ where $\{\mathbf{v}_i\}_{i=1}^m$ are independent vectors, either with $\mathcal{N}(0,1)$ entries or are uniformly distributed over the sphere with radius $\sqrt{d}$. Then, there exists absolute constants $c_1, c_2 > 0$ such that whenever $m < \frac{c_1 d}{\log d}$, with probability $1 - 2ed^{-2}$,*

$$\frac{\|\mathscr{A}(\bar{\mathbf{X}}_0)\|_2}{\sigma_{\min}(\mathbf{A}^T)} \leq \frac{c_2\sqrt{m}\log d}{d}$$

Proof: Let $\mathbf{V}_i = \mathbf{v}_i \mathbf{v}_i^T$. Without loss of generality, assume $\mathbf{v}_i$'s are uniformly distributed over sphere with radius $\sqrt{d}$. To lower bound $\sigma_{\min}(\mathbf{A}^T)$, we will estimate the coherence of its columns, defined by,

$$\mu(\mathbf{A}^T) = \max_{i \neq j} \frac{|\langle \mathbf{V}_i, \mathbf{V}_j \rangle|}{\|\mathbf{V}_i\|_F \|\mathbf{V}_j\|_F} = \frac{(\mathbf{v}_i^T \mathbf{v}_j)^2}{d^2}$$

Section 5.2.5 of [213] states that sub-gaussian norm of $\mathbf{v}_i$ is bounded by an absolute constant. Hence, conditioned on $\mathbf{v}_j$ (which satisfies $\|\mathbf{v}_j\|_2 = \sqrt{d}$), $\frac{(\mathbf{v}_i^T \mathbf{v}_j)^2}{d}$ is a subexponential random variable with mean 1. Hence, using Definition 5.3.1, there exists a constant $c > 0$ such that,

$$\mathbb{P}\left(\frac{(\mathbf{v}_i^T \mathbf{v}_j)^2}{d} > c \log d\right) \leq e d^{-4}$$

Union bounding over all i, j pairs ensure that with probability ed^{-2} we have $\mu(\mathbf{A}^T) \leq c\frac{\log d}{d}$. Next, we use the standard result that for a matrix with columns of equal length, $\sigma_{\min}(\mathbf{A}^T) \geq d(1-(m-1)\mu)$. The reader is referred to Proposition 1 of [205]. Hence, $m \leq \frac{d}{2c \log d}$, gives $\sigma_{\min}(\mathbf{A}^T) \geq \frac{d}{2}$.

It remains to upper bound $\|\mathscr{A}(\bar{\mathbf{X}}_0)\|_2$. The i'th entry of $\mathscr{A}(\bar{\mathbf{X}}_0)$ is equal to $|\mathbf{v}_i^T \bar{\mathbf{a}}|^2$, hence it is subexponential. Consequently, there exists a constant c' so that each entry is upper bounded by $\frac{c'}{2} \log d$ with probability $1 - ed^{-3}$. Union bounding, and using $m \leq d$, we find that $\|\mathscr{A}(\bar{\mathbf{X}}_0)\|_2 \leq \frac{c'}{2}\sqrt{m} \log d$ with probability $1 - ed^{-2}$. Combining with the $\sigma_{\min}(\mathbf{A}^T)$ estimate we can conclude. ■

Comparison to existing literature. Proposition 5.3.3 is useful to estimate the performance of the sparse phase retrieval problem, in which $\mathbf{a}$ is a k sparse vector, and we minimize a combination of the ℓ_1 norm and the nuclear norm to recover $\mathbf{X}_0$. Combined with Theorem 5.2.1, Proposition 5.3.3 gives that, whenever $m \leq \frac{c_1 d}{\log d}$ and $\frac{c_2 \sqrt{m} \log d}{d} \leq \min\{\frac{\|\bar{\mathbf{X}}_0\|_1}{d}, \frac{\|\bar{\mathbf{X}}_0\|_\star}{\sqrt{d}}\}$, the recovery fails with high probability. Since $\|\bar{\mathbf{X}}_0\|_\star = 1$ and $\|\bar{\mathbf{X}}_0\|_1 = \|\bar{\mathbf{a}}\|_1^2$, the failure condition reduces to,

$$m \leq \frac{c}{\log^2 d} \min\{\|\bar{\mathbf{a}}\|_1^4, d\}.$$

When $\bar{\mathbf{a}}$ is a k-sparse vector with ± 1 entries, in a similar flavor to Theorem 5.2.3, the right hand side has the form $\frac{c}{\log^2 d} \min\{k^2, d\}$.

We should emphasize that the lower bound provided in [134] is directly comparable to our results. Authors in [134] consider the same problem and give two results: first, if $m \geq \mathscr{O}\left(\|\bar{\mathbf{a}}\|_1^2 k \log d\right)$ then minimizing $\|\mathbf{X}\|_1 + \lambda \operatorname{tr}(\mathbf{X})$ for suitable value of λ over the set of PSD matrices will exactly recover $\mathbf{X}_0$ with high

probability. Secondly, their Theorem 1.3 gives a necessary condition (lower bound) on the number of measurements, under which the recovery program fails to recover $\mathbf{X}_0$ with high probability. In particular, their failure condition is $m \leq \min\{m_0, \frac{d}{40\log d}\}$ where $m_0 = \frac{\max(\|\bar{\mathbf{a}}\|_1^2 - k/2, 0)^2}{500\log^2 d}$.

First, observe that both results have $m \leq \mathscr{O}\left(\frac{d}{\log d}\right)$ condition. Focusing on the sparsity requirements, when the nonzero entries are sufficiently diffused (i.e. $\|\mathbf{a}\|_1^2 \approx k$) both results yield $\mathscr{O}\left(\frac{\|\bar{\mathbf{a}}\|^4}{\log^2 d}\right)$ as a lower bound. On the other hand, if $\|\bar{\mathbf{a}}\|_1 \leq \sqrt{\frac{k}{2}}$, their lower bound disappears while our lower bound still requires $\mathscr{O}\left(\frac{\|\bar{\mathbf{a}}\|^4}{\log^2 d}\right)$ measurements. $\|\bar{\mathbf{a}}\|_1 \leq \sqrt{\frac{k}{2}}$ can happen as soon as the nonzero entries are rather spiky, i.e. some of the entries are much larger than the rest. In this sense, our bounds are tighter. On the other hand, their lower bound includes the PSD constraint unlike ours.

5.3.4 Asymptotic regime

While we discussed two cases in the nonasymptotic setup, we believe significantly more general results can be stated asymptotically ($m, n \to \infty$). For instance, under finite fourth moment constraint, thanks to Bai-Yin law [8], asymptotically, the smallest singular value of a matrix with i.i.d. unit variance entries concentrate around $\sqrt{n} - \sqrt{m}$. Similarly, $\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2$ is sum of independent variables; hence thanks to the law of large numbers, we will have $\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2^2}{m} \to 1$. Together, these yield $\frac{\|\mathbf{A}\bar{\mathbf{x}}_0\|_2}{\sigma_{\min}(\mathbf{A}^T)} \to \frac{\sqrt{m}}{\sqrt{n}-\sqrt{m}}$.

5.4 Upper bounds

We now state an upper bound on the simultaneous optimization for Gaussian measurement ensemble. Our upper bound will be in terms of distance to the dilated subdifferentials.

To accomplish this, we will make use of the characterization of the linear inverse phase transitions via Gaussian width; which has been discussed in Chapter 2 and Chapter 3. The works due to Chandrasekaran et al. [50], Amelunxen et al. [4] and Donoho and Tanner [83] focus on signals with single structure and do not study properties of a penalty that is a combination of norms. The next theorem relates the phase transition point of the joint optimization (5.3) to the individual subdifferentials.

Theorem 5.4.1 *Suppose* $\mathbf{A}$ *has i.i.d.* $\mathscr{N}(0,1)$ *entries and let* $f(\mathbf{x}) = \sum_{i=1}^{\tau} \lambda_i \|\mathbf{x}\|_{(i)}$. *For positive scalars* $\{\alpha_i\}_{i=1}^{\tau}$, *let* $\bar{\lambda}_i = \frac{\lambda_i \alpha_i^{-1}}{\sum_{i=1}^{\tau} \lambda_i \alpha_i^{-1}}$ *and define,*

$$m_{up}(\{\alpha_i\}_{i=1}^{\tau}) := \left(\sum_i \bar{\lambda}_i \mathbf{D}(\alpha_i \partial \|\mathbf{x}_0\|_{(i)})^{1/2} \right)^2$$

If $m \geq (\sqrt{m_{up}} + t)^2 + 1$, *then program* (5.3) *will succeed with probability* $1 - 2\exp(-\frac{t^2}{2})$.

Proof: Fix $\mathbf{h}$ as an i.i.d. standard normal vector. Let $\mathbf{g}_i$ be so that $\alpha_i \mathbf{g}_i$ is closest to $\mathbf{h}$ over $\alpha_i \partial \|\mathbf{x}_0\|_{(i)}$. Let $\gamma = (\sum_i \frac{\lambda_i}{\alpha_i})^{-1}$. Then, we may write,

$$\begin{aligned}
\inf_{\mathbf{g}' \in \text{cone}(\partial f(\mathbf{x}_0))} \|\mathbf{h} - \mathbf{g}'\|_2 &\leq \inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} \|\mathbf{h} - \gamma \mathbf{g}\|_2 \\
&\leq \|\mathbf{h} - \gamma \sum_i \lambda_i \mathbf{g}_i\|_2 \\
&= \|\mathbf{h} - \gamma \sum_i \frac{\lambda_i}{\alpha_i} \alpha_i \mathbf{g}_i\|_2 = \|\mathbf{h} - \sum_i \bar{\lambda}_i \alpha_i \mathbf{g}_i\|_2 \\
&\leq \sum_i \bar{\lambda}_i \|\mathbf{h} - \alpha_i \mathbf{g}_i\|_2 \\
&= \sum_i \bar{\lambda}_i \inf_{\mathbf{g}_i' \in \partial \|\mathbf{x}_0\|_{(i)}} \|\mathbf{h} - \alpha_i \mathbf{g}_i'\|_2
\end{aligned}$$

Taking the expectations of both sides and using the definition of $\mathbf{D}(\cdot)$, we find,

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))^{1/2} \leq \sum_i \bar{\lambda}_i \mathbf{D}(\alpha_i \partial \|\mathbf{x}_0\|_{(i)})^{1/2}.$$

Using definition of $\mathbf{D}(\cdot)$, this gives, $m_{up} \geq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. The result then follows from Proposition 1.3, which gives that when $m \geq (\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))^{1/2} + t)^2 + 1$, recovery succeeds with probability $1 - 2\exp(-\frac{t^2}{2})$. To see this, recall that,

$$\mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0))) \geq \omega(\mathscr{T}_f(\mathbf{x}_0) \cap \mathscr{B}^{n-1})^2 \tag{5.7}$$

■

For Theorem 5.4.1 to be useful, choices of α_i should be made wisely. An obvious choice is letting,

$$\alpha_i^* = \arg\min_{\alpha_i \geq 0} \mathbf{D}(\alpha_i \partial \|\mathbf{x}_0\|_{(i)}). \tag{5.8}$$

With this choice, our upper bounds can be related to the individual sample complexities, which is equal to $\mathbf{D}(\text{cone}(\partial \|\mathbf{x}_0\|_{(i)}))$. Proposition 1 of [101] shows that, if $\|\cdot\|_{(i)}$ is a *decomposable* norm, then,

$$\mathbf{D}(\text{cone}(\partial \|\mathbf{x}_0\|_{(i)}))^{1/2} \leq \mathbf{D}(\alpha_i^* \partial \|\mathbf{x}_0\|_{(i)})^{1/2} \leq \mathbf{D}(\text{cone}(\partial \|\mathbf{x}_0\|_{(i)}))^{1/2} + 6$$

Decomposability is defined and discussed in detail in Section 5.5.4. In particular, $\ell_1, \ell_{1,2}$ and the nuclear norm are decomposable. With this assumption, our upper bound will suggest that, the sample complexity of the simultaneous optimization is smaller than a certain convex combination of individual sample complexities.

Corollary 5.4.1 *Suppose* $\mathbf{A}$ *has i.i.d* $\mathcal{N}(0,1)$ *entries and let* $f(\mathbf{x}) = \sum_{i=1}^{\tau} \lambda_i \|\mathbf{x}\|_{(i)}$ *for decomposable norms* $\{\|\cdot\|_{(i)}\}_{i=1}^{\tau}$. *Let* $\{\alpha_i^*\}_{i=1}^{\tau}$ *be as in* (5.8) *and assume they are strictly positive. Let* $\bar{\lambda}_i^* = \frac{\lambda_i(\alpha_i^*)^{-1}}{\sum_{i=1}^{\tau}\lambda_i(\alpha_i^*)^{-1}}$ *and define,*

$$\sqrt{m_{up}(\{\alpha_i^*\}_{i=1}^{\tau})} := \sum_i \bar{\lambda}_i^* \mathbf{D}(\text{cone}(\partial\|\mathbf{x}_0\|_{(i)}))^{1/2} + 6$$

If $m \geq (\sqrt{m_{up}} + t)^2 + 1$, *then program* (5.3) *will succeed with probability* $1 - 2\exp(-\frac{t^2}{2})$.

Here, we used the fact that $\sum_i \bar{\lambda}_i^* = 1$ to take 6 out of the sum over i. We note that Corollaries 5.2.1 and 5.4.1 can be related in the case of sparse and low-rank matrices. For norms of interest, roughly speaking,

- $n\kappa_i^2$ is proportional to the sample complexity $\mathbf{D}(\text{cone}(\partial\|\mathbf{x}_0\|_{(i)}))$.
- L_i is proportional to $\frac{\sqrt{n}}{\alpha_i^*}$.

Consequently, the sample complexity of (5.3) will be upper and lower bounded by similar convex combinations.

5.4.1 Upper bounds for the S&L model

We will now apply the bound obtained in Theorem 5.4.1 for S&L matrices. To obtain simple and closed form bounds, we will make use of the existing results in the literature.

- Table II of [101]: If $\mathbf{x}_0 \in \mathbb{R}^n$ is a k sparse vector, choosing $\alpha_{\ell_1} = \sqrt{2\log\frac{n}{k}}$, $\mathbf{D}(\alpha_{\ell_1}\partial\|\mathbf{x}_0\|_1) \leq 2k\log\frac{en}{k}$.
- Table 3 of [170]: If $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$ is a rank r matrix, choosing $\alpha_\star = 2\sqrt{d}$, $\mathbf{D}(\alpha_\star\partial\|\mathbf{X}_0\|_\star) \leq 6dr + 2d$.

Proposition 5.4.1 *Suppose* $\mathbf{A}$ *has i.i.d* $\mathcal{N}(0,1)$ *entries and* $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$ *is a rank* r *matrix whose nonzero entries lie on a* $k\times k$ *submatrix. For* $0 \leq \beta \leq 1$, *let* $f(\mathbf{X}) = \lambda_{\ell_1}\|\mathbf{X}\|_1 + \lambda_\star\|\mathbf{X}\|_\star$ *where* $\lambda_{\ell_1} = \beta\sqrt{\log\frac{d}{k}}$ *and* $\lambda_\star = (1-\beta)\sqrt{d}$. *Then, whenever,*

$$m \geq \left(2\beta k\sqrt{\log\frac{ed}{k}} + (1-\beta)\sqrt{6dr+2d} + t\right)^2 + 1,$$

$\mathbf{X}_0$ *can be recovered via* (5.3) *with probability* $1 - 2\exp(-\frac{t^2}{2})$.

Proof: To apply Theorem 5.4.1, we will choose $\alpha_{\ell_1} = \sqrt{4\log\frac{d}{k}}$ and $\alpha_\star = 2\sqrt{d}$. $\mathbf{X}_0$ is effectively an (at most) k^2 sparse vector of size d^2. Hence, $\alpha_{\ell_1} = \sqrt{2\log\frac{d^2}{k^2}}$ and $\mathbf{D}(\alpha_{\ell_1}\|\mathbf{X}_0\|_1) \leq 4k^2\log\frac{ed}{k}$.

Now, for the choice of $\alpha_\star$, we have, $\mathbf{D}(\alpha_\star\|\mathbf{X}_0\|_\star) \leq 6dr + 2d$. Observe that $\alpha_{\ell_1}^{-1}\lambda_{\ell_1} = \frac{\beta}{2}$, $\alpha_\star^{-1}\lambda_\star = \frac{1-\beta}{2}$ and apply Theorem 5.4.1 to conclude. ■

5.5 General Simultaneously Structured Model Recovery

Recall the setup from Section 5.1 where we consider a vector $\mathbf{x}_0 \in \mathbb{R}^n$ whose structures are associated with a family of norms $\{\|\cdot\|_{(i)}\}_{i=1}^{\tau}$ and $\mathbf{x}_0$ satisfies the cone constraint $\mathbf{x}_0 \in \mathscr{C}$. This section is dedicated to the proofs of theorems in Section 5.2.1 and additional side results where the goal is to find lower bounds on the required number of measurements to recover $\mathbf{x}_0$. For a subspace M, denote its orthogonal complement by $M^\perp$.

5.5.1 Preliminary Lemmas

We first show that the objective function $\max_{1\leq i\leq\tau}\frac{\|\mathbf{x}\|_{(i)}}{\|\mathbf{x}_0\|_{(i)}}$ can be viewed as the 'best' among the functions mentioned in (5.3) for recovery of $\mathbf{x}_0$.

Lemma 5.5.1 *Consider the class of recovery programs in* (5.3)*. If the program*

$$\begin{aligned} \underset{\mathbf{x}\in\mathscr{C}}{\textit{minimize}} \quad & f_{\text{best}}(\mathbf{x}) \triangleq \max_{i=1,\ldots,\tau}\frac{\|\mathbf{x}\|_{(i)}}{\|\mathbf{x}_0\|_{(i)}} \\ \textit{subject to} \quad & \mathscr{A}(\mathbf{x}) = \mathscr{A}(\mathbf{x}_0) \end{aligned} \tag{5.9}$$

fails to recover $\mathbf{x}_0$*, then any member of this class will also fail to recover* $\mathbf{x}_0$*.*

Proof: Suppose (5.9) does not have $\mathbf{x}_0$ as an optimal solution and there exists $\mathbf{x}'$ such that $f_{\text{best}}(\mathbf{x}') \leq f_{\text{best}}(\mathbf{x}_0)$, then

$$\frac{1}{\|\mathbf{x}_0\|_{(i)}}\|\mathbf{x}'\|_{(i)} \leq f_{\text{best}}(\mathbf{x}') \leq f_{\text{best}}(\mathbf{x}_0) = 1, \quad \text{for } i = 1,\ldots,\tau,$$

which implies,

$$\|\mathbf{x}'\|_{(i)} \leq \|\mathbf{x}_0\|_{(i)}, \quad \text{for all } i = 1,\ldots,\tau. \tag{5.10}$$

Conversely, given (5.10), we have $f_{\text{best}}(\mathbf{x}') \leq f_{\text{best}}(\mathbf{x}_0)$ from the definition of f_{best}.

Furthermore, since we assume $h(\cdot)$ in (5.3) is non-decreasing in its arguments and increasing in at least one of them, (5.10) implies $f(\mathbf{x}') \leq f(\mathbf{x}_0)$ for any such function $f(\cdot)$. Thus, failure of $f_{\text{best}}(\cdot)$ in recovery of

$\mathbf{x}_0$ implies failure of any other function in (5.3) in this task. ■

The following lemma gives necessary conditions for $\mathbf{x}_0$ to be a minimizer of the problem (5.3).

Lemma 5.5.2 *If $\mathbf{x}_0$ is a minimizer of the program (5.3), then there exist $\mathbf{v} \in \mathscr{C}^*$, $\mathbf{z}$, and $\mathbf{g} \in \partial f(\mathbf{x}_0)$ such that*

$$\mathbf{g} - \mathbf{v} - \mathbf{A}^T\mathbf{z} = 0 \quad \textit{and} \quad \langle \mathbf{x}_0, \mathbf{v} \rangle = 0.$$

The proof of Lemma 5.5.2 follows from the KKT conditions for (5.3) to have $\mathbf{x}_0$ as an optimal solution [17, Section 4.7].

The next lemma describes the subdifferential of any general function $f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)})$ as discussed in Section 5.1.1.

Lemma 5.5.3 *For any subgradient of the function $f(\mathbf{x}) = h(\|\mathbf{x}\|_{(1)}, \ldots, \|\mathbf{x}\|_{(\tau)})$ at $\mathbf{x} \neq 0$ defined by convex function $h(\cdot)$, there exists non-negative constants w_i, $i = 1, \ldots, \tau$ such that*

$$\mathbf{g} = \sum_{i=1}^{\tau} w_i \mathbf{g}_i$$

where $\mathbf{g}_i \in \partial \|\mathbf{x}_0\|_{(i)}$.

Proof: Consider the function $N(\mathbf{x}) = \begin{bmatrix} \|\mathbf{x}\|_{(1)}, & \ldots, & \|\mathbf{x}\|_{(\tau)} \end{bmatrix}^T$ by which we have $f(\mathbf{x}) = h(N(\mathbf{x}))$. By Theorem 10.49 in [183] we have

$$\partial f(\mathbf{x}) = \bigcup \{ \partial(\mathbf{y}^T N(\mathbf{x})) : \ \mathbf{y} \in \partial h(N(\mathbf{x})) \}$$

where we used the convexity of f and h. Now notice that any $\mathbf{y} \in \partial h(N(\mathbf{x}))$ is a non-negative vector because of the monotonicity assumption on $h(\cdot)$. This implies that any subgradient $\mathbf{g} \in \partial f(\mathbf{x})$ is in the form of $\partial(\mathbf{w}^T N(\mathbf{x}))$ for some nonnegative vector $\mathbf{w}$. The desired result simply follows because subgradients of conic combination of norms are conic combinations of their subgradients, (see e.g. [182]). ■

Using Lemmas 5.5.2 and 5.5.3, we now provide the proofs of Theorems 5.2.1 and 5.2.2.

5.5.2 Proof of Theorem 5.2.1

We prove the more general version of Theorem 5.2.1, which can take care of the cone constraint and alignment of subgradients over arbitrary subspaces. This will require us to extend the definition of correlation to

handle subspaces. For a linear subspace $\mathscr{R} \in \mathbb{R}^n$ and a set $S \in \mathbb{R}^n$, we define,

$$\rho(\mathscr{R}, S) = \inf_{0 \neq \mathbf{s} \in S} \frac{\|\text{Proj}_{\mathscr{R}}(\mathbf{s})\|_2}{\|\mathbf{s}\|_2}.$$

Proposition 5.5.1 *Let* $\sigma_{\mathscr{C}}(\mathbf{A}^T) = \inf_{\|\mathbf{z}\|_2=1} \frac{\|Proj_{\mathscr{C}}(\mathbf{A}^T\mathbf{z})\|_2}{\|\mathbf{A}^T\mathbf{z}\|_2}$. *Let* $\mathscr{R}$ *be an arbitrary linear subspace orthogonal to the following cone,*

$$\{\mathbf{y} \in \mathbb{R}^n \big| \mathbf{x}_0^T\mathbf{y} = 0,\ \mathbf{y} \in \mathscr{C}^*\}. \tag{5.11}$$

Suppose,

$$\rho(\mathscr{R}, \partial f(\mathbf{x}_0)) := \inf_{\mathbf{g} \in \partial f(\mathbf{x}_0)} \frac{\|Proj_{\mathscr{R}}(\mathbf{g})\|_2}{\|\mathbf{g}\|_2} > \frac{\sigma_{\max}(Proj_{\mathscr{R}}(\mathbf{A}^T))}{\sigma_{\mathscr{C}}(\mathbf{A}^T)\sigma_{\min}(\mathbf{A}^T)}.$$

Then, $\mathbf{x}_0$ *is not a minimizer of* (5.3).

Proof: Suppose $\mathbf{x}_0$ is a minimizer of (5.3). From Lemma 5.5.2, there exist a $\mathbf{g} \in \partial f(\mathbf{x}_0)$, $\mathbf{z} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathscr{C}^*$ such that

$$\mathbf{g} = \mathbf{A}^T\mathbf{z} + \mathbf{v} \tag{5.12}$$

and $\langle \mathbf{x}_0, \mathbf{v} \rangle = 0$. We will first eliminate the contribution of $\mathbf{v}$ in equation (5.12). Projecting both sides of (5.12) onto the subspace $\mathscr{R}$ gives,

$$\text{Proj}_{\mathscr{R}}(\mathbf{g}) = \text{Proj}_{\mathscr{R}}(\mathbf{A}^T\mathbf{z}) = \text{Proj}_{\mathscr{R}}(\mathbf{A}^T)\mathbf{z} \tag{5.13}$$

Taking the ℓ_2 norms,

$$\|\text{Proj}_{\mathscr{R}}(\mathbf{g})\|_2 = \|\text{Proj}_{\mathscr{R}}(\mathbf{A}^T)\mathbf{z}\|_2 \leq \sigma_{\max}(\text{Proj}_{\mathscr{R}}(\mathbf{A}^T))\|\mathbf{z}\|_2. \tag{5.14}$$

Since $\mathbf{v} \in \mathscr{C}^*$, from Fact 2.1 we have $\text{Proj}_{\mathscr{C}}(-\mathbf{v}) = \text{Proj}_{\mathscr{C}}(\mathbf{A}^T\mathbf{z} - \mathbf{g}) = 0$. Using Corollary B.1.1,

$$\|\mathbf{g}\|_2 \geq \|\text{Proj}_{\mathscr{C}}(\mathbf{A}^T\mathbf{z})\|_2. \tag{5.15}$$

From the initial assumption, for any $\mathbf{z} \in \mathbb{R}^m$, we have,

$$\sigma_{\mathscr{C}}(\mathbf{A}^T)\|\mathbf{A}^T\mathbf{z}\|_2 \leq \|\text{Proj}_{\mathscr{C}}(\mathbf{A}^T\mathbf{z})\|_2 \tag{5.16}$$

Combining (5.15) and (5.16) yields $\|\mathbf{g}\|_2 \geq \sigma_{\mathscr{C}}(\mathbf{A}^T)\|\mathbf{A}^T\mathbf{z}\|_2$. Further incorporating (5.14), we find,

$$\frac{\|\text{Proj}_{\mathscr{R}}(\mathbf{g})\|_2}{\sigma_{\max}(\text{Proj}_{\mathscr{R}}(\mathbf{A}^T))} \leq \|\mathbf{z}\|_2 \leq \frac{\|\mathbf{A}^T\mathbf{z}\|_2}{\sigma_{\min}(\mathbf{A}^T)} \leq \frac{\|\mathbf{g}\|_2}{\sigma_{\mathscr{C}}(\mathbf{A}^T)\sigma_{\min}(\mathbf{A}^T)}.$$

Hence, if $\mathbf{x}_0$ is recoverable, there exists $\mathbf{g} \in \partial f(\mathbf{x}_0)$ satisfying,

$$\frac{\|\text{Proj}_{\mathscr{R}}(\mathbf{g})\|_2}{\|\mathbf{g}\|_2} \leq \frac{\sigma_{\max}(\text{Proj}_{\mathscr{R}}(\mathbf{A}^T))}{\sigma_{\mathscr{C}}(\mathbf{A}^T)\sigma_{\min}(\mathbf{A}^T)}.$$

■

To obtain Theorem 5.2.1, choose $\mathscr{R} = \text{span}(\{\mathbf{x}_0\})$ and $\mathscr{C} = \mathbb{R}^n$. This choice of $\mathscr{R}$ yields $\sigma_{\max}(\text{Proj}_{\mathscr{R}}(\mathbf{A}^T)) = \|\bar{\mathbf{x}}_0\bar{\mathbf{x}}_0^T\mathbf{A}^T\|_2 = \|\mathbf{A}\bar{\mathbf{x}}_0\|_2$ and $\|\text{Proj}_{\mathscr{R}}(\mathbf{g})\|_2 = |\bar{\mathbf{x}}_0^T\mathbf{g}|$. Choice of $\mathscr{C} = \mathbb{R}^n$ yields $\sigma_{\mathscr{C}}(\mathbf{A}) = 1$. Also note that, for any choice of $\mathscr{C}$, $\mathbf{x}_0$ is orthogonal to (5.11) by definition.

5.5.3 Proof of Theorem 5.2.2

Rotational invariance of Gaussian measurements allow us to make full use of Proposition 5.5.1. The following is a generalization of Theorem 5.2.2.

Proposition 5.5.2 *Consider the setup in Proposition 5.5.1 where* $\mathbf{A}$ *has i.i.d* $\mathscr{N}(0,1)$ *entries. Let,*

$$m_{low} = \frac{n(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})\rho(\mathscr{R},\partial f(\mathbf{x}_0))^2}{100},$$

and suppose $\dim(\mathscr{R}) \leq m_{low}$*. Then, whenever* $m \leq m_{low}$*, with probability* $1-10\exp(-\frac{1}{16}\min\{m_{low},(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})^2 n\})$*,* (5.3) *will fail for all functions* $f(\cdot)$*.*

Proof: More measurements can only increase the chance of success. Hence, without losing generality, assume $m = m_{low}$ and $\dim(\mathscr{R}) \leq m$. The result will follow from Proposition 5.5.1. Recall that $m \leq \frac{(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})n}{100}$.

- $\text{Proj}_{\mathscr{R}}(\mathbf{A}^T)$ is statistically identical to a $\dim(\mathscr{R}) \times m$ matrix with i.i.d. $\mathscr{N}(0,1)$ entries under proper unitary rotation. Hence, using Corollary 5.35 of [213], with probability $1-2\exp(-\frac{m}{8})$, $\sigma_{\max}(\text{Proj}_{\mathscr{R}}(\mathbf{A}^T)) \leq 1.5\sqrt{m}+\sqrt{\dim(\mathscr{R})} \leq 2.5\sqrt{m}$. With the same probability, $\sigma_{\min}(\mathbf{A}^T) \geq \sqrt{n}-1.5\sqrt{m}$.
- From Theorem B.1.2, using $m \leq \frac{(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})n}{100}$, with probability $1-6\exp(-\frac{(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})^2 n}{16})$, $\sigma^2_{\mathscr{C}}(\mathbf{A}^T) \geq \frac{1-\bar{\mathbf{D}}(\mathscr{C})^{1/2}}{4(1+\bar{\mathbf{D}}(\mathscr{C})^{1/2})} \geq \frac{1-\bar{\mathbf{D}}(\mathscr{C})^{1/2}}{8}$.

Since $\frac{m}{n} \leq \frac{1}{30}$, combining these, with the desired probability,

$$\frac{\sigma_{\max}(\text{Proj}_{\mathscr{R}}(\mathbf{A}^T))}{\sigma_{\mathscr{C}}(\mathbf{A}^T)\sigma_{\min}(\mathbf{A}^T)} \leq \sqrt{\frac{8}{1-\bar{\mathbf{D}}(\mathscr{C})^{1/2}}} \frac{2.5\sqrt{m}}{\sqrt{n}-1.5\sqrt{m}} < \frac{10\sqrt{m}}{\sqrt{(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})n}}.$$

Finally, using Proposition 5.5.1 and $m \leq \frac{n(1-\bar{\mathbf{D}}(\mathscr{C})^{1/2})}{100}\rho(\mathscr{R}, \partial f(\mathbf{x}_0))^2$, with the same probability (5.3) fails.

■

To achieve Theorem 5.2.2, choose $\mathscr{R} = \text{span}(\{\mathbf{x}_0\})$ and use the first statement of Proposition 5.2.1.

To achieve Corollary 5.2.1, choose $\mathscr{R} = \text{span}(\{\mathbf{x}_0\})$ and use the second statement of Proposition 5.2.1.

5.5.4 Enhanced lower bounds

From our initial results, it may look like our lower bounds are suboptimal. For instance, considering only ℓ_1 norm, $\kappa = \frac{\|\bar{\mathbf{x}}_0\|_1}{\sqrt{n}}$ lies between $\frac{1}{\sqrt{n}}$ and $\sqrt{\frac{k}{n}}$ for a k sparse signal. Combined with Theorem 5.2.2, this would give a lower bound of $\|\bar{\mathbf{x}}_0\|_1^2$ measurements. On the other hand, clearly, we need at least $\mathscr{O}(k)$ measurements to estimate a k sparse vector.

Indeed, Proposition 5.5.1 gives such a bound with a better choice of $\mathscr{R}$. In particular, let us choose $\mathscr{R} = \text{span}(\{\text{sign}(\mathbf{x}_0)\})$. For any $\mathbf{g} \in \partial\|\mathbf{x}_0\|_1$, we have that,

$$\frac{\left\langle \mathbf{g}, \frac{\text{sign}(\mathbf{x}_0)}{\sqrt{k}} \right\rangle}{L} = \sqrt{\frac{k}{n}} \implies \rho(\text{sign}(\mathbf{x}_0), \partial\|\mathbf{x}_0\|_1) = \sqrt{\frac{k}{n}}$$

Hence, we immediately have $m \geq \mathscr{O}(k)$ as a lower bound. The idea of choosing such sign vectors can be generalized to the so-called decomposable norms.

Definition 5.5.1 (Decomposable Norm) *A norm $\|\cdot\|$ is decomposable at $\mathbf{x} \in \mathbb{R}^n$ if there exist a subspace $T \subset \mathbb{R}^n$ and a vector $\mathbf{e} \in T$ such that the subdifferential at $\mathbf{x}$ has the form*

$$\partial\|\mathbf{x}\| = \{\mathbf{z} \in \mathbb{R}^n \ : \ Proj_T(\mathbf{z}) = \mathbf{e}\,, \ \|\mathscr{P}_{T^\perp}(\mathbf{z})\|^* \leq 1\}.$$

We refer to T as the support *and $\mathbf{e}$ as the* sign vector *of $\mathbf{x}$ with respect to $\|\cdot\|$.*

Similar definitions are used in [30] and [221]. Our definition is simpler and less strict compared to these works. Note that L is a global property of the norm while $\mathbf{e}$ and T depend on both the norm and the point under consideration (decomposability is a local property in this sense).

To give some intuition for Definition 5.5.1, we review examples of norms that arise when considering simultaneously sparse and low rank matrices. For a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, let $\mathbf{X}_{i,j}$, $\mathbf{X}_{i,\cdot}$ and $\mathbf{X}_{\cdot,j}$ denote its (i,j) entry, ith row and jth column respectively.

Lemma 5.5.4 (see [30]) *The ℓ_1 norm, the $\ell_{1,2}$ norm and the nuclear norm are decomposable as follows.*

• **ℓ_1 norm** *is decomposable at every $\mathbf{x} \in \mathbb{R}^n$, with sign $\mathbf{e} = \operatorname{sgn}(\mathbf{x})$, and support as*

$$T = \operatorname{supp}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \ \mathbf{x}_i = 0 \ \Rightarrow \ \mathbf{y}_i = 0 \ \textit{for } i = 1, \ldots, n\} .$$

• **$\ell_{1,2}$ norm** *is decomposable at every $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. The support is*

$$T = \left\{\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2} : \ \mathbf{X}_{\cdot,i} = \mathbf{0} \ \Rightarrow \ \mathbf{Y}_{\cdot,i} = \mathbf{0} \ \textit{for } i = 1, \ldots, d_2\right\},$$

and the sign vector $\mathbf{e} \in \mathbb{R}^{d_1 \times d_2}$ is obtained by normalizing the columns of $\mathbf{X}$ present in the support, $\mathbf{e}_{\cdot,j} = \frac{\mathbf{X}_{\cdot,j}}{\|\mathbf{X}_{\cdot,j}\|_2}$ if $\|\mathbf{X}_{\cdot,j}\|_2 \neq 0$, and setting the rest of the columns to zero.

• **Nuclear norm** *is decomposable at every $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$. For a matrix $\mathbf{X}$ with rank r and compact singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$, we have $\mathbf{e} = \mathbf{U}\mathbf{V}^T$ and*

$$\begin{aligned} T &= \left\{\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2} : \ (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{Y}(\mathbf{I} - \mathbf{V}\mathbf{V}^T) = \mathbf{0}\right\} \\ &= \left\{\mathbf{Z}_1\mathbf{V}^T + \mathbf{U}\mathbf{Z}_2^T \mid \mathbf{Z}_1 \in \mathbb{R}^{d_1 \times r}, \mathbf{Z}_2 \in \mathbb{R}^{d_2 \times r}\right\}. \end{aligned}$$

The next lemma shows that the sign vector $\mathbf{e}$ will yield the largest correlation with the subdifferential and the best lower bound for such norms.

Lemma 5.5.5 *Let $\|\cdot\|$ be a decomposable norm with support T and sign vector $\mathbf{e}$. For any $\mathbf{v} \neq 0$, we have that,*

$$\rho(\mathbf{v}, \partial\|\mathbf{x}_0\|) \leq \rho(\mathbf{e}, \partial\|\mathbf{x}_0\|) \tag{5.17}$$

Also $\rho(\mathbf{e}, \partial\|\mathbf{x}_0\|) \geq \frac{\|\mathbf{e}\|_2}{L}$.

Proof: Let $\mathbf{v}$ be a unit vector. Without losing generality, assume $\mathbf{v}^T\mathbf{e} \geq 0$. Pick a vector $\mathbf{z} \in T^\perp$ with $\|\mathbf{z}\|^* = 1$ such that $\mathbf{z}^T\mathbf{v} \leq 0$ (otherwise pick $-\mathbf{z}$). Now, consider the class of subgradients $\mathbf{g}(\alpha) = \mathbf{e} + \alpha\mathbf{z}$ for

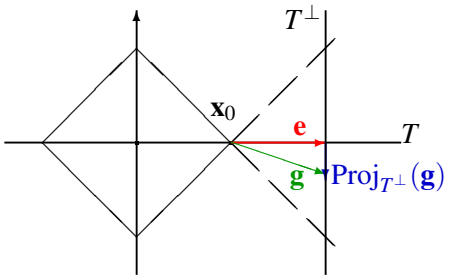


Figure 5.4: An example of a decomposable norm: ℓ_1 norm is decomposable at $\mathbf{x}_0 = (1,0)$. The sign vector $\mathbf{e}$, the support T, and shifted subspace $T^\perp$ are illustrated. A subgradient $\mathbf{g}$ at $\mathbf{x}_0$ and its projection onto $T^\perp$ are also shown.

$1 \geq \alpha \geq -1$. Then,

$$\inf_{-1\leq\alpha\leq 1} \frac{|\mathbf{v}^T\mathbf{g}(\alpha)|}{\|\mathbf{g}(\alpha)\|_2} = \inf_{0\leq\alpha\leq 1} \frac{|\mathbf{v}^T\mathbf{g}(\alpha)|}{\|\mathbf{g}(\alpha)\|_2} = \inf_{0\leq\alpha\leq 1} \frac{|\mathbf{e}^T\mathbf{v} - \alpha|\mathbf{z}^T\mathbf{v}||}{(\|\mathbf{e}\|_2^2 + \alpha^2\|\mathbf{z}\|_2^2)^{1/2}}$$

If $|\mathbf{z}^T\mathbf{v}| \geq \mathbf{e}^T\mathbf{v}$, then, the numerator can be made 0 and $\rho(\mathbf{v}, \partial\|\mathbf{x}_0\|) = 0$. Otherwise, the right hand side is decreasing function of α, hence the minimum is achieved at $\alpha = 1$, which gives,

$$\inf_{-1\leq\alpha\leq 1} \frac{|\mathbf{v}^T\mathbf{g}(\alpha)|}{\|\mathbf{g}(\alpha)\|_2} = \frac{|\mathbf{e}^T\mathbf{v} - |\mathbf{z}^T\mathbf{v}||}{(\|\mathbf{e}\|_2^2 + \|\mathbf{z}\|_2^2)^{1/2}} \leq \frac{|\mathbf{e}^T\mathbf{v}|}{(\|\mathbf{e}\|_2^2 + \|\mathbf{z}\|_2^2)^{1/2}} \leq \frac{\|\mathbf{e}\|_2}{(\|\mathbf{e}\|_2^2 + \|\mathbf{z}\|_2^2)^{1/2}} = \inf_{-1\leq\alpha\leq 1} \frac{|\bar{\mathbf{e}}^T\mathbf{g}(\alpha)|}{\|\mathbf{g}(\alpha)\|_2}$$

where we used $\mathbf{e}^T\mathbf{g}(\alpha) = \mathbf{e}^T\mathbf{e} = \|\mathbf{e}\|_2^2$. Hence, along any direction $\mathbf{z}$, $\mathbf{e}$ yields a higher minimum correlation than $\mathbf{v}$. To obtain (5.17), further take infimum over all $\mathbf{z} \in T^\perp, \|\mathbf{z}\|^* \leq 1$ which will yield infimum over $\partial\|\mathbf{x}_0\|$. Finally, use $\|g(\alpha)\|_2 \leq L$ to lower bound $\rho(\mathbf{e}, \partial\|\mathbf{x}_0\|)$.

■

Based on Lemma 5.5.5, the individual lower bound would be $\mathscr{O}\left(\frac{\|\mathbf{e}\|_2^2}{L^2}\right)n$. Calculating $\frac{\|\mathbf{e}\|_2^2}{L^2}n$ for the norms in Lemma 5.5.4, reveals that, this quantity is k for a k sparse vector, cd_1 for a c-column sparse matrix and $r\max\{d_1,d_2\}$ for a rank r matrix. Compared to bounds obtained by using $\bar{\mathbf{x}}_0$, these new quantities are directly proportional to the true model complexities. Finally, we remark that, these new bounds correspond to choosing $\mathbf{x}_0$ that maximizes the value of $\|\bar{\mathbf{x}}_0\|_1, \|\bar{\mathbf{x}}_0\|_\star$ or $\|\bar{\mathbf{x}}_0\|_{1,2}$ while keeping sparsity, rank or column sparsity fixed. In particular, in these examples, $\mathbf{e}$ has the same sparsity, rank, column sparsity as $\mathbf{x}_0$.

The next lemma gives a correlation bound for the combination of decomposable norms as well as a simple lower bound on the sample complexity.

Proposition 5.5.3 *Given decomposable norms $\|\cdot\|_{(i)}$ with supports T_i and sign vectors $\mathbf{e}_i$. Let $T_\cap = \bigcap_{1\leq i\leq\tau} T_i$. Choose the subspace $\mathscr{R}$ to be a subset of $T_\cap$.*

- *Assume* $\left\langle Proj_{\mathscr{R}}(\mathbf{e}_i), Proj_{\mathscr{R}}(\mathbf{e}_j)\right\rangle \geq 0$ *for all* i, j *and* $\min_{1\leq i\leq \tau} \frac{\|Proj_{\mathscr{R}}(\mathbf{e}_i)\|_2}{\|\mathbf{e}_i\|_2} \geq \upsilon$. *Then,*

$$\rho(\mathscr{R}, \partial f(\mathbf{x}_0)) \geq \frac{\upsilon}{\sqrt{\tau}} \min_{1\leq i\leq \tau} \rho(\mathbf{e}_i, \partial \|\mathbf{x}_0\|_{(i)}).$$

- *Consider Proposition 5.5.1 with Gaussian measurements and suppose* $\mathscr{R}$ *is orthogonal to the set (5.11). Let* $f(\mathbf{x}) = \sum_{i=1}^{\tau} \lambda_i \|\mathbf{x}\|_{(i)}$ *for nonnegative* $\{\lambda_i\}$*'s. Then, if* $m < dim(\mathscr{R})$, *(5.3) fails with probability* 1.

Proof: Let $\mathbf{g} = \sum_{i=1}^{\tau} w_i \mathbf{g}_i$ for some $\mathbf{g}_i \in \partial \|\mathbf{x}_0\|_{(i)}$. First, $\|\mathbf{g}\|_2 \leq \sum_{i=1}^{\tau} w_i \|\mathbf{g}_i\|_2$. Next,

$$\|\text{Proj}_{\mathscr{R}}(\mathbf{g})\|_2^2 = \|\sum_{i=1}^{\tau} w_i \text{Proj}_{\mathscr{R}}(\mathbf{e}_i)\|_2^2 \geq \sum_{i=1}^{\tau} w_i^2 \|\text{Proj}_{\mathscr{R}}(\mathbf{e}_i)\|_2^2 \geq \upsilon^2 \sum_{i=1}^{\tau} w_i^2 \|\mathbf{e}_i\|_2^2 \geq \frac{\upsilon^2}{\tau} (\sum_{i=1}^{\tau} w_i \|\mathbf{e}_i\|_2)^2.$$

To see the second statement, consider the line (5.13) from the proof of Proposition 5.5.1. $\text{Proj}_{\mathscr{R}}(\mathbf{g}) = \sum_{i=1}^{\tau} \lambda_i \text{Proj}_{\mathscr{R}}(\mathbf{e}_i)$. On the other hand, column space of $\text{Proj}_{\mathscr{R}}(\mathbf{A}^T)$ is an m-dimensional random subspace of $\mathscr{R}$. If $m < \dim(\mathscr{R})$, $\text{Proj}_{\mathscr{R}}(\mathbf{g})$ is linearly independent with $\text{Proj}_{\mathscr{R}}(\mathbf{A}^T)$ with probability 1 and (5.13) will not hold. ■

In the next section, we will show how better choices of $\mathscr{R}$ (based on the decomposability assumption) can improve the lower bounds for S&L recovery.

5.6 Proofs for Section 5.2.2

Using the general framework provided in Section 5.2.1, in this section we present the proof of Theorem 5.2.3, which states various convex and nonconvex recovery results for the S&L models. We start with the proofs of the convex recovery.

5.6.1 Convex recovery results for *S&L*

In this section, we prove the statements of Theorem 5.2.3 regarding convex approaches, using Theorem 5.2.2 and Proposition 5.5.2. We will make use of the decomposable norms to obtain better lower bounds. Hence, we first state a result on the sign vectors and the supports of the S&L model following Lemma 5.5.4. The proof is provided in Appendix B.2.

Lemma 5.6.1 *Denote the norm* $\|\mathbf{X}^T\|_{1,2}$ *by* $\|\cdot^T\|_{1,2}$. *Given a matrix* $\mathbf{X}_0 \in \mathbb{R}^{d_1 \times d_2}$, *let* $\mathbf{E}_\star, \mathbf{E}_c, \mathbf{E}_r$ *and* $T_\star, T_c, T_r$ *be the sign vectors and supports for the norms* $\|\cdot\|_\star$, $\|\cdot\|_{1,2}$, $\|\cdot^T\|_{1,2}$ *respectively. Then,*

- $\mathbf{E}_\star, \mathbf{E}_r, \mathbf{E}_c \in T_\star \cap T_c \cap T_r$,
- $\langle \mathbf{E}_\star, \mathbf{E}_r \rangle \geq 0$, $\langle \mathbf{E}_\star, \mathbf{E}_c \rangle \geq 0$, *and* $\langle \mathbf{E}_c, \mathbf{E}_r \rangle \geq 0$.

5.6.1.1 Proof of Theorem 5.2.3: Convex cases

Proof of (a1) We use the functions $\|\cdot\|_{1,2}, \|\cdot^T\|_{1,2}$ and $\|\cdot\|_\star$ without the cone constraint, i.e., $\mathscr{C} = \mathbb{R}^{d_1 \times d_2}$. We will apply Proposition 5.5.2 with $\mathscr{R} = T_\star \cap T_c \cap T_r$. From Lemma 5.6.1 all the sign vectors lie on $\mathscr{R}$ and they have pairwise nonnegative inner products. Consequently, applying Proposition 5.5.3,

$$\rho(\mathscr{R}, \partial f(\mathbf{X}_0))^2 \geq \frac{1}{3} \min\{\frac{k_1}{d_1}, \frac{k_2}{d_2}, \frac{r}{\min\{d_1, d_2\}}\}$$

If $m < \dim(\mathscr{R})$, we have failure with probability 1. Hence, assume $m \geq \dim(\mathscr{R})$. Now, apply Proposition 5.5.2 with the given m_{low}.

Proof of (b1) In this case, we apply Lemma B.2.2. We choose $\mathscr{R} = T_\star \cap T_c \cap T_r \cap \mathbb{S}^n$, the norms are the same as in the general model, and $\upsilon \geq \frac{1}{\sqrt{2}}$. Also, pairwise inner products are positive, hence, using Proposition 5.5.3, $\rho(\mathscr{R}, \partial f(\mathbf{X}_0))^2 \geq \frac{1}{4} \min\{\frac{k}{d}, \frac{r}{d}\}$. Again, we may assume $m \geq \dim(\mathscr{R})$. Finally, based on Corollary B.1.1, for the PSD cone we have $\bar{\mathbf{D}}(\mathscr{C}) \geq \frac{3}{4}$. The result follows from Proposition 5.5.2 with the given m_{low}.

Proof of (c1) For PSD cone, $\bar{\mathbf{D}}(\mathscr{C}) \geq \frac{3}{4}$ and we simply use Theorem 5.2.2 to obtain the result by using $\kappa_{\ell_1}^2 = \frac{\|\bar{\mathbf{X}}_0\|_1^2}{d^2}$ and $\kappa_\star^2 = \frac{\|\bar{\mathbf{X}}_0\|_\star^2}{d}$.

5.6.1.2 Proof of Corollary 5.2.2

To show this, we will simply use Theorem 5.2.1 and will substitute κ's corresponding to ℓ_1 and the nuclear norm. $\kappa_\star = \frac{\|\bar{\mathbf{X}}_0\|_\star}{\sqrt{d}}$ and $\kappa_{\ell_1} = \frac{\|\bar{\mathbf{X}}_0\|_{\ell_1}}{d}$. Also observe that, $\lambda_{\ell_1} L_{\ell_1} = \beta d$ and $\lambda_\star L_\star = (1-\beta)d$. Hence, $\sum_{i=1}^2 \bar{\lambda}_i \kappa_i = \alpha \|\bar{\mathbf{X}}_0\|_1 + (1-\alpha)\|\bar{\mathbf{X}}_0\|_\star \sqrt{d}$. Use Proposition 5.3.1 to conclude with sufficiently small $c_1, c_2 > 0$.

5.6.2 Nonconvex recovery results for S&L

While Theorem 5.2.3 states the result for Gaussian measurements, we prove the nonconvex recovery for the more general sub-gaussian measurements. We first state a lemma that will be useful in proving the nonconvex results. The proof is provided in the Appendix B.3 and uses standard arguments.

Lemma 5.6.2 *Consider the set of matrices M in $\mathbb{R}^{d_1\times d_2}$ that are supported over an $s_1\times s_2$ submatrix with rank at most q. There exists a constant $c>0$ such that whenever $m\geq c\min\{(s_1+s_2)q, s_1\log\frac{d_1}{s_1}, s_2\log\frac{d_2}{s_2}\}$, with probability $1-2\exp(-cm)$, $\mathscr{A}(\cdot):\mathbb{R}^{d_1\times d_2}\to\mathbb{R}^m$ with i.i.d. zero-mean and isotropic sub-gaussian rows will satisfy the following,*

$$\mathscr{A}(\mathbf{X})\neq 0, \quad \textit{for all} \quad \mathbf{X}\in M. \tag{5.18}$$

5.6.2.1 Proof of Theorem 5.2.3: Nonconvex cases

Denote the sphere in $\mathbb{R}^{d_1\times d_2}$ with unit Frobenius norm by $\mathscr{S}^{d_1\times d_2}$.

Proof of (a2) Observe that the function $f(\mathbf{X})=\frac{\|\mathbf{X}\|_{0,2}}{\|\mathbf{X}_0\|_{0,2}}+\frac{\|\mathbf{X}^T\|_{0,2}}{\|\mathbf{X}_0^T\|_{0,2}}+\frac{\text{rank}(\mathbf{X})}{\text{rank}(\mathbf{X}_0)}$ satisfies the triangle inequality and we have $f(\mathbf{X}_0)=3$. Hence, if all null space elements $\mathbf{W}\in\text{Null}(\mathscr{A})$ satisfy $f(\mathbf{W})>6$, we have

$$f(\mathbf{X})\geq f(\mathbf{X}-\mathbf{X}_0)-f(-\mathbf{X}_0)>3,$$

for all feasible $\mathbf{X}$ which implies $\mathbf{X}_0$ being the unique minimizer.

Consider the set M of matrices, which are supported over a $6k_1\times 6k_2$ submatrix with rank at most $6r$. Observe that any $\mathbf{Z}$ satisfying $f(\mathbf{Z})\leq 6$ belongs to M. Hence ensuring $\text{Null}(\mathscr{A})\cap M=\{0\}$ would ensure $f(\mathbf{W})>6$ for all $\mathbf{W}\in\text{Null}(\mathscr{A})$. Since M is a cone, this is equivalent to $\text{Null}(\mathscr{A})\cap(M\cap\mathscr{S}^{d_1\times d_2})=\emptyset$. Now, applying Lemma 5.6.2 with set M and $s_1=6k_1$, $s_2=6k_2$, $q=6r$ we find the desired result.

Proof of (b2) Observe that due to the symmetry constraint,

$$f(\mathbf{X})=\frac{\|\mathbf{X}\|_{0,2}}{\|\mathbf{X}_0\|_{0,2}}+\frac{\|\mathbf{X}^T\|_{0,2}}{\|\mathbf{X}_0^T\|_{0,2}}+\frac{\text{rank}(\mathbf{X})}{\text{rank}(\mathbf{X}_0)}.$$

Hence, the minimization is the same as **(a2)**, the matrix is rank r contained in a $k\times k$ submatrix and we additionally have the positive semidefinite constraint which can only reduce the amount of required measurements compared to **(a2)**. Consequently, the result follows by applying Lemma 5.6.2, similar to **(a2)**.

Proof of (c2) Let $C=\{\mathbf{X}\neq 0\big|f(\mathbf{X})\leq f(\mathbf{X}_0)\}$. Since $\text{rank}(\mathbf{X}_0)=1$, if $f(\mathbf{X})\leq f(\mathbf{X}_0)=2$, $\text{rank}(\mathbf{X})=1$. With the symmetry constraint, this means $\mathbf{X}=\pm\mathbf{x}\mathbf{x}^T$ for some l-sparse $\mathbf{x}$. Observe that $\mathbf{X}-\mathbf{X}_0$ has rank at most 2 and is contained in a $2k\times 2k$ submatrix as $l\leq k$. Let M be the set of matrices that are symmetric and whose support lies in a $2k\times 2k$ submatrix. Using Lemma 5.6.2 with $q=2$, $s_1=s_2=2k$, whenever

$m \geq ck \log \frac{n}{k}$, with desired probability all nonzero $\mathbf{W} \in M$ will satisfy $\mathscr{A}(\mathbf{W}) \neq 0$. Consequently, any $\mathbf{X} \in C$ will have $\mathscr{A}(\mathbf{X}) \neq \mathscr{A}(\mathbf{X}_0)$, hence $\mathbf{X}_0$ will be the unique minimizer.

5.6.3 Existence of a matrix with large κ's

We now argue that, there exists an *S&L* matrix that have large $\kappa_{\ell_1}, \kappa_{\ell_{1,2}}$ and $\kappa_\star$ simultaneously. We will have a deterministic construction that is close to optimal. Our construction will be based on Hadamard matrices. $\mathbf{H}_n \in \mathbb{R}^{n \times n}$ is called a Hadamard matrix if it has ± 1 entries and orthogonal rows. Hadamard matrices exist for n that is an integer power of 2.

Using $\mathbf{H}_n$, our aim will be to construct a $d_1 \times d_2$ *S&L* (k_1, k_2, r) matrix $\mathbf{X}_0$ that satisfy $\|\bar{\mathbf{X}}_0\|_1^2 \approx k_1 k_2$, $\|\bar{\mathbf{X}}_0\|_\star^2 \approx r$, $\|\bar{\mathbf{X}}_0\|_{1,2}^2 \approx k_2$ and $\|\bar{\mathbf{X}}_0^T\|_{1,2}^2 \approx k_1$. To do this, we will construct a $k_1 \times k_2$ matrix and then plant it into a larger $d_1 \times d_2$ matrix. The following lemma summarizes the construction.

Lemma 5.6.3 *Without loss of generality, assume $k_2 \geq k_1 \geq r$. Let $\mathbf{H} := \mathbf{H}_{\lfloor \log_2 k_2 \rfloor}$. Let $\mathbf{X} \in \mathbb{R}^{k_1 \times k_2}$ be so that, i'th row of $\mathbf{X}$ is equal to $[i-1 \ (mod\ r)]+1$'th row of $\mathbf{H}$ followed by 0's for $1 \leq i \leq k_1$. Then,*

$$\|\bar{\mathbf{X}}_0\|_1^2 \geq \frac{k_1 k_2}{2},\ \|\bar{\mathbf{X}}_0\|_\star^2 \geq \frac{r}{2},\ \|\bar{\mathbf{X}}_0\|_{1,2}^2 \geq \frac{k_2}{2},\ \|\bar{\mathbf{X}}_0^T\|_{1,2}^2 = k_1.$$

In particular, if $k_1 \equiv 0 \ (mod\ r)$ and k_2 is an integer power of 2, then,

$$\|\bar{\mathbf{X}}_0\|_1^2 = k_1 k_2,\ \|\bar{\mathbf{X}}_0\|_\star^2 = r,\ \|\bar{\mathbf{X}}_0\|_{1,2}^2 = k_2,\ \|\bar{\mathbf{X}}_0^T\|_{1,2}^2 = k_1.$$

Proof: The left $k_1 \times 2^{\lfloor \log_2 k_2 \rfloor}$ entries of $\mathbf{X}$ are ± 1, and the remaining entries are 0. This makes the calculation of ℓ_1 and $\ell_{1,2}$ and Frobenius norms trivial.

In particular, $\|\mathbf{X}_0\|_F^2 = \|\mathbf{X}_0\|_1 = k_1 2^{\lfloor \log_2 k_2 \rfloor}$, $\|\mathbf{X}_0\|_{1,2} = \sqrt{k_1} 2^{\lfloor \log_2 k_2 \rfloor}$ and $\|\mathbf{X}_0^T\|_{1,2} = k_1 2^{\frac{\lfloor \log_2 k_2 \rfloor}{2}}$. Substituting these yield the results for these norms.

To lower bound the nuclear norm, observe that, each of the first r rows of the $\mathbf{H}$ are repeated at least $\lfloor \frac{k_1}{r} \rfloor$ times in $\mathbf{X}$. Combined with the orthogonality, this ensures that each singular value of $\mathbf{X}$ that is associated with the j'th row of $\mathbf{H}$ is at least $\sqrt{2^{\lfloor \log_2 k_2 \rfloor} \lfloor \frac{k_1}{r} \rfloor}$ for all $1 \leq j \leq r$. Consequently,

$$\|\mathbf{X}\|_\star \geq r \sqrt{2^{\lfloor \log_2 k_2 \rfloor} \lfloor \frac{k_1}{r} \rfloor}$$

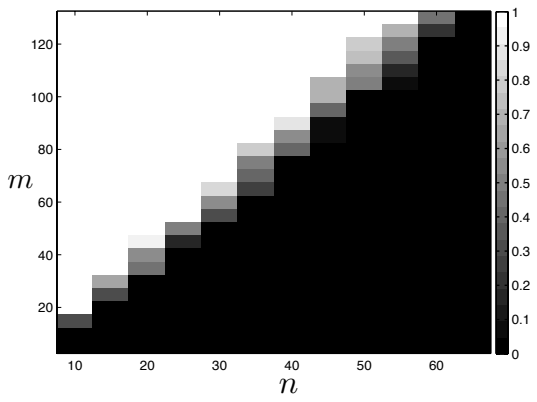


Figure 5.5: Performance of the recovery program minimizing $\max\{\frac{\mathrm{tr}(\mathbf{X})}{\mathrm{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_{1,2}}{\|\mathbf{X}_0\|_{1,2}}\}$ with a PSD constraint. The dark region corresponds to the experimental region of failure due to insufficient measurements. As predicted by Theorem 5.2.3, the number of required measurements increases linearly with rd.

Hence,

$$\|\bar{\mathbf{X}}\|_\star \geq \frac{r\sqrt{2^{\lfloor \log_2 k_2 \rfloor} \lfloor \frac{k_1}{r} \rfloor}}{\sqrt{k_1 2^{\lfloor \log_2 k_2 \rfloor}}} = \frac{r\sqrt{2^{\lfloor \log_2 k_2 \rfloor} \lfloor \frac{k_1}{r} \rfloor}}{\sqrt{2^{\lfloor \log_2 k_2 \rfloor}}} = r\sqrt{\frac{1}{k_1}\lfloor \frac{k_1}{r} \rfloor}$$

Use the fact that $\lfloor \frac{k_1}{r} \rfloor \geq \frac{k_1}{2r}$ as $k_1 \geq r$. ■

If we are allowed to use complex numbers, one can apply the same idea with the Discrete Fourier Transform (DFT) matrix. Similar to $\mathbf{H}_n$, DFT has orthogonal rows and its entries have the same absolute value. However, it exists for any $n \geq 1$; which would make the argument more concise.

5.7 Numerical Experiments

In this section, we numerically verify our theoretical bounds on the number of measurements for the Sparse and Low-rank recovery problem. We demonstrate the empirical performance of the weighted maximum of the norms f_{best} (see Lemma 5.5.1), as well as the weighted sum of norms.

The experimental setup is as follows. Our goal is to explore how the number of required measurements m scales with the size of the matrix d. We consider a grid of (m,d) values, and generate at least 100 test instances for each grid point (in the boundary areas, we increase the number of instances to at least 200).

We generate the target matrix $\mathbf{X}_0$ by generating a $k \times r$ i.i.d. Gaussian matrix $\mathbf{G}$, and inserting the $k \times k$ matrix $\mathbf{G}\mathbf{G}^T$ in an $d \times d$ matrix of zeros. We take $r = 1$ and $k = 8$ in all of the following experiments; even with these small values, we can observe the scaling predicted by our bounds. In each test, we measure the

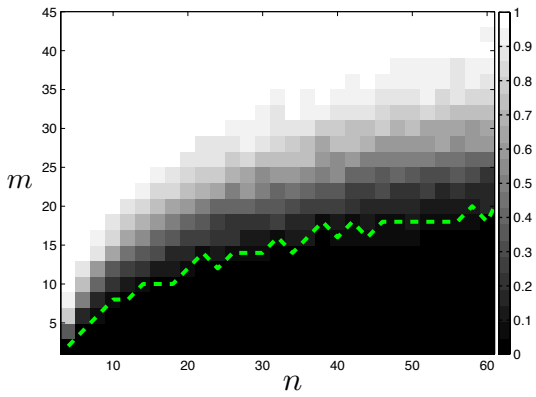


Figure 5.6: Performance of the recovery program minimizing $\max\{\frac{\text{tr}(\mathbf{X})}{\text{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$ with a PSD constraint. $r = 1, k = 8$ and d is allowed to vary. The plot shows m versus d to illustrate the lower bound $\Omega(\min\{k^2, dr\})$ predicted by Theorem 5.2.3.

normalized recovery error $\frac{\|\mathbf{X}-\mathbf{X}_0\|_F}{\|\mathbf{X}_0\|_F}$ and declare successful recovery when this error is less than 10^{-4}. The optimization programs are solved using the CVX package [113], which calls the SDP solver SeDuMi [196].

We first test our bound in part (b) of Theorem 5.2.3, $\Omega(rd)$, on the number of measurements for recovery in the case of minimizing $\max\{\frac{\text{tr}(\mathbf{X})}{\text{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_{1,2}}{\|\mathbf{X}_0\|_{1,2}}\}$ over the set of positive semi-definite matrices. Figure 5.5 shows the results, which demonstrates m scaling linearly with d (note that $r = 1$).

Next, we replace $\ell_{1,2}$ norm with ℓ_1 norm and consider a recovery program that emphasizes entry-wise sparsity rather than block sparsity. Figure 5.6 demonstrates the lower bound $\Omega(\min\{k^2, d\})$ in Part (c) of Theorem 5.2.3 where we attempt to recover a rank-1 positive semi-definite matrix $\mathbf{X}_0$ by minimizing $\max\{\frac{\text{tr}(\mathbf{X})}{\text{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$ subject to the measurements and a PSD constraint. The green curve in the figure shows the empirical 95% failure boundary, depicting the region of failure with high probability that our results have predicted. It starts off growing linearly with d, when the term rd dominates the term k^2, and then saturates as d grows and the k^2 term (which is a constant in our experiments) becomes dominant.

The penalty function $\max\{\frac{\text{tr}(\mathbf{X})}{\text{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$ depends on the norm of $\mathbf{X}_0$. In practice the norm of the solution is not known beforehand, a weighted sum of norms is used instead. In Figure 5.7 we examine the performance of the weighted sum of norms penalty in recovery of a rank-1 PSD matrix, for different weights. We pick $\lambda = 0.20$ and $\lambda = 0.35$ for a randomly generated matrix $\mathbf{X}_0$, and it can be seen that we get a reasonable result which is comparable to the performance of $\max\{\frac{\text{tr}(\mathbf{X})}{\text{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$.

In addition, we consider the *amount of error* in the recovery when the program fails. Figure 5.8 shows two curves below which we get a 90% percent failure, where for the green curve the normalized error

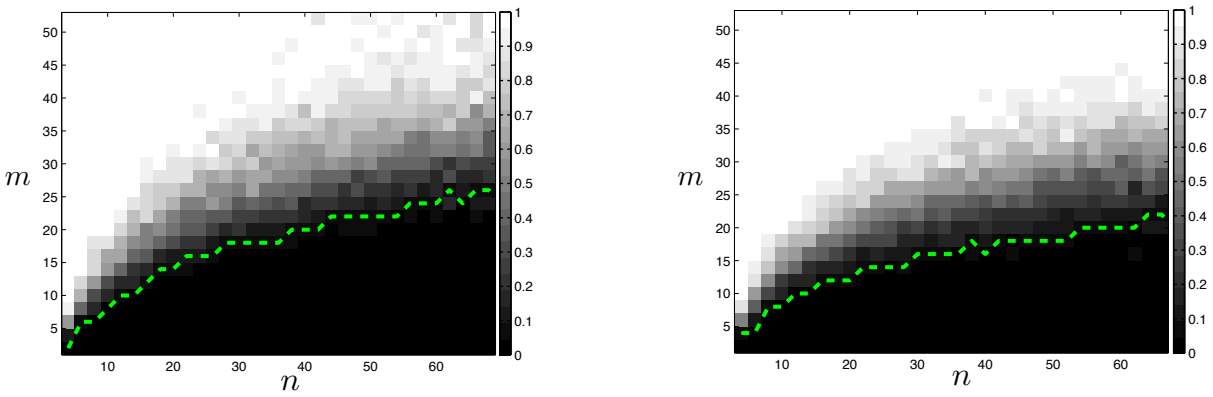


Figure 5.7: Performance of the recovery program minimizing $\mathrm{tr}(\mathbf{X}) + \lambda \|\mathbf{X}\|_1$ with a PSD constraint, for $\lambda = 0.2$ (left) and $\lambda = 0.35$ (right).

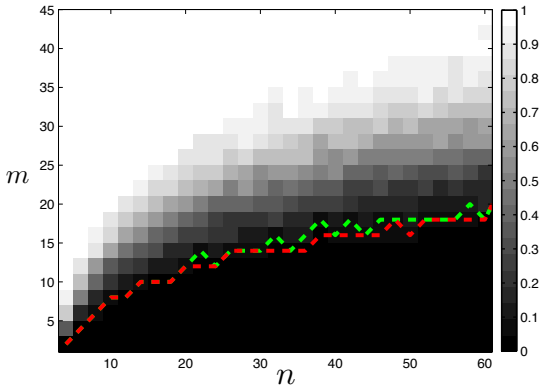


Figure 5.8: 90% frequency of failure where the threshold of recovery is 10^{-4} for the green and 0.05 for the red curve. $\max\{\frac{\mathrm{tr}(\mathbf{X})}{\mathrm{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$ is minimized subject to the PSD constraint and the measurements.

threshold for declaring failure is 10^{-4}, and for the red curve it is a larger value of 0.05. We minimize $\max\{\frac{\mathrm{tr}(\mathbf{X})}{\mathrm{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$ as the objective. We observe that when the recovery program has an error, it is very likely that this error is large, as the curves for 10^{-4} and 0.05 almost overlap. Thus, when the program fails, it fails badly. This observation agrees with intuition from similar problems in compressed sensing where sharp phase transition is observed.

As a final comment, observe that, in Figures 5.6, 5.7 and 5.8 the required amount of measurements slowly increases even when d is large and $k^2 = 64$ is the dominant constant term. While this is consistent with our lower bound of $\Omega(k^2, d)$, the slow increase for constant k, can be explained by the fact that, as d gets larger, sparsity becomes the dominant structure and ℓ_1 minimization by itself requires $\mathcal{O}\left(k^2 \log \frac{d}{k}\right)$

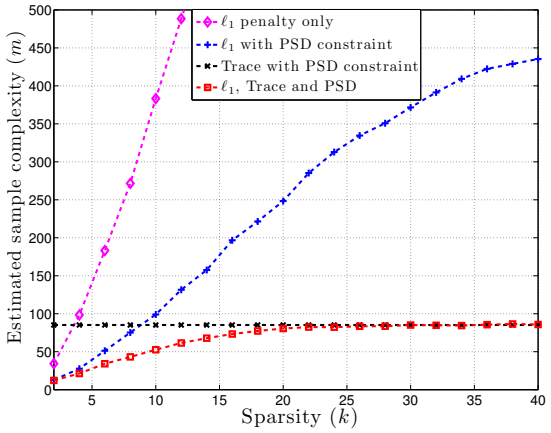


Figure 5.9: We compare sample complexities of different approaches for a rank 1, 40×40 matrix as function of sparsity. The sample complexities were estimated by a search over m, where we chose the m with success rate closest to 50% (over 100 iterations).

measurements rather than $\mathcal{O}\left(k^2\right)$. Hence for large d, the number of measurements can be expected to grow logarithmically in d.

In Figure 5.9, we compare the estimated phase transition points for different approaches for varying sparsity levels. The algorithms we compare are,

- Minimize ℓ_1 norm,
- Minimize ℓ_1 norm subject to the positive-semidefinite constraint,
- Minimize trace norm subject to the positive-semidefinite constraint,
- Minimize $\max\{\frac{\mathrm{tr}(\mathbf{X})}{\mathrm{tr}(\mathbf{X}_0)}, \frac{\|\mathbf{X}\|_1}{\|\mathbf{X}_0\|_1}\}$ subject to the positive-semidefinite constraint

Not surprisingly, the last option outperforms the rest in all cases. On the other hand, its performance is highly comparable to the minimum of the second and third approaches. For all regimes of sparsity, we observe that, measurements required by the last method is at least half as much as the minimum of second and third methods.

5.8 Discussion

We have considered the problem of recovery of a simultaneously structured object from limited measurements. It is common in practice to combine known norm penalties corresponding to the individual structures (also known as regularizers in statistics and machine learning applications), and minimize this combined objective in order to recover the object of interest. The common use of this approach motivated us to analyze

its performance, in terms of the smallest number of generic measurements needed for correct recovery. We showed that, under a certain assumption on the norms involved, the combined penalty requires more generic measurements than one would expect based on the degrees of freedom of the desired object. Our lower bounds on the required number of measurements implies that the combined norm penalty cannot perform significantly better than the best individual norm.

These results raise several interesting questions, and lead to directions for future work. We briefly outline some of these directions, as well as connections to some related problems.

Defining new atoms for simultaneously structured models. Our results show that combinations of individual norms do not exhibit a strong recovery performance. On the other hand, the paper [50] proposes a remarkably general construction for an appropriate penalty given a set of atoms. Can we revisit a simultaneously structured recovery problem, and define new atoms that capture all structures at the same time? And can we obtain a new norm penalty induced by the convex hull of the atoms? Abstractly, the answer is yes, but such convex hulls may be hard to characterize, and the corresponding penalty may not be efficiently computable. It is interesting to find special cases where this construction can be carried out and results in a tractable problem. Recent developments in this direction include the "square norm" proposed by [158] for the low-rank tensor recovery; which provably outperforms (5.2) for Gaussian measurements and the (k,q)-trace norm introduced by Richard et al. to estimate S&L matrices [180].

Algorithms for minimizing combination of norms. Despite the limitation in their theoretical performance, in practice one may still need to solve convex relaxations that combine the different norms, i.e., problem (5.3). Consider the special case of sparse and low-rank matrix recovery. All corresponding optimization problems mentioned in Theorem 5.2.3 can be expressed as a semidefinite program and solved by standard solvers; for example, for the numerical experiments in Section 5.7 we used the interior-point solver SeDuMi [196] via the modeling environment CVX [113]. However, interior point methods do not scale for problems with tens of thousands of matrix entries, which are common in machine learning applications. One future research direction is to explore first-order methods, which have been successful in solving problems with a single structure (for example ℓ_1 or nuclear norm regularization alone). In particular, Alternating Directions Methods of Multipliers (ADMM) appears to be a promising candidate.

Connection to Sparse PCA. The sparse PCA problem (see, e.g. [64, 124, 236]) seeks sparse principal components given a (possibly noisy) data matrix. Several formulations for this problem exist, and many algorithms have been proposed. In particular, a popular algorithm is the SDP relaxation proposed in [64], which is based on the following formulation.

For the first principal component to be sparse, we seek an $\mathbf{x} \in \mathbb{R}^n$ that maximizes $\mathbf{x}^T\mathbf{A}\mathbf{x}$ for a given data matrix $\mathbf{A}$, and minimizes $\|\mathbf{x}\|_0$. Similar to the sparse phase retrieval problem, this problem can be reformulated in terms of a rank-1, PSD matrix $\mathbf{X} = \mathbf{x}\mathbf{x}^T$ which is also row- and column-sparse. Thus we seek a simultaneously low-rank and sparse $\mathbf{X}$. This problem is different from the recovery problem studied in this chapter, since we do not have m random measurements of $\mathbf{X}$. Yet, it will be interesting to connect this chapter's results to the sparse PCA problem to potentially provide new insights for sparse PCA.

Chapter 6

Graph Clustering via Low-Rank and Sparse Decomposition

Given an unweighted graph, finding nodes that are well-connected with each other is a very useful problem with applications in social networks [69, 99], data mining [228], bioinformatics [229, 230], computer networks, sensor networks. Different versions of this problem have been studied as graph clustering [52,98,106,185], correlation clustering [10,66,108], graph partitioning on planted partition model [20]. Developments in convex optimization techniques to recover low-rank matrices [36,40,43,50,51] via nuclear norm minimization has recently led to the development of several convex algorithms to recover clusters in a graph [3,5–7,61,62,164,223][1].

Let us assume that a given graph has dense clusters; we can look at its adjacency matrix as a low-rank matrix with sparse noise. That is, the graph can be viewed as a union of cliques with some edges missing inside the cliques and extra edges between the cliques. Our aim is to recover the low-rank matrix since it is equivalent to finding clusters. A standard approach is the following convex program which decomposes the adjacency matrix ($\mathbf{A}$) as the sum of a low-rank ($\mathbf{L}$) and a sparse ($\mathbf{S}$) component.

Simple Convex Program:

$$\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \;\; \|\mathbf{L}\|_\star + \lambda \|\mathbf{S}\|_1 \tag{6.1}$$

subject to

$$1 \geq \mathbf{L}_{i,j} \geq 0 \;\; \text{for all } i,j \in \{1,2,\ldots n\} \tag{6.2}$$

$$\mathbf{L} + \mathbf{S} = \mathbf{A}$$

[1] This chapter is based on the works [164,214,215].

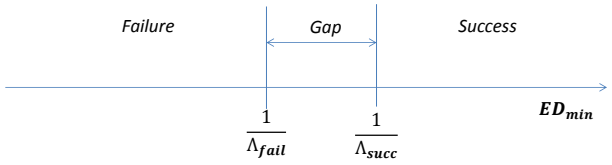


Figure 6.1: Feasibility of Program 6.1 in terms of the minimum effective density (ED_{min}).

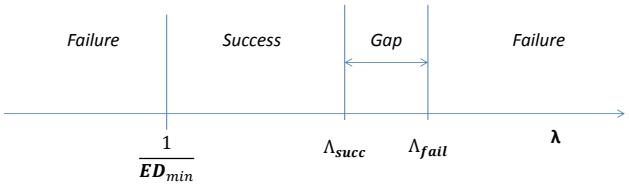


Figure 6.2: Feasibility of Program 6.1 in terms of the regularization parameter (λ).

Figure 6.3: Characterization of the feasibility of Program (6.1) in terms of the minimum effective density and the value of the regularization parameter. The feasibility is determined by the values of these parameters in comparison with two constants Λ_{succ} and Λ_{fail}, derived in Theorem 6.1 and Theorem 6.2. The thresholds guaranteeing the success or failure of Program 6.1 derived in this chapter are fairly close to each other.

where $\lambda > 0$ is a regularization parameter. This program is very intuitive and requires the knowledge of only the adjacency matrix. Program 6.1 has been studied in several works [36, 51, 61, 62], including [164] which is due to the author.

While (6.1) requires only the knowledge of adjacency matrix, it is not difficult to see that, when the edge probability inside the cluster is $p < 1/2$, (as $n \to \infty$) Program 6.1 will return $\mathbf{L}^0 = 0$ as the optimal solution (since if the cluster is not dense completing the missing edges is more costly compared to treating the cluster as sparse). As a result our analysis of Program 6.1, and the corresponding Theorems 6.1 and 6.2, assumes $p > 1/2$. Clearly, there are many instances of graphs we would like to cluster where $p < 1/2$, most notably social networking. If the total size of the cluster region (i.e, the total number of edges in the cluster, denoted by $|\mathscr{R}|$) is known, then the following convex program can be used, and can be shown to work for $p < 1/2$ (see Theorem 6.3).

Improved Convex Program:

$$\underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \quad \|\mathbf{L}\|_\star + \lambda \|\mathbf{S}\|_1 \tag{6.3}$$

subject to

$$1 \geq \mathbf{L}_{i,j} \geq \mathbf{S}_{i,j} \geq 0 \quad \text{for all } i,j \in \{1,2,\ldots n\} \tag{6.4}$$

$$\mathbf{L}_{i,j} = \mathbf{S}_{i,j} \text{ whenever } \mathbf{A}_{i,j} = 0 \tag{6.5}$$

$$\text{sum}(\mathbf{L}) \geq |\mathscr{R}| \tag{6.6}$$

As before, $\mathbf{L}$ is the low-rank matrix corresponding to the ideal cluster structure and $\lambda \geq 0$ is the regularization parameter. However, $\mathbf{S}$ will now correspond to the sparse error matrix that accounts only for the missing edges inside the clusters. This program was first proposed by the author in [164]. The similar approaches include work by Ames [5, 7].

If $\mathscr{R}$ is not known, it is possible to solve Problem 6.3 for several values of $\mathscr{R}$ until the desired performance is obtained. Intuitively, as the right hand side of (6.6) is increased, Improved Program should return a larger group of clusters. Hence, one can start from a small estimate of $|\mathscr{R}|$ and increase it as long as he gets a reasonable performance. Our analytic results, however, will assume the exact knowledge of $|\mathscr{R}|$.

6.0.1 Our Contributions

- We define the "effective density" of a cluster (which depends on the program we solve). Effective density is a function of the size and density of the cluster as well as inter-cluster edges. Our results are in terms of the effective density, i.e. when a cluster has large effective density, it is recoverable via the proposed convex approaches.
- We analyze the Simple Convex Program 6.1 for the SBM. We provide *explicit bounds* on the regularization parameter as a function of the parameters of the SBM, that characterizes the success and failure conditions of Program 6.1 (see results in Section 6.2.1). Our success and failure conditions show a good match (the gap is a constant factor of 4); hence our analysis is helpful in understanding the **phase transition** from failure to success for the simple approach.
- We also analyze the Improved Convex Program 6.3. We explicitly characterize the conditions on the parameters of the SBM and the regularization parameter for successfully recovering clusters using this approach (see results in Section 6.2.3). Our bounds not only reflect the correct relation between the problem parameters; but also have small constants in front.
- Our findings are shown to match well with the numerical experiments.

We consider the popular *stochastic block model* (also called the planted partition model) for the graph. Under this model of generating random graphs, the existence of an edge between any pair of vertices is independent of the other edges. The probability of the existence of an edge is identical within any individual cluster, but may vary across clusters. One may think of this as a heterogeneous form of the Erdös-Renyi model. We characterize the conditions under which Programs 6.1 and 6.3 can successfully recover the correct clustering, and when it cannot. Our analysis reveals the dependence of its success on a metric that

we term the *minimum effective density* of the graph. While defined more formally later in the chapter, in a nutshell, the minimum effective density of a random graph tries to capture the density of edges in the sparsest cluster. We derive explicit upper and lower bounds on the value of this metric that determine the success or failure of Program 6.1 (as illustrated in Fig. 6.1).

A second contribution of this chapter is to explicitly characterize the efficacy of Programs 6.1 and 6.3 with respect to the regularization parameter λ. We obtain bounds on the values of λ that permit the recovery of the clusters, or those that necessitate Program 6.1 to fail (as illustrated in Fig. 6.2). Our results thus lead to a more principled approach towards the choice of the regularization parameter for the problem at hand.

Most of the convex algorithms proposed for graph clustering, for example, the recent works by Xu et al. [223], Ames and Vavasis [6, 7], Jalali et al. [61], Chen et al. [62], Ames [5], Ailon et al. [3] are variants of Program 6.1. These results show that planted clusters can be identified via tractable convex programs as long as the cluster size is proportional to the square-root of the size of the adjacency matrix. However, the exact requirements on the cluster size are not known. In this chapter, we find sharp bounds for the identifiability as a function of cluster sizes, inter cluster density and intra cluster density. To the best of our knowledge, this is the first explicit characterization of the feasibility of the convex optimization based approaches (6.1) and (6.3) towards this problem.

The rest of the chapter is organized as follows. Section 6.1 formally introduces the model considered in this chapter. Section 6.2.1 presents the main results of the chapter: an analytical characterization of the feasibility of the low rank plus sparse based approximation for identifying clusters. Section 6.3 presents simulations that corroborate our theoretical results. Finally, the proofs of the technical results are deferred to Sections C.1, C.2 and C.3.

6.1 Model

For any positive integer m, let $[m]$ denote the set $\{1, 2, \ldots, m\}$. Let $\mathbf{G}$ be an unweighted graph on n nodes, $[n]$, with K disjoint (dense) clusters. Let $\mathscr{C}_i$ denote the set of nodes in the i^{th} cluster. Let n_i denote the size of the i^{th} cluster, i.e., the number of nodes in $\mathscr{C}_i$. We shall term the set of nodes that do not fall in any of these K clusters as *outliers* and denote them as $\mathscr{C}_{K+1} := [n] - \bigcup_{i=1}^{K} \mathscr{C}_i$. The number of outliers is thus $n_{K+1} := n - \sum_{i=1}^{K} n_i$. Since the clusters are assumed to be disjoint, we have $\mathscr{C}_i \cap \mathscr{C}_j = \emptyset$ for all $i, j \in [n]$.

Let $\mathscr{R}$ be the region corresponding to the union of regions induced by the clusters, i.e., $\mathscr{R} = \bigcup_{i=1}^{K} \mathscr{C}_i \times \mathscr{C}_i \subseteq [n] \times [n]$. So, $\mathscr{R}^c = [n] \times [n] - \mathscr{R}$ is the region corresponding to out of cluster regions. Note that

$|\mathscr{R}| = \sum_{i=1}^{K} n_i^2$ and $|\mathscr{R}^c| = n^2 - \sum_{i=1}^{K} n_i^2$. Let $n_{min} := \min_{1 \le i \le K} n_i$.

Let $\mathbf{A} = \mathbf{A}^T$ denote the adjacency matrix of the graph $\mathbf{G}$. The diagonal entries of $\mathbf{A}$ are 1. The adjacency matrix will follow a probabilistic model, in particular, a more general version of the popular stochastic block model [63, 117].

Definition 6.1 (Stochastic Block Model) *Let $\{p_i\}_{i=1}^{K}, q$ be constants between 0 and 1. Then, a random graph $\mathbf{G}$, generated according to stochastic block model, has the following adjacency matrix. Entries of $\mathbf{A}$ on the lower triangular part are independent random variables and for any $i > j$:*

$$\mathbf{A}_{i,j} = \begin{cases} Bernoulli(p_l) & \text{if both } \{i,j\} \in \mathscr{C}_l \text{ for some } l \le K \\ Bernoulli(q) & \text{otherwise.} \end{cases}$$

So, an edge inside i^{th} cluster exists with probability p_i and an edge outside the clusters exists with probability q. Let $p_{min} := \min_{1 \le i \le K} p_i$. We assume that the clusters are dense and the density of edges inside clusters is greater than outside, i.e., $p_{\min} > \frac{1}{2} > q > 0$. We note that the Program 6.1 does not require the knowledge of $\{p_i\}_{i=1}^{K}, q$ or K, and uses only the adjacency matrix $\mathbf{A}$ for its operation. However, the knowledge of $\{p_i\}_{i=1}^{K}, q$ will help us tune λ in a better way.

6.2 Main Results

6.2.1 Results on the Simple Convex Program

The desired solution to Program 6.1 is $(\mathbf{L}^0, \mathbf{S}^0)$ where $\mathbf{L}^0$ corresponds to the full cliques, when missing edges inside $\mathscr{R}$ are completed, and $\mathbf{S}^0$ corresponds to the missing edges and the extra edges between the clusters. In particular we want:

$$\mathbf{L}^0_{i,j} = \begin{cases} 1 & \text{if both } \{i,j\} \in \mathscr{C}_l \text{ for some } l \le K, \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

$$\mathbf{S}^0_{i,j} = \begin{cases} -1 & \text{if both } \{i,j\} \in \mathscr{C}_l \text{ for some } l \le K, \text{ and } \mathbf{A}_{i,j} = 0, \\ 1 & \text{if } \{i,j\} \text{ are not in the same cluster and } \mathbf{A}_{i,j} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that the $(\mathbf{L}^0, \mathbf{S}^0)$ pair is feasible. We say that Program 6.1 *succeeds* when $(\mathbf{L}^0, \mathbf{S}^0)$ is the optimal solution to Program 6.1. In this section we present two theorems which give the conditions under which Program 6.1 succeeds or fails.

The following definitions are critical to our results.

- Define $\mathbf{ED}_i := n_i\,(2p_i - 1)$ as the effective density of cluster $\mathscr{C}_i$ and $\mathbf{ED}_{\min} = \min_{1 \le i \le K} \mathbf{ED}_i$.
- Let $\gamma_{\text{succ}} := \max_{1 \le i \le K} 4\sqrt{(q(1-q) + p_i(1-p_i))n_i}$,
  $\gamma_{\text{fail}} := \sum_{i=1}^{K} \frac{n_i^2}{n}$
- $\Lambda_{\text{fail}} := \frac{1}{\sqrt{q(n-\gamma_{\text{fail}})}}$ and $\Lambda_{\text{succ}} := \frac{1}{4\sqrt{q(1-q)n} + \gamma_{\text{succ}}}$.

Theorem 6.1 *Let* $\mathbf{G}$ *be a random graph generated according to the Stochastic Block Model 6.1 with* K *clusters of sizes* $\{n_i\}_{i=1}^K$ *and probabilities* $\{p_i\}_{i=1}^K$ *and* q*, such that* $p_{min} > \frac{1}{2} > q > 0$*. Given* $\varepsilon > 0$*, there exists positive constants* δ, c_1, c_2 *such that,*

1. *For any given* $\lambda \ge 0$*, if* $\mathbf{ED}_{\min} \le (1-\varepsilon)\Lambda_{fail}^{-1}$ *then Program 6.1 fails with probability* $1 - c_1 \exp(-c_2 |\mathscr{R}^c|)$.
2. *Whenever* $\mathbf{ED}_{\min} \ge (1+\varepsilon)\Lambda_{succ}^{-1}$*, for* $\lambda = (1-\delta)\Lambda_{succ}$*, Program 6.1 succeeds with probability* $1 - c_1 n^2 \exp(-c_2 n_{\min})$.

As it will be discussed in Sections C.1 and C.2, Theorem 6.1 is actually a special case of the following result, which characterizes success and failure as a function of λ.

Theorem 6.2 *Let* $\mathbf{G}$ *be a random graph generated according to the Stochastic Block Model 6.1 with* K *clusters of sizes* $\{n_i\}_{i=1}^K$ *and probabilities* $\{p_i\}_{i=1}^K$ *and* q*, such that* $p_{min} > \frac{1}{2} > q > 0$*. Given* $\varepsilon > 0$*, there exists positive constants* c_1', c_2' *such that,*

1. *If* $\lambda \ge (1+\varepsilon)\Lambda_{fail}$*, then Program 6.1 fails with probability* $1 - c_1' \exp(-c_2' |\mathscr{R}^c|)$.
2. *If* $\lambda \le (1-\varepsilon)\Lambda_{succ}$ *then,*
   - *If* $\mathbf{ED}_{\min} \le (1-\varepsilon)\frac{1}{\lambda}$*, then Program 6.1 fails with probability* $1 - c_1' \exp(-c_2' n_{\min})$.
   - *If* $\mathbf{ED}_{\min} \ge (1+\varepsilon)\frac{1}{\lambda}$*, then Program 6.1 succeeds with probability* $1 - c_1' n^2 \exp(-c_2' n_{\min})$.

We see that the minimum effective density $\mathbf{ED}_{\min}, \Lambda_{\text{succ}}$ and Λ_{fail} play a fundamental role in determining the success of Program 6.1. Theorem 6.1 gives a criteria for the inherent success of Program 6.1, whereas Theorem 6.2 characterizes the conditions for the success of Program 6.1 as a function of the regularization parameter λ. We illustrate these results in Figures 6.1 and 6.2.

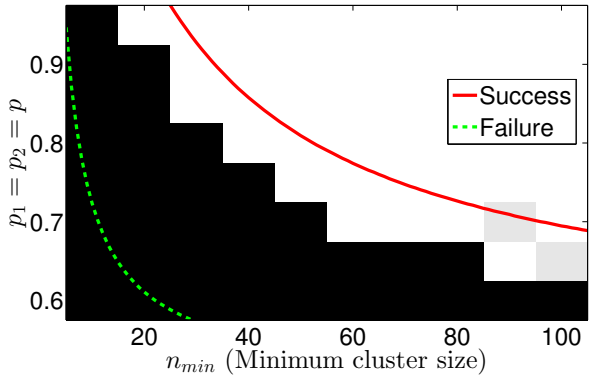


Figure 6.4: Simulation results showing the region of success (white region) and failure (black region) of Program 6.1 with $\lambda = 0.99\Lambda_{\text{succ}}$. Also depicted are the thresholds for success (solid red curve on the top-right) and failure (dashed green curve on the bottom-left) predicted by Theorem 6.1.

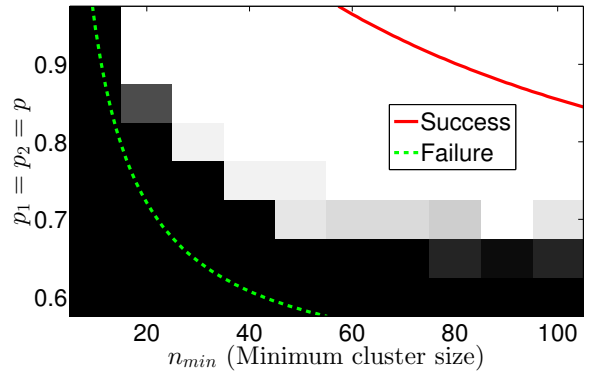


Figure 6.5: Simulation results showing the region of success (white region) and failure (black region) of Program 6.1 with $\lambda = 2\mathbf{ED}_{\text{min}}^{-1}$. Also depicted are the thresholds for success (solid red curve on the top-right) and failure (dashed green curve on the bottom-left) predicted by Theorem 6.2.

6.2.2 Sharp Performance Bounds

From our forward and converse results, we see that there is a gap between Λ_{fail} and Λ_{succ}. The gap is $\frac{\Lambda_{\text{fail}}}{\Lambda_{\text{succ}}} = \frac{4\sqrt{q(1-q)n}+\gamma_{\text{succ}}}{\sqrt{q(n-\gamma_{\text{fail}})}}$ times. In the small cluster regime where $\max_{1\le i\le K} n_i = o(n)$ and $\sum_{i=1}^{K} n_i^2 = o(n^2)$, the ratio $\frac{\Lambda_{\text{fail}}}{\Lambda_{\text{succ}}}$ takes an extremely simple form as we have $\gamma_{\text{fail}} \ll n$ and $\gamma_{\text{succ}} \ll \sqrt{n}$. In particular, $\frac{\Lambda_{\text{fail}}}{\Lambda_{\text{succ}}} = 4\sqrt{1-q} + o(1)$, which is at most 4 times in the worst case.

6.2.3 Results on the Improved Convex Program

The following definitions are critical to describe our results.

- Define $\tilde{\mathbf{ED}}_i := n_i\,(p_i - q)$ as the effective density of cluster i and $\tilde{\mathbf{ED}}_{\text{min}} = \min_{1\le i\le K} \tilde{\mathbf{ED}}_i$.

- $\tilde{\gamma}_{\text{succ}} := 2 \max_{1 \le i \le K} \sqrt{n_i} \sqrt{(1-p_i)p_i + (1-q)q}$
- $\tilde{\Lambda}^{-1}_{\text{succ}} := 2\sqrt{nq(1-q)} + \tilde{\gamma}_{\text{succ}}$.

We note that the threshold, $\tilde{\Lambda}_{\text{succ}}$ depends only on the parameters of the model.

Theorem 6.3 (Improved Program) *Consider a random graph generated according to the SBM of Definition 6.1, with adjacency matrix* $\mathbf{A}$*, K disjoint clusters of sizes* $\{n_i\}_{i=1}^K$*, and probabilities* $\{p_i\}_{i=1}^K$ *and q, such that* $p_{min} > q > 0$*. Further assume each edge of* $\mathbf{A}$ *is independently observed with probability r. Given* $\varepsilon > 0$*, there exists positive constants* c_1', c_2' *such that: If* $0 < \lambda \le (1-\varepsilon)\tilde{\Lambda}_{succ}$ *and* $\tilde{\mathbf{ED}}_{\min} \ge (1+\varepsilon)\frac{1}{\lambda}$*, then Program 6.3 succeeds in recovering the clusters with probability* $1 - c_1' n^2 \exp(-c_2' n_{\min})$.

Discussion:

1. Theorem 6.3 gives a sufficient condition for the success of Program 6.3 as a function of λ. In particular, for any $\lambda > 0$, we succeed if $\tilde{\mathbf{ED}}^{-1}_{\min} < \lambda < \tilde{\Lambda}_{\text{succ}}$

2. **Small Cluster Regime:**

   When $n_{\max} = o(n)$, we have $\tilde{\Lambda}^{-1}_{\text{succ}} = 2\sqrt{nq(1-q)}$. For simplicity let $p_i = p$, $\forall\, i$, which yields $\tilde{\mathbf{ED}}_{\min} = n_{\min}(p-q)$. Then $\tilde{\mathbf{ED}}_{\min} > \tilde{\Lambda}^{-1}_{\text{succ}}$ implies,

   $$n_{\min} > \frac{2\sqrt{nq(1-q)}}{p-q}, \tag{6.8}$$

   which gives a lower bound on the minimum cluster size that is sufficient for success. This lower bound on the minimum cluster size will increase as the noise term q increases or as the clusters get sparser (smaller p) which is intuitive.

Note: The proofs for Theorems 6.1, 6.2 and 6.3 are provided in Appendices C.1, C.2 and C.3.

6.2.4 Comparison to the literature on graph theory

While we approached the clustering problem from a convex optimization view point, it is crucial that our results compare well with the related literature. Fixing densities, and focusing on the minimum recoverable cluster size, we find that, we need $n_{\min} = O(\sqrt{n})$ for the programs to be successful. This compares well with the convex optimization based results of [5–7, 61, 62, 164].

Now, let us allow the cluster sizes to be $O(n)$ and investigate how small $p = p_1 = \cdots = p_K$ and q can be. For comparison, we will make use of McSherry's result which is based on spectral techniques [146].

Corollary 1 of [146] essentially states that, their proposed algorithm will succeed if,

$$\frac{p-q}{p} > c\sqrt{\frac{\log n}{pn}} \iff \sqrt{\frac{n}{\log n}} > c\frac{\sqrt{p}}{p-q}. \tag{6.9}$$

In this regime, we have to take $\tilde{\gamma}_{\text{succ}}$ term in the account as n_i, n is comparable. Hence, our bound yields the condition,

$$n_i = O(n) > c'\frac{\sqrt{pn}}{p-q}. \tag{6.10}$$

Ignoring the $\log n$ term in (6.9), it can be seen that the two bounds are identical. We remark that the extra $\log n$ term in (6.9) and the minimum density of $p \sim O(\frac{\log n}{n})$ is indeed necessary. For instance, to be able to identify a cluster, we require it to be connected, which requires $p \sim O(\frac{\log n}{n})$ [93]. The reason it does not exist in our bounds is because we assume p, q are constants independent of n. We believe $p, q \sim O(\frac{\log n}{n})$ regime can be handled as well, however, our arguments should be modified to accommodate sparse random matrices, which will change the spectral norm estimates in our proof.

6.3 Simulations

We implement Program 6.1 using the inexact augmented Lagrangian multiplier method algorithm by Lin et al. [136]. We note that this algorithm solves the program approximately. Moreover, numerical imprecision prevents the output of the algorithm from being strictly 1 or 0. Hence we round each entry to 1 or 0 by comparing it with the mean of all entries of the output. In other words, if an entry is greater than the overall mean, we round it to 1 and to 0 otherwise. We declare success if the number of entries that are wrong in the rounded output compared to $\mathbf{L}^0$ (recall from (6.7)) is less than 0.1%.

We consider the set up with $n = 200$ nodes and two clusters of equal sizes, $n_1 = n_2$. We vary the cluster sizes from 10 to 100 in steps of 10. We fix $q = 0.1$ and vary the probability of edge inside clusters $p_1 = p_2 = p$ from 0.6 to 0.95 in steps of 0.05. We run the experiments 20 times and average over the outcomes. In the first set of experiments, we run the program with $\lambda = 0.99\Lambda_{succ}$ which ensures that $\lambda < \Lambda_{\text{succ}}$. Figure 6.4 shows the region of success (white region) and failure (black region) for this experiment. From Theorem 6.1, we expect the program to succeed when $\mathbf{ED}_{\text{min}} > \Lambda_{\text{succ}}^{-1}$, which is the region above the solid red curve in Figure 6.4, and fail when $\mathbf{ED}_{\text{min}} < \Lambda_{\text{fail}}^{-1}$, which is the region below the dashed green curve in Figure 6.4.

In the second set of experiments, we run the program with $\lambda = \frac{2}{\mathbf{ED}_{\min}}$. This ensures that $\mathbf{ED}_{\min} > \frac{1}{\lambda}$. Figure 6.5 shows the region of success (white region) and failure (black region) for this experiment. From Theorem 6.2, we expect the program to succeed when $\lambda < \Lambda_{\text{succ}}$ which is the region above the solid red curve in Figure 6.5 and fail when $\lambda > \Lambda_{\text{fail}}$ which is the region below the dashed green curve in Figure 6.5.

We see that the transition indeed happens between the solid red curve and the dashed green curve in both Figure 6.4 and Figure 6.5 as predicted by Theorem 6.1 and Theorem 6.2 respectively.

6.4 Discussion and Conclusion

To tackle the task of graph clustering, we proposed convex programs 6.1 and 6.3 based on decomposing the adjacency matrix into low-rank and sparse components. (6.1) was already being used for low-rank and sparse decomposition task, however, we showed its viability for the specific problem of clustering and also developed tight conditions it fails. For sparse graphs, Improved Program 6.3 is shown to have a good performance comparable with existing literature. We believe, our technique can be extended to tightly analyze variants of this approach. These results can be extended in a few ways, for instance studying the case where the adjacency matrix is partially observed, or modifying the Programs 6.1 and 6.3 for clustering weighted graphs, where the adjacency with $\{0,1\}$-entries is replaced by a similarity matrix with real entries.

Chapter 7

Conclusions

There are several directions in which our results can be extended.

7.1 Generalized Lasso

In Chapter 3, we considered three formulation of the lasso problem. We give a full proof for the case of constrained lasso. For the ℓ_2-lasso, our analysis holds in the large SNR regime. For arbitrary SNR levels, the results in Section 3.10 show that $\frac{2\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{(\sqrt{m}-\sqrt{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))})^2}$ is an upper bound; however, in simulation $\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{m-\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}$ looks to be the correct upper bound. It would be interesting to prove that this is indeed the case.

We also state the conjecture on ℓ_2^2-lasso, which is the most popular variation. The proof of this conjecture would complete the missing piece in Chapter 3. Chapter 3 can be extended in several ways. An obvious direction is to investigate different loss functions, in particular, the problem,

$$\min_{\mathbf{x}'} \mathscr{L}(\mathbf{y}-\mathbf{A}\mathbf{x}')+\lambda f(\mathbf{x}').$$

The typical choices of $\mathscr{L}$ would be ℓ_1 and ℓ_∞ norms. ℓ_1 norm has resilience to sparse noise and ℓ_∞ is advantageous when the noise is known to be bounded (for instance $\pm\sigma$). Minimax formulation we used for ℓ_2-loss would work in these setups as well by making use of the corresponding dual norms. For instance, for the ℓ_1 loss, we have that

$$\min_{\mathbf{x}'} \|\mathbf{y}-\mathbf{A}\mathbf{x}'\|_1+\lambda f(\mathbf{x}') = \min_{\mathbf{x}'} \max_{\|\mathbf{v}\|_\infty\leq 1} \left\langle \mathbf{y}-\mathbf{A}\mathbf{x}', \mathbf{v}\right\rangle + \lambda f(\mathbf{x}').$$

Consequently, Gaussian Min-Max Theorem would be applicable to find bounds on the objective function.

Another direction is the exact characterization of the noise behavior in the arbitrary SNR regime. While high SNR error bound is a function of subdifferential, when the SNR is arbitrary we need more than first order statistics. For an arbitrary convex function $f(\cdot)$, it is not possible to give a simple formula as we do not know the curvature of the function. However, from Section 3.7, we have argued that the error bounds can be related to the solution of the key optimization (3.77). Hence, one possibility is to give the error in form of an outcome of the simpler optimization (3.77). The other option is the explicit calculation for well-known functions with a reasonable signal model. For instance, in [82], Donoho et al. attempts to analyze ℓ_1 recovery (lasso) where they assume $\mathbf{x}$ to be a sparse signal whose nonzero entries are i.i.d. Gaussian with a certain variance, which determines the SNR.

7.2 Universality of the Phase Transitions

There is a mature theory on phase transitions when the measurement ensemble is i.i.d. Gaussian. In particular, we have seen that Gaussian width and statistical dimension can fully capture the behavior of linear inverse problems. The usefulness of these quantities extends to the more challenging problems such as the demixing problem and noise analysis [101, 144, 170]. In Chapter 3, we have seen that the noisy problem (lasso formulation) requires the related quantity Gaussian distance. For linear inverse problems, these quantities are based only on the subdifferential $\partial f(\mathbf{x})$ (i.e., first order characteristics). In summary, with few parameters we are able to know what to expect from structured signal recovery problems.

It is widely accepted that the phase transitions exhibit universality [71]. In the simplest case, for noiseless ℓ_1 minimization, Donoho-Tanner bound characterizes the PT for a wide variety of ensembles including i.i.d. matrices. In Chapter 3, we have seen that numerical experiments indicate:

- Phase transitions are universal for different problems.
- Error (robustness) behaviors are universal.

Based on these, an obvious direction is to study the precise characteristics of i.i.d. subgaussian measurement ensembles. We expect Gaussian width and Gaussian distance to asymptotically (large dimensions) capture the behavior for these ensembles. In fact, there is already a significant amount of work dedicated toward this goal. In [129, 152], Mendelson and coauthors introduce comparison theorems for subgaussians. Using these techniques, Tropp shows that for a wide class of measurement ensembles, the required oversampling is proportional to the Gaussian width (up to a possibly large constant) [206]. Along the similar lines, in Chapter 4.1, we gave a short argument to obtain small constants for the Bernoulli ensemble. While these

results are powerful, they are far from capturing the true behavior: the asymptotic behavior is the *same* as Gaussians. Further study and generalization of message passing algorithms and their relation to BP is an alternative approach towards this goal. As of today, the universality phenomenon is proven for the case of sparse recovery thanks to the significant advances in AMP [12]. We believe proper applications of standard techniques such as the Lindeberg replacement might be promising directions towards this goal [55, 198]. Related directions include the investigation of universal behavior when we have nonlinear measurements. The notable examples are the phase retrieval and the one bit compressed sensing problems [35, 172].

7.3 Simultaneously structured signals

In Chapter 5, we have shown that for simultaneously structured signal (SSS), performance of convex optimization can be far from being optimal. This creates a challenge for several applications involving sparse and low-rank matrices and low-rank tensors. This bottleneck is also theoretically intriguing as for sparse and low-rank recovery, the performance of convex relaxation is on par with the performance of minimizing the (non-relaxed) cardinality and rank objectives. A natural direction is to overcome this limitation. This can be pursued in two ways.

- Find a better convex relaxation for SSS: Linear combination of the individual structure inducing norms is a tempting approach; however, it does not have to be the best convex relaxation. Can we construct a better norm which exploits the fact that the signal is simultaneously structured? If so, is this norm computationally efficient?
- We know that nonconvex approaches are information theoretically optimal; however, they have exponential complexity. Can we find an efficient algorithm tailored for SSS?

These problems are already getting attention. For instance, [132] proposes a new algorithm for the recovery of S&L matrices that can outperform the standard convex methods in certain settings (also see [158] and [180]).

In Chapter 5, we focused on the compressive sensing of simultaneously sparse and low-rank matrices, which shows up in quadratic CS and sparse phase retrieval. These signals also show up in important estimation problems such as *sparse principal component analysis* and *graph clustering* [64, 164, 236] (also see Chapter 6). It would be interesting to relate the computational and statistical challenges arising in these problems and have a unified theory that connects them.

7.4 Structured signal recovery beyond convexity

As we have argued throughout, for a structured signal estimation problem, when the problem is convex or it can be relaxed into a convex one, we have a good idea on how to approach it. For instance, in the case of linear inverse problems, subdifferential determines the characteristics of the problem. In various important problems, there is no nice convex formulation or the convex approaches are not computationally efficient. These include important machine learning problems such as the dictionary learning, nonnegative matrix factorization, and several formulations of the matrix completion [88, 125, 135, 141]. The formulations of these problems are often based on alternating minimization; however, we can additionally take advantage of the fact that the signal to be recovered is structured. As an example, consider the following formulation of the matrix completion problem. We wish to represent $\mathbf{X} \in \mathbb{R}^{n \times n}$ as $\mathbf{X} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are variables and r is an upper bound to $\text{rank}(\mathbf{X})$. Assume we have the measurements $\mathbf{y} = \mathscr{A}(\mathbf{X}) \in \mathbb{R}^m$. We wish to solve the problem

$$\min_{\mathbf{U},\mathbf{V} \in \mathbb{R}^{n \times r}} \|\mathbf{y} - \mathscr{A}(\mathbf{U}\mathbf{V}^T)\|_2.$$

The problem is not convex, however, it is convex in $\mathbf{U}$ for fixed $\mathbf{V}$ and vice versa. Hence, we can start from an initial point $\mathbf{U}_0, \mathbf{V}_0$ and alternate between $\mathbf{U}$ and $\mathbf{V}$ until convergence. This formulation has been studied extensively, and it has been recently shown that one can provably recover the low-rank $\mathbf{X}$ with a "good" initialization and for reasonable measurements $\mathscr{A}(\cdot)$ [123]. It would be interesting to have a general theory of such approaches when the objective signal is structured in some sense.

Bibliography

[1] Alekh Agarwal, Sahand Negahban, Martin J Wainwright, et al. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

[2] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[3] Nir Ailon, Yudong Chen, and Xu Huan. Breaking the small cluster barrier of graph clustering. *arXiv preprint arXiv:1302.4549*, 2013.

[4] DENNIS Amelunxen, MARTIN Lotz, MICHAEL B McCoy, and JOEL A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 2014.

[5] Brendan PW Ames. Robust convex relaxation for the planted clique and densest k-subgraph problems. *arXiv preprint arXiv:1305.4891*, 2013.

[6] Brendan PW Ames and Stephen A Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011.

[7] Brendan PW Ames and Stephen A Vavasis. Convex optimization for the planted k-disjoint-clique problem. *Mathematical Programming*, 143(1-2):299–337, 2014.

[8] ZD Bai and YQ Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The annals of Probability*, pages 1275–1294, 1993.

[9] Radu Balan, Bernhard G Bodmann, Peter G Casazza, and Dan Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4):488–501, 2009.

[10] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[11] Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 2008.

[12] Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *arXiv preprint arXiv:1207.7321*, 2012.

[13] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *Information Theory, IEEE Transactions on*, 57(2):764–785, 2011.

[14] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4):1997–2017, 2012.

[15] Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[16] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[17] Dimitri P Bertsekas, Angelia Nedić, and Asuman E Ozdaglar. *Convex analysis and optimization*. Athena Scientific Belmont, 2003.

[18] Badri Narayan Bhaskar and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 261–268. IEEE, 2011.

[19] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[20] Béla Bollobás and Alex D Scott. Max cut for random graphs with a planted partition. *Combinatorics, Probability and Computing*, 13(4-5):451–474, 2004.

[21] JM Borwein. A note on the existence of subgradients. *Mathematical Programming*, 24(1):225–228, 1982.

[22] Jonathan M Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*, volume 3. Springer, 2010.

[23] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[24] Florentina Bunea, Alexandre Tsybakov, Marten Wegkamp, et al. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[25] Florentina Bunea, Alexandre Tsybakov, Marten Wegkamp, et al. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[26] Jian-Feng Cai and Weiyu Xu. Guarantees of total variation minimization for signal recovery. *arXiv preprint arXiv:1301.6791*, 2013.

[27] T.T. Cai, G. Xu, and J. Zhang. New bounds for restricted isometry constants. *IEEE Tran. Info. Theory*, 56:4388–4394, 2010. Technical Report, Available Online.

[28] T.T. Cai, G. Xu, and J. Zhang. Shifting inequality and recovery of sparse signals. *IEEE Tran. Signal Processing*, 58:1300–1308, 2010. To appear in *IEEE Transactions on Signal Processing*.

[29] E.J. Candes, M.B. Wakin, and S. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.

[30] Emmanuel Candès and Benjamin Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

[31] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[32] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[33] Emmanuel J Candes and Mark A Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[34] Emmanuel J Candes, Yonina C Eldar, Deanna Needell, and Paige Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

[35] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[36] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[37] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[38] Emmanuel J Candes and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

[39] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[40] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[41] Emmanuel J Candes and Justin Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2):227–254, 2006.

[42] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[43] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[44] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[45] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[46] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

[47] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[48] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[49] Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

[50] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[51] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[52] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with qualitative information. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 524–533. IEEE, 2003.

[53] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24:1–14, 2008.

[54] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[55] Sourav Chatterjee et al. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6):2061–2076, 2006.

[56] Jie Chen and Xiaoming Huo. Theoretical results on sparse representations of multiple-measurement vectors. *Signal Processing, IEEE Transactions on*, 54(12):4634–4643, 2006.

[57] Scott Chen and David L Donoho. Examples of basis pursuit. In *SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*, pages 564–574. International Society for Optics and Photonics, 1995.

[58] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[59] Shaobing Chen and David Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

[60] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning*, pages 674–682, 2014.

[61] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *arXiv preprint arXiv:1104.4803*, 2011.

[62] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212, 2012.

[63] Anne Condon and Richard M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms*, 18(2):116–140, 2001.

[64] Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

[65] I. Daubechies, R. DeVore, M. Fornasier, and C.S. Gunturk. Iteratively re-weighted least squares minimization for sparse recovery. *Commun. Pure Appl. Math*, 63(1), 2010.

[66] Erik D Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2):172–187, 2006.

[67] Yash Deshpande and Andrea Montanari. Finding hidden cliques of size\ sqrt {N/e} in nearly linear time. *arXiv preprint arXiv:1304.7047*, 2013.

[68] Ronald A DeVore and Vladimir N Temlyakov. Some remarks on greedy algorithms. *Advances in computational Mathematics*, 5(1):173–187, 1996.

[69] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[70] David Donoho, Iain Johnstone, and Andrea Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *arXiv preprint arXiv:1111.1041*, 2011.

[71] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[72] David L Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

[73] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[74] David L Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete & Computational Geometry*, 35(4):617–652, 2006.

[75] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[76] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

[77] David L Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *arXiv preprint arXiv:1304.2085*, 2013.

[78] David L Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from gaussian measurements matches the minimax mse of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.

[79] David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2001.

[80] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[81] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[82] David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011.

[83] David L Donoho and Jared Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

[84] David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.

[85] David L Donoho and Jared Tanner. Thresholds for the recovery of sparse solutions via l1 minimization. In *Information Sciences and Systems, 2006 40th Annual Conference on*, pages 202–206. IEEE, 2006.

[86] Sever Silvestru Dragomir, Ravi P Agarwal, and Neil S Barnett. Inequalities for beta and gamma functions via some classical and new integral inequalities. *Journal of Inequalities and Applications*, 5(2):103–165, 1900.

[87] K. Dvijotham and M. Fazel. A nullspace analysis of the nuclear norm heuristic for rank minimization. In *Proc. of ICASSP*, 2010.

[88] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[89] Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007.

[90] Michael Elad, J-L Starck, Philippe Querre, and David L Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis*, 19(3):340–358, 2005.

[91] Yonina C Eldar, Patrick Kuppinger, and Helmut Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *Signal Processing, IEEE Transactions on*, 58(6):3042–3054, 2010.

[92] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[93] Paul Erdös and A Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[94] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

[95] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

[96] A. Feuer and A. Nemirovski. On sparse sepresentation in pairs of bases. *IEEE Tran. Info. Theory*, 49(6):1579 – 1581, 2003.

[97] James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[98] Gary William Flake, Robert E Tarjan, and Kostas Tsioutsiouliklis. Graph clustering and minimum cut trees. *Internet Mathematics*, 1(4):385–408, 2004.

[99] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[100] Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

[101] Rina Foygel and Lester Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *Information Theory, IEEE Transactions on*, 60(2):1223–1247, 2014.

[102] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[103] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[104] R. Garg and R.Khandekar. Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *Proc. of 26th Annual ICML*, 2009.

[105] Ralph W Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237, 1972.

[106] Joachim Giesen and Dieter Mitsche. Reconstructing many partitions using spectral techniques. In *Fundamentals of Computation Theory*, pages 433–444. Springer, 2005.

[107] Anna C Gilbert, Sudipto Guha, Piotr Indyk, S Muthukrishnan, and Martin Strauss. Near-optimal sparse fourier representations via sampling. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 152–161. ACM, 2002.

[108] Ioannis Giotis and Venkatesan Guruswami. Correlation clustering with a fixed number of clusters. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1167–1176. ACM, 2006.

[109] Mohammad Golbabaee and Pierre Vandergheynst. Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2741–2744. Ieee, 2012.

[110] Mohammad Golbabaee and Pierre Vandergheynst. Joint trace/tv norm minimization: A new efficient approach for spectral compressive imaging. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 933–936. IEEE, 2012.

[111] Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[112] Yehoram Gordon. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

[113] Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[114] Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013.

[115] Robert W Harrison. Phase problem in crystallography. *JOSA A*, 10(5):1046–1055, 1993.

[116] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I: Part 1: Fundamentals*, volume 305. Springer, 1996.

[117] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[118] RA Horn and CR Johnson. Topics in matrix analysis. *Cambridge University Presss, Cambridge*, 1991.

[119] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[120] Norman E Hurt. *Phase Retrieval and Zero Crossings: Mathematical Methods in Image Reconstruction*, volume 52. Springer, 2001.

[121] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium On*, pages 1473–1477. IEEE, 2012.

[122] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1022–1026. IEEE, 2013.

[123] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[124] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010.

[125] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, pages 952–960, 2009.

[126] M Amin Khajehnejad, Alexandros G Dimakis, Weiyu Xu, and Babak Hassibi. Sparse recovery of nonnegative signals with minimal expansion. *Signal Processing, IEEE Transactions on*, 59(1):196–208, 2011.

[127] M Amin Khajehnejad, Weiyu Xu, Amir Salman Avestimehr, and Babak Hassibi. Weighted ℓ_1 minimization for sparse recovery with prior information. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 483–487. IEEE, 2009.

[128] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[129] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *arXiv preprint arXiv:1312.3580*, 2013.

[130] M.J. Lai. An improved estimate for restricted isometry constant for the ℓ_1 minimization. *IEEE Trans Info. Theory*, 2010. Available at http://www.math.uga.edu/~mjlai/papers/Lai10CSfinal.pdf.

[131] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

[132] Kiryung Lee, Yihong Wu, and Yoram Bresler. Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. *arXiv preprint arXiv:1312.0525*, 2013.

[133] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013.

[134] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[135] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[136] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[137] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013.

[138] Miguel Sousa Lobo, Lieven Vandenberghe, Stephen Boyd, and Hervé Lebret. Applications of second-order cone programming. *Linear algebra and its applications*, 284(1):193–228, 1998.

[139] Yue M Lu and Martin Vetterli. Sparse spectral factorization: Unicity and reconstruction algorithms. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5976–5979. IEEE, 2011.

[140] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine*, 58(6):1182–1195, 2007.

[141] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009.

[142] Arian Maleki, Laura Anitori, Zai Yang, and Richard G Baraniuk. Asymptotic analysis of complex lasso via complex approximate message passing (camp). *Information Theory, IEEE Transactions on*, 59(7):4290–4308, 2013.

[143] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[144] Michael B McCoy and Joel A Tropp. Sharp recovery bounds for convex deconvolution, with applications. *arXiv preprint arXiv:1205.1580*, 2012.

[145] Michael B McCoy and Joel A Tropp. The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478*, 2013.

[146] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[147] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[148] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

[149] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

[150] R. Meka, P. Jain, and I.S. Dhillon. Guaranteed rank minimization via singular value projection. In *Proc. of NIPS*, 2010.

[151] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

[152] Shahar Mendelson. Learning without concentration. *arXiv preprint arXiv:1401.0304*, 2014.

[153] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[154] RP Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[155] K. Mohan and M. Fazel. Iterative reweighted least squares for matrix rank minimization. In *Proc. Allerton conference on Controls and communications*, 2010.

[156] K. Mohan and M. Fazel. New restricted isometry results for noisy recovery. In *Proc. International Symposium Information Theory*, 2010.

[157] Jean-Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

[158] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013.

[159] D. Needell. Noisy signal recovery via iterative reweighted l1-minimization. In *Proc. Asilomar conference on Signals, Systems and Computers*, 2009.

[160] Deanna Needell and Rachel Ward. Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058, 2013.

[161] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[162] Henrik Ohlsson, Allen Y Yang, Roy Dong, and Shankar S Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv preprint arXiv*, 1111, 2011.

[163] Samet Oymak and Babak Hassibi. New null space results and recovery thresholds for matrix rank minimization. *arXiv preprint arXiv:1011.6326*, 2010.

[164] Samet Oymak and Babak Hassibi. Finding dense clusters via "low rank + sparse" decomposition. *arXiv preprint arXiv:1104.5186*, 2011.

[165] Samet Oymak and Babak Hassibi. Tight recovery thresholds and robustness analysis for nuclear norm minimization. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2323–2327. IEEE, 2011.

[166] Samet Oymak and Babak Hassibi. On a relation between the minimax risk and the phase transitions of compressed recovery. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1018–1025. IEEE, 2012.

[167] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *arXiv preprint arXiv:1305.2714*, 2013.

[168] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

[169] Samet Oymak, Karthik Mohan, Maryam Fazel, and Babak Hassibi. A simplified approach to recovery conditions for low rank matrices. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2318–2322. IEEE, 2011.

[170] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*, 2013.

[171] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2233–2246, 2012.

[172] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

[173] S James Press. *Applied multivariate analysis: using Bayesian and frequentist methods of inference*. Courier Dover Publications, 2012.

[174] Bhaskar D Rao and Kenneth Kreutz-Delgado. An affine scaling methodology for best basis selection. *Signal Processing, IEEE Transactions on*, 47(1):187–200, 1999.

[175] Nikhil Rao, Benjamin Recht, and Robert Nowak. Tight measurement bounds for exact recovery of structured sparse signals. *arXiv preprint arXiv:1106.4355*, 2011.

[176] Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[177] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[178] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[179] Benjamin Recht, Weiyu Xu, and Babak Hassibi. Null space conditions and thresholds for rank minimization. *Mathematical programming*, 127(1):175–202, 2011.

[180] Emile Richard, Guillaume Obozinski, and Jean-Philippe Vert. Tight convex relaxations for sparse matrix factorization. *arXiv preprint arXiv:1407.5158*, 2014.

[181] Emile Richard, Pierre-André Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. *arXiv preprint arXiv:1206.6474*, 2012.

[182] R Tyrrell Rockafellar. *Convex analysis*. Princeton university press, 1997.

[183] R Tyrrell Rockafellar, Roger J-B Wets, and Maria Wets. *Variational analysis*, volume 317. Springer, 1998.

[184] Mark Rudelson and Roman Vershynin. Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In *Information Sciences and Systems, 2006 40th Annual Conference on*, pages 207–212. IEEE, 2006.

[185] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[186] LR Scott. *Numerical Analysis*. Princeton University Press, 2011.

[187] Oguz Semerci, Ning Hao, Misha E Kilmer, and Eric L Miller. Tensor-based formulation and nuclear norm regularization for multi-energy computed tomography. *arXiv preprint arXiv:1307.5348*, 2013.

[188] Yoav Shechtman, Amir Beck, and Yonina C Eldar. Efficient phase retrieval of sparse signals. In *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pages 1–5. IEEE, 2012.

[189] Yoav Shechtman, Yonina C Eldar, Alexander Szameit, and Mordechai Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics express*, 19(16):14807–14822, 2011.

[190] Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *Signal Processing, IEEE Transactions on*, 59(9):4183–4198, 2011.

[191] Mihailo Stojnic. Block-length dependent thresholds in block-sparse compressed sensing. *arXiv preprint arXiv:0907.3679*, 2009.

[192] Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[193] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[194] Mihailo Stojnic. Upper-bounding ℓ_1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013.

[195] Mihailo Stojnic, Farzad Parvaresh, and Babak Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *Signal Processing, IEEE Transactions on*, 57(8):3075–3085, 2009.

[196] Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

[197] A Szameit, Yoav Shechtman, E Osherovich, E Bullkich, P Sidorenko, H Dana, S Steiner, Ernst B Kley, S Gazit, T Cohen-Hyams, et al. Sparsity-based single-shot subwavelength coherent diffractive imaging. *Nature materials*, 11(5):455–459, 2012.

[198] Terence Tao and Van Vu. Random matrices: universality of local eigenvalue statistics. *Acta mathematica*, 206(1):127–204, 2011.

[199] Chris Thrampoulidis, Samet Oymak, and Babak Hassibi. A tight version of the gaussian min-max theorem in the presence of convexity. *in preparation*, 2014.

[200] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Simple error bounds for regularized noisy linear inverse problems. *arXiv preprint arXiv:1401.6578*, 2014.

[201] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[202] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[203] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.

[204] Joel A Tropp. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing*, 86(3):589–602, 2006.

[205] Joel A Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

[206] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. *arXiv preprint arXiv:1405.1102*, 2014.

[207] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[208] Joel A Tropp, Anna C Gilbert, and Martin J Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

[209] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[210] M. Uchiyama. Subadditivity of eigenvalue sums. *Proc. American Mathematical Society*, 134(5):1405–1412, 2005.

[211] Samuel Vaiter, Gabriel Peyré, Charles Dossal, and Jalal Fadili. Robust sparse analysis regularization. *Information Theory, IEEE Transactions on*, 59(4):2001–2016, 2013.

[212] Namrata Vaswani and Wei Lu. Modified-cs: Modifying compressive sensing for problems with partially known support. *Signal Processing, IEEE Transactions on*, 58(9):4595–4607, 2010.

[213] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[214] Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi. Sharp performance bounds for graph clustering via convex optimization, 2013.

[215] Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi. Graph clustering with missing data : Convex algorithms and analysis, 2014.

[216] Van H Vu. Spectral norm of random matrices. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 423–430. ACM, 2005.

[217] MJ Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[218] Adriaan Walther. The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49, 1963.

[219] Meng Wang, Weiyu Xu, and Ao Tang. On the performance of sparse recovery via-minimization. *Information Theory, IEEE Transactions on*, 57(11):7255–7278, 2011.

[220] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

[221] John Wright, Arvind Ganesh, Kerui Min, and Yi Ma. Compressive principal component pursuit. *Information and Inference*, 2(1):32–68, 2013.

[222] John Wright, Arvind Ganesh, Kerui Min, and Yi Ma. Compressive principal component pursuit. *Information and Inference*, 2(1):32–68, 2013.

[223] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

[224] Weiyu Xu and Babak Hassibi. Compressed sensing over the grassmann manifold: A unified analytical framework. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 562–567. IEEE, 2008.

[225] Weiyu Xu and Babak Hassibi. On sharp performance bounds for robust sparse signal recoveries. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 493–497. IEEE, 2009.

[226] Weiyu Xu and Babak Hassibi. Precise stability phase transitions for minimization: A unified geometric framework. *Information Theory, IEEE Transactions on*, 57(10):6894–6919, 2011.

[227] Weiyu Xu, M Amin Khajehnejad, Amir Salman Avestimehr, and Babak Hassibi. Breaking through the thresholds: an analysis for iterative reweighted ℓ_1 minimization via the grassmann angle framework. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5498–5501. IEEE, 2010.

[228] Xiaowei Xu, Jochen Jäger, and Hans-Peter Kriegel. A fast parallel clustering algorithm for large spatial databases. In *High Performance Data Mining*, pages 263–290. Springer, 2002.

[229] Ying Xu, Victor Olman, and Dong Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, 2002.

[230] Qiaofeng Yang and Stefano Lonardi. A parallel algorithm for clustering protein-protein interaction networks. In *Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE*, pages 174–177. IEEE, 2005.

[231] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[232] Gesen Zhang, Shuhong Jiao, Xiaoli Xu, and Lan Wang. Compressed sensing and reconstruction with bernoulli matrices. In *Information and Automation (ICIA), 2010 IEEE International Conference on*, pages 455–460. IEEE, 2010.

[233] Y. Zhang. Theory of compressive sensing via l_1 minimization: A Non-RIP analysis and extensions. *IEEE Trans Info. Theory*, 2008. Technical Report TR08-11 revised. Available at http://www.caam.rice.edu/~zhang/reports/tr0811_revised.pdf.

[234] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[235] Zihan Zhou, Xiaodong Li, John Wright, Emmanuel Candes, and Yi Ma. Stable principal component pursuit. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1518–1522. IEEE, 2010.

[236] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

List of Figures

List of Tables

Appendix A

Further Proofs for Chapter 3

A.1 Auxiliary Results

Lemma A.1 *Let $f(\cdot):\mathbb{R}^p\to\mathbb{R}$ be an L-Lipschitz function and $\mathbf{g}\sim\mathcal{N}(0,I_p)$. Then, with probability $1-2\exp(-\frac{t^2}{2L^2})$,*

$$\sqrt{\mathbb{E}\left[(f(\mathbf{g}))^2\right]-L^2}-t\leq f(\mathbf{g})\leq\sqrt{\mathbb{E}\left[(f(\mathbf{g}))^2\right]}+t.$$

Proof: From Fact 2.4, $|f(\mathbf{g})-\mathbb{E}[f(\mathbf{g})]|\leq t$ holds with probability $1-2\exp(-\frac{t^2}{2L^2})$. Furthermore,

$$\mathbb{E}[(f(\mathbf{g}))^2]-L^2\leq(\mathbb{E}[f(\mathbf{g})])^2\leq\mathbb{E}[(f(\mathbf{g}))^2]. \tag{A.1}$$

The left hand side inequality in (A.1) follows from an application of Fact 2.3 and the right hand side follows from Jensen's Inequality. Combining $|f(\mathbf{g})-\mathbb{E}[f(\mathbf{g})]|\leq t$ and (A.1) completes the proof. ■

For the statements of the lemmas below, recall the definitions of $\mathbf{D}(\mathscr{C})$,$\mathbf{P}(\mathscr{C})$ and $\mathbf{C}(\mathscr{C})$ in Section 3.5.2.

Lemma A.2 *Let $\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_m),\mathbf{h}\sim\mathcal{N}(0,\mathbf{I}_n)$ and let $\mathscr{C}\in\mathbb{R}^n$ be a closed and convex set. Given $t>0$, each of the followings hold with probability $1-2\exp\left(\frac{-t^2}{2}\right)$.*

- $\sqrt{m-1}-t\leq\|\mathbf{g}\|_2\leq\sqrt{m}+t$
- $\sqrt{\mathbf{D}(\mathscr{C})-1}-t\leq dist(\mathbf{h},\mathscr{C})\leq\sqrt{\mathbf{D}(\mathscr{C})}+t$
- $\sqrt{\mathbf{P}(\mathscr{C})-1}-t\leq\|Proj(\mathbf{h},\mathscr{C})\|_2\leq\sqrt{\mathbf{P}(\mathscr{C})}+t$

Proof: The result is an immediate application of Lemma A.1. The functions $\|\cdot\|_2$, $\|\text{Proj}(\cdot,\mathscr{C})\|_2$ and $\text{dist}(\cdot,\mathscr{C})$ are all 1-Lipschitz. Furthermore, $\mathbb{E}[\|\mathbf{g}\|_2^2]=m$ and $\mathbb{E}[\text{Proj}(\mathbf{h},\mathscr{C})^2]=\mathbf{P}(\mathscr{C})$, $\mathbb{E}[\text{dist}(\mathbf{h},\mathscr{C})^2]=\mathbf{D}(\mathscr{C})$ by definition. ■

Lemma A.3 *Let* $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *and let* $\mathscr{C} \in \mathbb{R}^n$ *be a convex and closed set. Then, given* $t > 0$*,*

- $|dist(\mathbf{h}, \mathscr{C})^2 - \mathbf{D}(\mathscr{C})| \leq 2t\sqrt{\mathbf{D}(\mathscr{C})} + t^2 + 1.$
- $|\|Proj(\mathbf{h}, \mathscr{C})\|_2^2 - \mathbf{P}(\mathscr{C})| \leq 3t\sqrt{n + \mathbf{D}(\mathscr{C})} + t^2 + 1.$
- $|corr(\mathbf{h}, \mathscr{C}) - \mathbf{C}(\mathscr{C})| \leq 3t\sqrt{n + \mathbf{D}(\mathscr{C})} + t^2 + 1.$

with probability $1 - 4\exp(-\frac{t^2}{2})$.

Proof: The first two statements follow trivially from Lemma A.2. For the second statement, use again Lemma A.2 and also upper bound $\mathbf{P}(\mathscr{C})$ by $2(n + \mathbf{D}(\mathscr{C}))$ via Lemma A.4. To obtain the third statement, we write,

$$\text{corr}(\mathbf{h}, \mathscr{C}) = \frac{n - (\|\text{Proj}(\mathbf{h}, \mathscr{C})\|_2^2 + \text{dist}(\mathbf{h}, \mathscr{C})^2)}{2}$$

and use the fact that first two statements hold with probability $1 - 4\exp(-\frac{t^2}{2})$. This will give,

$$|\text{corr}(\mathbf{h}, \mathscr{C}) - \mathbf{C}(\mathscr{C})| \leq t(\sqrt{\mathbf{D}(\mathscr{C})} + \sqrt{\mathbf{P}(\mathscr{C})}) + t^2 + 1,$$

which when combined with Lemma A.4 concludes the proof. ■

Lemma A.4 *Let* $\mathscr{C} \in \mathbb{R}^n$ *be a convex and closed set. Then, the following holds,*

$$\max\{\mathbf{C}(\mathscr{C}), \mathbf{P}(\mathscr{C})\} \leq 2(n + \mathbf{D}(\mathscr{C})).$$

Proof: From triangle inequality, for any $\mathbf{h} \in \mathbb{R}^n$,

$$\|\text{Proj}(\mathbf{h}, \mathscr{C})\|_2 \leq \|\mathbf{h}\|_2 + \text{dist}(\mathbf{h}, \mathscr{C}).$$

We also have,

$$\mathbb{E}[\|\mathbf{h}\|_2 \cdot \text{dist}(\mathbf{h}, \mathscr{C})] \leq \frac{1}{2}(\mathbb{E}[\|\mathbf{h}\|_2^2] + \mathbb{E}[\text{dist}(\mathbf{h}, \mathscr{C})^2]) = \frac{n + \mathbf{D}(\mathscr{C})}{2}.$$

From these, we may write,

$$\begin{aligned}
\mathbf{C}(\mathscr{C}) &= \mathbb{E}[\langle \Pi(\mathbf{h}, \mathscr{C}), \text{Proj}(\mathbf{h}, \mathscr{C}) \rangle] \\
&\leq \mathbb{E}[\text{dist}(\mathbf{h}, \mathscr{C}) \|\text{Proj}(\mathbf{h}, \mathscr{C})\|_2] \\
&\leq \frac{n + 3\mathbf{D}(\mathscr{C})}{2}.
\end{aligned}$$

Similarly, we have,

$$\mathbf{P}(\mathscr{C}) = \mathbb{E}[\|\text{Proj}(\mathbf{h},\mathscr{C})\|_2^2] \leq \mathbb{E}[\|\mathbf{h}\|_2 + \text{dist}(\mathbf{h},\mathscr{C})^2] \leq 2(n+\mathbf{D}(\mathscr{C})).$$

■

Lemma A.5 *Let $\mathbf{g} \sim \mathcal{N}(0,\mathbf{I}_m)$ and $\mathbf{h} \sim \mathcal{N}(0,\mathbf{I}_n)$. Let $\mathscr{C}$ be a closed and convex set in $\mathbb{R}^n$. Assume $m(1-\varepsilon_L) > \mathbf{D}(\mathscr{C}) > \varepsilon_L m$ for some constant $\varepsilon_L > 0$ and m is sufficiently large. Then, for any constant $\varepsilon > 0$, each of the following holds with probability $1-\exp(-\mathscr{O}(m))$,*

- $\|\mathbf{g}\|_2 > dist(\mathbf{h},\mathscr{C})$.
- $\left|\frac{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\mathscr{C})}{m-\mathbf{D}(\mathscr{C})} - 1\right| < \varepsilon$.
- $\left|\frac{dist^2(\mathbf{h},\mathscr{C})}{\|\mathbf{g}\|_2^2 - dist^2(\mathbf{h},\mathscr{C})} \times \frac{m-\mathbf{D}(\mathscr{C})}{\mathbf{D}(\mathscr{C})} - 1\right| < \varepsilon$.

Proof: Let δ be a constant to be determined. For sufficiently large m, using Lemma A.2, with probability $1-\exp(-\mathscr{O}(m))$, we have,

$$|\|\mathbf{g}\|_2^2 - m| < \delta m, \quad |\text{dist}(\mathbf{h},\mathscr{C})^2 - \mathbf{D}(\mathscr{C})| < \delta m$$

Now, choose $\delta < \frac{\varepsilon_L}{2}$, which gives,

$$\|\mathbf{g}\|_2 \geq \sqrt{m(1-\delta)} > \sqrt{\mathbf{D}(\mathscr{C}) + \varepsilon_L m - \delta m} > \sqrt{\mathbf{D}(\mathscr{C}) + \delta m} \geq \text{dist}(\mathbf{h},\mathscr{C})$$

This gives the first statement. For the second statement, observe that,

$$1 + \frac{2\delta}{\varepsilon_L} \geq \frac{m - \mathbf{D}(\mathscr{C}) + 2\delta}{m - \mathbf{D}(\mathscr{C})} \geq \frac{\|\mathbf{g}\|_2^2 - \text{dist}^2(\mathbf{h},\mathscr{C})}{m - \mathbf{D}(\mathscr{C})} \geq \frac{m - \mathbf{D}(\mathscr{C}) - 2\delta}{m - \mathbf{D}(\mathscr{C})} \geq 1 - \frac{2\delta}{\varepsilon_L}.$$

Choose $\frac{\delta}{\varepsilon_L} < \frac{\varepsilon}{2}$ to ensure the desired result. For the last statement, we similarly have,

$$\frac{1+\frac{\delta}{\varepsilon_L}}{1-\frac{2\delta}{\varepsilon_L}} \geq \frac{\text{dist}^2(\mathbf{h},\mathscr{C})}{\|\mathbf{g}\|_2^2 - \text{dist}^2(\mathbf{h},\mathscr{C})} \times \frac{m-\mathbf{D}(\mathscr{C})}{\mathbf{D}(\mathscr{C})} \geq \frac{1-\frac{\delta}{\varepsilon_L}}{1+\frac{2\delta}{\varepsilon_L}} \tag{A.2}$$

To conclude, notice that we can choose $\frac{\delta}{\varepsilon_L}$ sufficiently small (constant) to ensure that the left and right bounds in (A.2) above are between $1 \pm \varepsilon$.

■

Proof: [**Proof of Lemma 3.30**] We will show the results for $L_{\mathbf{P}}(\lambda)$ and $L_{\mathbf{D}}(\lambda)$. $L_{\mathbf{C}}(\lambda)$ follows from the fact that $\mathbf{P}(\lambda\mathscr{C})+\mathbf{D}(\lambda\mathscr{C})+2\mathbf{C}(\lambda\mathscr{C})=n$. Let $\mathbf{h}\in\mathbb{R}^n$. Then, for $\lambda+\varepsilon,\lambda>0$,

$$\|\mathrm{Proj}(\mathbf{h},(\lambda+\varepsilon)\mathscr{C})\|_2=\frac{\lambda+\varepsilon}{\lambda}\|\mathrm{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\varepsilon},\lambda\mathscr{C})\|_2=\|\mathrm{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\varepsilon},\lambda\mathscr{C})\|_2+\frac{\varepsilon}{\lambda}\|\mathrm{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\varepsilon},\lambda\mathscr{C})\|_2$$

This gives,

$$\left|\|\mathrm{Proj}(\mathbf{h},(\lambda+\varepsilon)\mathscr{C})\|_2-\|\mathrm{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\varepsilon},\lambda\mathscr{C})\|_2\right|\leq|\varepsilon|R$$

Next, observe that,

$$\left|\|\mathrm{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\varepsilon},\lambda\mathscr{C})\|_2-\|\mathrm{Proj}(\mathbf{h},\lambda\mathscr{C})\|_2\right|\leq\frac{|\varepsilon|\|\mathbf{h}\|_2}{\lambda+\varepsilon}$$

Combining, letting $\mathbf{h}\sim\mathscr{N}(0,\mathbf{I}_n)$ and using $\|\mathrm{Proj}(\mathbf{h},\lambda\mathscr{C})\|_2\leq\lambda R$, we find,

$$\begin{aligned}\mathbf{P}((\lambda+\varepsilon)\mathscr{C})&\leq\mathbb{E}[(\|\mathrm{Proj}(\mathbf{h},\lambda\mathscr{C})\|_2+\frac{|\varepsilon|\|\mathbf{h}\|_2}{\lambda+\varepsilon}+|\varepsilon|R)^2]\\&\leq\mathbf{P}(\lambda\mathscr{C})+2\lambda R|\varepsilon|(\frac{\mathbb{E}[\|\mathbf{h}\|_2]}{\lambda+\varepsilon}+R)+|\varepsilon|^2\mathbb{E}[(\frac{\|\mathbf{h}\|_2}{\lambda+\varepsilon}+R)^2]\end{aligned}$$

Obtaining the similar lower bound on $\mathbf{P}((\lambda+\varepsilon)\mathscr{C})$ and letting $\varepsilon\to0$,

$$L_{\mathbf{P}}(\lambda)=\limsup_{\varepsilon\to0}\left|\frac{\mathbf{P}((\lambda+\varepsilon)\mathscr{C})-\mathbf{P}(\lambda\mathscr{C})}{\varepsilon}\right|\leq\lim_{\varepsilon\to0}2\lambda R(\frac{\mathbb{E}[\|\mathbf{h}\|_2]}{\lambda+\varepsilon}+R+\mathscr{O}(|\varepsilon|))\leq2R(\sqrt{n}+\lambda R)$$

For $\lambda=0$, observe that for any $\varepsilon>0,\mathbf{h}\in\mathbb{R}^n$, $\|\mathrm{Proj}(\mathbf{h},\varepsilon\mathscr{C})\|_2\leq\varepsilon R$ which implies $\mathbf{P}(\varepsilon\mathscr{C})\leq\varepsilon^2R^2$. Hence,

$$L_{\mathbf{P}}(0)=\lim_{\varepsilon\to0^+}\varepsilon^{-1}(\mathbf{P}(\varepsilon\mathscr{C})-\mathbf{P}(0))=0\tag{A.3}$$

Next, consider $\mathbf{D}(\lambda\mathscr{C})$. Using differentiability of $\mathbf{D}(\lambda\mathscr{C})$, for $\lambda>0$,

$$L_{\mathbf{D}}(\lambda)=|\mathbf{D}(\lambda\mathscr{C})'|=\frac{2}{\lambda}|\mathbf{C}(\lambda\mathscr{C})|\leq\frac{2\cdot\mathbb{E}[\|\mathrm{Proj}(\mathbf{h},\lambda\mathscr{C})\|_2\cdot\mathrm{dist}(\mathbf{h},\lambda\mathscr{C})]}{\lambda}\leq2R\cdot\mathbb{E}[\mathrm{dist}(\mathbf{h},\lambda\mathscr{C})]\leq2R(\sqrt{n}+\lambda R)$$

For $\lambda=0$, see the "Continuity at zero" part of the proof of Lemma B.2 in [4], which gives the upper bound $2R\sqrt{n}$ on $L_{\mathbf{D}}(0)$. ■

A.2 Proof of Proposition 2.7

In this section we prove Proposition 2.7. The proposition is a consequence of Theorem 2.2. We repeat the statement of the proposition for ease of reference.

Proposition A.1 (Modified Gordon's Lemma) *Let* $\mathbf{G} \in \mathbb{R}^{m\times n}$, $\mathbf{g} \in \mathbb{R}^m$, $\mathbf{h} \in \mathbb{R}^n$ *be independent with i.i.d* $\mathcal{N}(0,1)$ *entries and let* $\Phi_1 \subset \mathbb{R}^n$ *be arbitrary and* $\Phi_2 \subset \mathbb{R}^m$ *be a compact set. Also, assume* $\psi(\cdot,\cdot) : \Phi_1 \times \Phi_2 \to \mathbb{R}$ *is a continuous function. Then, for any* $c \in \mathbb{R}$*:*

$$\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\mathbf{a}^T\mathbf{G}\mathbf{x}-\psi(\mathbf{x},\mathbf{a})\right\}\geq c\right)\geq 2\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a}-\|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x}-\psi(\mathbf{x},\mathbf{a})\right\}\geq c\right)-1.$$

Our proof will closely parallel the proof of Gordon's original variation Lemma 3.1 of [112].

Proof: Let S_1, S_2 be finite subsets of Φ_1, Φ_2. For $\mathbf{x} \in S_1$ and $\mathbf{a} \in S_2$ define the two processes,

$$Y_{\mathbf{x},\mathbf{a}} = \mathbf{x}^T\mathbf{G}\mathbf{a} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g \quad \text{and} \quad X_{\mathbf{x},\mathbf{a}} = \|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x}$$

where $\mathbf{G}, \mathbf{g}, \mathbf{h}$ are as defined in the statement of the proposition and $g \sim \mathcal{N}(0,1)$ and independent of the other. We show that the processes defined satisfy the conditions of Gordon's Theorem 2.2:

$$\mathbb{E}[X_{\mathbf{x},\mathbf{a}}^2] = \|\mathbf{x}\|_2^2\|\mathbf{a}\|_2^2 + \|\mathbf{a}\|_2^2\|\mathbf{x}\|_2^2 = \mathbb{E}[Y_{\mathbf{x},\mathbf{a}}^2],$$

and

$$\begin{aligned}\mathbb{E}[X_{\mathbf{x},\mathbf{a}}X_{\mathbf{x}',\mathbf{a}'}] - \mathbb{E}[Y_{\mathbf{x},\mathbf{a}}Y_{\mathbf{x}',\mathbf{a}'}] &= \|\mathbf{x}\|_2\|\mathbf{x}'\|_2(\mathbf{a}^T\mathbf{a}') + \|\mathbf{a}\|_2^2(\mathbf{x}^T\mathbf{x}') - (\mathbf{x}^T\mathbf{x}')(\mathbf{a}^T\mathbf{a}') - \|\mathbf{a}\|_2\|\mathbf{a}'\|_2\|\mathbf{x}\|_2\|\mathbf{x}'\|_2 \\ &= \left(\underbrace{\|\mathbf{x}\|_2\|\mathbf{x}'\|_2 - (\mathbf{x}^T\mathbf{x}')}_{\geq 0}\right)\left(\underbrace{(\mathbf{a}^T\mathbf{a}') - \|\mathbf{a}\|_2\|\mathbf{a}'\|_2}_{\leq 0}\right),\end{aligned}$$

which is non positive and equal to zero when $x = x'$. Also, on the way of applying Theorem 2.2 for the two processes defined above, let

$$\lambda_{\mathbf{x},\mathbf{a}} = \psi(\mathbf{x},\mathbf{a}) + c.$$

Now, for finite Φ_1 and Φ_2, we can apply Theorem 2.2 and then use Section A.2.2 to conclude. Hence, we are interested in moving from finite sets to the case where Φ_1 is arbitrary and Φ_2 is compact. This technicality is addressed by Gordon in [112] (see Lemma 3.1 therein), for the case where Φ_1 is arbitrary, Φ_2 is the unit sphere and $\psi(\cdot)$ is only a function of $\mathbf{x}$.

We will give two related proofs for this, one of which follows directly that of Gordon's argument and the other one is based on an elementary covering technique but requires Φ_1 to be compact.

A.2.1 From discrete to continuous sets

• Moving from a finite set to a compact set can be done via Lemma A.6. Hence, we may replace S_2 with Φ_2. To move from S_1 to Φ_1, repeating Gordon's argument, we will first show that, for fixed $\mathbf{x}$, the set,

$$\{[\mathbf{G},g]\in\mathbb{R}^{mn+1}\big|\max_{\mathbf{a}\in\Phi_2}\{Y_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0\}=\{\max_{\mathbf{a}\in\Phi_2}\{\mathbf{x}^T\mathbf{G}\mathbf{a}+\|\mathbf{a}\|_2\|\mathbf{x}\|_2 g-\psi(\mathbf{x},\mathbf{a})-c\}\geq 0\}$$

is closed in the probability space $\{\mathbb{R}^{mn+1},P\}$ where P is the standard Gaussian measure of $\mathbb{R}^{mn+1}$. To see this, observe that, for fixed $\mathbf{x}$ using the boundedness of Φ_2, $\gamma_{\mathbf{x},\mathbf{a}}(\mathbf{G},g)=\max_{\mathbf{a}\in\Phi_2}\{\mathbf{x}^T\mathbf{G}\mathbf{a}+\|\mathbf{a}\|_2\|\mathbf{x}\|_2 g-\psi(\mathbf{x},\mathbf{a})-c\}$ is a continuous function of $[\mathbf{G},g]$. Hence, if $\gamma_{\mathbf{x},\mathbf{a}}(\mathbf{G},g)<0$, a sufficiently small neighborhood of $[\mathbf{G},g]$ will be strictly negative as well. Hence, the set is indeed closed. The same argument applies to $X_{\mathbf{x},\mathbf{a}}$ (i.e. $[\mathbf{g},\mathbf{h}]$). Let F be the collection of finite subsets S_1 of Φ_1 ordered by inclusion. From Theorem 2.2, we know that,

$$\lim_F \mathbb{P}(\min_{\mathbf{x}\in S_1}\max_{\mathbf{a}\in\Phi_2}\{Y_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0)\geq\lim_F\mathbb{P}(\min_{\mathbf{x}\in S_1}\max_{\mathbf{a}\in\Phi_2}\{X_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0). \tag{A.4}$$

Here, by inclusion, the (left and right) sequences are decreasing hence the limits exist. Now, the sets $\{\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\{X_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0\}$ and $\{\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\{Y_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0\}$ are intersection of closed sets, and hence are closed themselves (in $\{\mathbb{R}^{m+n},P\}$, $\{\mathbb{R}^{mn+1},P\}$). Additionally P is a regular measure, therefore, it follows that, the limits over F will yield the intersections over Φ_1, i.e. (A.4) is identical to,

$$\mathbb{P}(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\{Y_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0)\geq\mathbb{P}(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\{X_{\mathbf{x},\mathbf{a}}-\lambda_{\mathbf{x},\mathbf{a}}\}\geq 0).$$

• Argument for compact sets: When Φ_1 is also compact, we can state the following lemma.

Lemma A.6 *Let $\mathbf{G}\in\mathbb{R}^{m\times n},\mathbf{g}\in\mathbb{R}^m,\mathbf{h}\in\mathbb{R}^n,g\in\mathbb{R}$ be independent with i.i.d. standard normal entries. Let $\Phi_1\subset\mathbb{R}^n,\Phi_2\subset\mathbb{R}^m$ be compact sets. Let $\psi(\cdot,\cdot):\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ be a continuous function. Assume, for all*

finite sets $S_1 \subset \Phi_1, S_2 \subset \Phi_2$ and $c \in \mathbb{R}$, we have,

$$\mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2} \{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\} \geq c) \geq \mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2} \{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c).$$

Then,

$$\mathbb{P}(\min_{\mathbf{x}\in \Phi_1} \max_{\mathbf{a}\in \Phi_2} \{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\} \geq c) \geq \mathbb{P}(\min_{\mathbf{x}\in \Phi_1} \max_{\mathbf{a}\in \Phi_2} \{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} > c).$$

Proof: Let $R(\Phi_i) = \sup_{\mathbf{v}\in\Phi_i} \|\mathbf{v}\|_2$ for $1 \leq i \leq 2$. Let $S_1 \subset \Phi_1, S_2 \subset \Phi_2$ be arbitrary ε-coverings of the sets Φ_1, Φ_2 so that, for any $\mathbf{v} \in \Phi_i$, there exists $\mathbf{v}' \in S_i$ satisfying $\|\mathbf{v}' - \mathbf{v}\|_2 \leq \varepsilon$. Furthermore, using continuity of ψ over the compact set $\Phi_1 \times \Phi_2$, for any $\delta > 0$, we can choose ε sufficiently small to guarantee that $|\psi(\mathbf{x},\mathbf{a}) - \psi(\mathbf{x}',\mathbf{a}')| < \delta$. Here δ can be made arbitrarily small as a function of ε. Now, for any $\mathbf{x} \in \Phi_1, \mathbf{a} \in \Phi_2$, pick $\mathbf{x}', \mathbf{a}'$ in the ε-coverings S_1, S_2. This gives,

$$|[\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})] - [\mathbf{a}'^T\mathbf{G}\mathbf{x}' - \psi(\mathbf{x}',\mathbf{a}')]| \leq \varepsilon(R(\Phi_1) + R(\Phi_2) + \varepsilon)\|\mathbf{G}\| + \delta \tag{A.5}$$

$$\begin{aligned}|[\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})] - [\|\mathbf{x}'\|_2\mathbf{g}^T\mathbf{a}' - \|\mathbf{a}'\|_2\mathbf{h}^T\mathbf{x}' - \psi(\mathbf{x}',\mathbf{a}')]| \\ \leq \varepsilon(R(\Phi_1) + R(\Phi_2) + \varepsilon)(\|\mathbf{g}\|_2 + \|\mathbf{h}\|_2) + \delta\end{aligned} \tag{A.6}$$

Next, using Lipschitzness of $\|\mathbf{g}\|_2, \|\mathbf{h}\|_2, \|\mathbf{G}\|$ and Fact 2.4, for $t > 1$, we have,

$$\mathbb{P}(\max\{\|\mathbf{g}\|_2 + \|\mathbf{h}\|_2, \|\mathbf{G}\|\} \leq t(\sqrt{n} + \sqrt{m})) \geq 1 - 4\exp(-\frac{(t-1)^2(m+n)}{2}) := p(t) \tag{A.7}$$

Let $C(t,\varepsilon) = t\varepsilon(R(\Phi_1) + R(\Phi_2) + \varepsilon)(\sqrt{m} + \sqrt{n}) + \delta$. Then, since (A.5) and (A.6) holds for all $\mathbf{a}, \mathbf{x}$, using (A.7),

$$\begin{aligned}\mathbb{P}(\min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{a}\in\Phi_2} \{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\} \geq c - C(t,\varepsilon)) \\ \geq \mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2} \{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c) - p(t)\end{aligned} \tag{A.8}$$

$$\begin{aligned}\mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2} \{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c) \\ \geq \mathbb{P}(\min_{\mathbf{x}\in \Phi_1} \max_{\mathbf{a}\in \Phi_2} \{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c + C(t,\varepsilon)) - p(t)\end{aligned} \tag{A.9}$$

Combining (A.8) and (A.9), for all $\varepsilon > 0, t > 1$, the following holds,

$$\mathbb{P}(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\} \geq c - C(t,\varepsilon)) \geq$$
$$\mathbb{P}(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c + C(t,\varepsilon)) - 2p(t)$$

Setting $t = \varepsilon^{-1/2}$ and letting $\varepsilon \to 0$, we obtain the desired result as $C(t,\varepsilon), \delta \to 0$ and $p(t) \to 1$. Here, we implicitly use the standard continuity results on the limits of decreasing and increasing sequence of events. ■

A.2.2 Symmetrization

To conclude, using Theorem 2.2 and Section A.2.1 we have,

$$\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\ \max_{\mathbf{a}\in\Phi_2}\ \left\{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\right\} \geq c\right) \geq$$
$$\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\ \max_{\mathbf{a}\in\Phi_2}\ \left\{\|\mathbf{x}\|_2\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|_2\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\right\} \geq c\right) := q. \tag{A.10}$$

Since $g \sim \mathcal{N}(0,1)$, we can write the left hand side of (A.10) as, $p = \frac{p_+ + p_-}{2}$ where we define p_+, p_-, p_0 as,

$$p_- = \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\ \max_{\mathbf{a}\in\Phi_2}\ \left\{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\right\} \geq c \ \middle|\ g \leq 0\right),$$
$$p_+ = \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\ \max_{\mathbf{a}\in\Phi_2}\ \left\{\mathbf{a}^T\mathbf{G}\mathbf{x} + \|\mathbf{a}\|_2\|\mathbf{x}\|_2 g - \psi(\mathbf{x},\mathbf{a})\right\} \geq c \ \middle|\ g > 0\right),$$
$$p_0 = \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\ \max_{\mathbf{a}\in\Phi_2}\ \left\{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\right\} \geq c\right)$$

By construction and independence of $g, \mathbf{G}$; $1 \geq p_+ \geq p_0 \geq p_-$. On the other hand, $1 - q \geq 1 - p \geq \frac{1-p_-}{2}$ which implies, $p_- \geq 2q - 1$. This further yields $p_0 \geq 2q - 1$, which is what we want. ■

A.3 The Dual of the LASSO

To derive the dual we write the problem in (3.33) equivalently as

$$\begin{aligned}\mathscr{F}(\mathbf{A},\mathbf{v}) = &\min_{\mathbf{w},\mathbf{b}} \{\|\mathbf{b}\|_2 + p(\mathbf{w})\} \\ &\text{s.t.} \quad \mathbf{b} = \mathbf{A}\mathbf{w} - \sigma\mathbf{v},\end{aligned}$$

and then reduce it to

$$\min_{\mathbf{w},\mathbf{b}} \max_{\boldsymbol{\mu}} \left\{ \|\mathbf{b}\|_2 + \boldsymbol{\mu}^T (\mathbf{b} - \mathbf{A}\mathbf{w} + \sigma\mathbf{v}) + p(\mathbf{w}) \right\}.$$

The dual of the problem above is

$$\max_{\boldsymbol{\mu}} \min_{\mathbf{w},\mathbf{b}} \{\|\mathbf{b}\|_2 + \boldsymbol{\mu}^T (\mathbf{b} - \mathbf{A}\mathbf{w} + \sigma\mathbf{v}) + p(\mathbf{w})\}. \tag{A.11}$$

The minimization over $\mathbf{b}$ above is easy to perform. A simple application of the Cauchy–Schwarz inequality gives

$$\begin{aligned}\|\mathbf{b}\|_2 + \boldsymbol{\mu}^T \mathbf{b} &\geq \|\mathbf{b}\|_2 - \|\mathbf{b}\|_2 \|\boldsymbol{\mu}\|_2 \\ &= (1 - \|\boldsymbol{\mu}\|_2) \|\mathbf{b}\|_2.\end{aligned}$$

Thus,

$$\min_{\mathbf{b}} \left\{ \|\mathbf{b}\|_2 + \boldsymbol{\mu}^T \mathbf{b} \right\} = \begin{cases} 0 & , \|\boldsymbol{\mu}\|_2 \leq 1, \\ -\infty & , o.w.. \end{cases}$$

Combining this with (A.11) we conclude that the dual problem of the problem in (3.33) is the following:

$$\max_{\|\boldsymbol{\mu}\|_2 \leq 1} \min_{\mathbf{w}} \left\{ \boldsymbol{\mu}^T (-\mathbf{A}\mathbf{w} + \sigma\mathbf{v}) + p(\mathbf{w}) \right\}.$$

We equivalently rewrite the dual problem in the format of a minimization problem as follows:

$$-\min_{\|\boldsymbol{\mu}\|_2\leq 1}\max_{\mathbf{w}}\left\{\boldsymbol{\mu}^T\left(\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\right)-p(\mathbf{w})\right\}. \tag{A.12}$$

If $p(\mathbf{w})$ is a finite convex function from $\mathbb{R}^n\to\mathbb{R}$, the problem in (3.33) is convex and satisfies Slater's conditions. When $p(\mathbf{w})$ is the indicator function of a convex set $\{\mathbf{w}\big|g(\mathbf{w})\leq 0\}$, the problem can be viewed as $\min_{g(\mathbf{w})\leq 0,\mathbf{b}}\left\{\|\mathbf{b}\|_2+\boldsymbol{\mu}^T\left(\mathbf{b}-\mathbf{A}\mathbf{w}+\sigma\mathbf{v}\right)\right\}$. For strong duality, we need strict feasibility, i.e., there must exist $\mathbf{w}$ satisfying $g(\mathbf{w})<0$. In our setup, $g(\mathbf{w})=f(\mathbf{x}_0+\mathbf{w})-f(\mathbf{x}_0)$ and $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$, hence strong duality holds and thus problems in (3.33) and (A.12) have the same optimal cost $\mathscr{F}(\mathbf{A},\mathbf{v})$.

A.4 Proofs for Section 3.5

A.4.1 Proof of Lemma 3.6

We prove the statements of the lemma in the order that they appear.

A.4.1.1 Scalarization

The first statement of Lemma 3.6 claims that the optimization problem in (3.43) can be reduced into a one dimensional optimization problem. To see this begin by evaluating the optimization over $\mathbf{w}$ for fixed $\|\mathbf{w}\|_2$:

$$\begin{aligned}
\mathscr{L}(\mathbf{g},\mathbf{h}) &= \min_{\mathbf{w}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2-\mathbf{h}^T\mathbf{w}+\max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}\\
&= \min_{\substack{\mathbf{w}:\|\mathbf{w}\|_2=\alpha\\ \alpha\geq 0}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2-\mathbf{h}^T\mathbf{w}+\max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}\\
&= \min_{\alpha\geq 0}\left\{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2+\min_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha}\left\{-\mathbf{h}^T\mathbf{w}+\max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}\right\}\\
&= \min_{\alpha\geq 0}\left\{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2-\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha}\left\{\mathbf{h}^T\mathbf{w}-\min_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\}\right\}\\
&= \min_{\alpha\geq 0}\left\{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2-\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha}\min_{\mathbf{s}\in\mathscr{C}}\left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\}\right\}
\end{aligned} \tag{A.13}$$

To further simplify (A.13), we use the following key observation as summarized in the Lemma below.

Lemma A.7 *Let $\mathscr{C} \in \mathbb{R}^n$ be a nonempty convex set in $\mathbb{R}^n$, $\mathbf{h} \in \mathbb{R}^n$ and $\alpha \geq 0$. Then,*

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} = \min_{\mathbf{s}\in\mathscr{C}} \max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\}.$$

Thus,

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} = \alpha \cdot \mathit{dist}(\mathbf{h},\mathscr{C}),$$

and the optimum is attained at $\mathbf{w}^* = \alpha \cdot \frac{\Pi(\mathbf{h},\mathscr{C})}{\mathit{dist}(\mathbf{h},\mathscr{C})}$.

Proof: First notice that

$$\min_{\mathbf{s}\in\mathscr{C}} \max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} (\mathbf{h}-\mathbf{s})^T\mathbf{w} = \min_{\mathbf{s}\in\mathscr{C}} \alpha\|\mathbf{h}-\mathbf{s}\|_2 = \alpha \cdot \mathrm{dist}(\mathbf{h},\mathscr{C}).$$

Furthermore, MinMax is never less than MaxMin [23]. Thus,

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} \leq \min_{\mathbf{s}\in\mathscr{C}} \max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} = \alpha \cdot \mathrm{dist}(\mathbf{h},\mathscr{C}).$$

It suffices to prove that

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} \geq \alpha \cdot \mathrm{dist}(\mathbf{h},\mathscr{C}).$$

Consider $\mathbf{w}^* = \alpha \cdot \frac{\Pi(\mathbf{h},\mathscr{C})}{\mathrm{dist}(\mathbf{h},\mathscr{C})}$. Clearly,

$$\max_{\mathbf{w}:\|\mathbf{w}\|_2=\alpha} \min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} \geq \min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}^*\right\}.$$

But,

$$\begin{aligned}
\min_{\mathbf{s}\in\mathscr{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}^*\right\} &= \frac{\alpha}{\mathrm{dist}(\mathbf{h},\mathscr{C})} \cdot \left(\mathbf{h}^T\Pi(\mathbf{h},\mathscr{C}) - \max_{\mathbf{s}\in\mathscr{C}} \mathbf{s}^T\Pi(\mathbf{h},\mathscr{C})\right) && \text{(A.14)}\\
&= \frac{\alpha}{\mathrm{dist}(\mathbf{h},\mathscr{C})} \cdot \left(\mathbf{h}^T\Pi(\mathbf{h},\mathscr{C}) - \mathrm{Proj}(\mathbf{h},\mathscr{C})^T\Pi(\mathbf{h},\mathscr{C})\right) && \text{(A.15)}\\
&= \alpha \cdot \mathrm{dist}(\mathbf{h},\mathscr{C}),
\end{aligned}$$

where (A.15) follows from Fact 2.2. This completes the proof of the Lemma. ■

Applying the result of Lemma A.7 to (A.13), we conclude that

$$\begin{aligned}\mathscr{L}(\mathbf{g},\mathbf{h}) &= \min_{\mathbf{w}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{h}^T\mathbf{w} + \max_{\mathbf{s}\in\mathscr{C}}\mathbf{s}^T\mathbf{w}\right\} \\ &= \min_{\alpha\geq 0}\left\{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2 - \alpha\cdot\text{dist}(\mathbf{h},\mathscr{C})\right\} \end{aligned} \tag{A.16}$$

A.4.1.2 Deterministic Result

The optimization problem in (A.16) is one dimensional and easy to handle. Setting the derivative of its objective function equal to zero and solving for the optimal α^*, under the assumption that

$$\|\mathbf{g}\|_2^2 > \text{dist}(\mathbf{h},\mathscr{C})^2, \tag{A.17}$$

it only takes a few simple calculations to prove the second statement of Lemma 3.6, i.e.

$$(\alpha^*)^2 = \|\mathbf{w}_{low}^*(\mathbf{g},\mathbf{h})\|_2^2 = \sigma^2\frac{\text{dist}^2(\mathbf{h},\mathscr{C})}{\|\mathbf{g}\|_2^2 - \text{dist}^2(\mathbf{h},\mathscr{C})}$$

and,

$$\mathscr{L}(\mathbf{g},\mathbf{h}) = \sigma\sqrt{\|\mathbf{g}\|_2^2 - \text{dist}^2(\mathbf{h},\mathscr{C})}. \tag{A.18}$$

A.4.1.3 Probabilistic Result

Next, we prove the high probability lower bound for $\mathscr{L}(\mathbf{g},\mathbf{h})$ implied by the last statement of Lemma 3.6. To do this, we will make use of concentration results for specific functions of Gaussian vectors as they are stated in Lemma A.3. Setting $t = \delta\sqrt{m}$ in Lemma A.3, with probability $1-8\exp(-c_0\delta^2 m)$,

$$\begin{aligned}&|\|\mathbf{g}\|_2^2 - m| \leq 2\delta m + \delta^2 m + 1, \\ &|\text{dist}^2(\mathbf{h},\mathscr{C}) - \mathbf{D}(\mathscr{C})| \leq 2\delta\sqrt{\mathbf{D}(\mathscr{C})m} + \delta^2 m + 1 \leq 2\delta m + \delta^2 m + 1.\end{aligned}$$

Combining these and using the assumption that $m \geq \mathbf{D}(\mathscr{C}) + \varepsilon_L m$, we find that

$$\begin{aligned}\|\mathbf{g}\|_2^2 - \mathrm{dist}^2(\mathbf{h}, \mathscr{C}) &\geq m - \mathbf{D}(\mathscr{C}) - [(2\delta^2 + 4\delta)m + 2] \\ &\geq m - \mathbf{D}(\mathscr{C}) - [(2\delta^2 + 4\delta)\frac{m - \mathbf{D}(\mathscr{C})}{\varepsilon_L} + 2] \\ &\geq (m - \mathbf{D}(\mathscr{C}))[1 - \frac{(2\delta^2 + 4\delta)}{\varepsilon_L}] - 2,\end{aligned}$$

with the same probability. Choose ε' so that $\sqrt{1-\varepsilon'} = 1 - \varepsilon$. Also, choose δ such that $\frac{(2\delta^2+4\delta)}{\varepsilon_L} < \frac{\varepsilon'}{2}$ and m sufficiently large to ensure $\varepsilon_L \varepsilon' m > 4$. Combined,

$$\|\mathbf{g}\|_2^2 - \mathrm{dist}^2(\mathbf{h}, \mathscr{C}) \geq (m - \mathbf{D}(\mathscr{C}))(1 - \frac{\varepsilon'}{2}) - 2 \geq (m - \mathbf{D}(\mathscr{C}))(1 - \varepsilon'), \tag{A.19}$$

with probability $1 - 8\exp(-c_0 \delta^2 m)$. Since the right hand side in (A.19) is positive, it follows from the second statement of Lemma 3.6 that

$$\mathscr{L}(\mathbf{g}, \mathbf{h}) \geq \sigma \sqrt{(m - \mathbf{D}(\mathscr{C}))(1 - \varepsilon')} = \sigma(1 - \varepsilon)\sqrt{m - \mathbf{D}(\mathscr{C})},$$

with the same probability. This concludes the proof.

A.4.2 Proof of Lemma 3.7

A.4.2.1 Scalarization

We have

$$\begin{aligned}\hat{\mathscr{U}}(\mathbf{g}, \mathbf{h}) &= -\min_{\|\mu\|_2 \leq 1} \max_{\|\mathbf{w}\|_2 = C_{up}} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + \|\boldsymbol{\mu}\|_2 \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathscr{C}} \mathbf{s}^T \mathbf{w} \right\} \\ &= -\min_{\|\mu\|_2 \leq 1} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + \max_{\|\mathbf{w}\|_2 = C_{up}} \left\{ \|\boldsymbol{\mu}\|_2 \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathscr{C}} \mathbf{s}^T \mathbf{w} \right\} \right\}. \end{aligned} \tag{A.20}$$

Notice that

$$\begin{aligned}\max_{\|\mathbf{w}\|_2 = C_{up}} \left\{ \|\boldsymbol{\mu}\|_2 \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathscr{C}} \mathbf{s}^T \mathbf{w} \right\} &= \max_{\|\mathbf{w}\|_2 = C_{up}} \min_{\mathbf{s} \in \mathscr{C}} (\|\boldsymbol{\mu}\|_2 \mathbf{h} - \mathbf{s})^T \mathbf{w} \\ &= C_{up} \mathrm{dist}(\|\boldsymbol{\mu}\|_2 \mathbf{h}, \mathscr{C}). \end{aligned} \tag{A.21}$$

where (A.21) follows directly from Lemma A.7. Combine (A.20) and (A.21) to conclude that

$$\begin{aligned}\hat{\mathscr{U}}(\mathbf{g},\mathbf{h}) &= -\min_{\|\mu\|_2\le 1}\left\{\sqrt{C_{up}^2+\sigma^2}\,\mathbf{g}^T\boldsymbol{\mu}+C_{up}\text{dist}(\|\boldsymbol{\mu}\|_2\mathbf{h},\mathscr{C})\right\}\\ &= -\min_{0\le\alpha\le 1}\left\{-\alpha\cdot\sqrt{C_{up}^2+\sigma^2}\,\|\mathbf{g}\|_2+C_{up}\text{dist}(\alpha\mathbf{h},\mathscr{C})\right\}. \end{aligned} \tag{A.22}$$

A.4.2.2 Deterministic Result

For convenience denote the objective function of problem (A.22) as

$$\phi(\alpha)=C_{up}\text{dist}(\alpha\mathbf{h},\mathscr{C})-\alpha\sqrt{C_{up}^2+\sigma^2}\|\mathbf{g}\|_2.$$

Notice that $\phi(\cdot)$ is convex. By way of justification, $\text{dist}(\alpha\mathbf{h},\mathscr{C})$ is a convex function for $\alpha\ge 0$ [182], and $\alpha\sqrt{C^2+\sigma^2}\|\mathbf{g}\|_2$ is linear in α. Denote $\alpha^*=\text{argmin}\,\phi(\alpha)$. Clearly, it suffices to show that $\alpha^*=1$. First, we prove that $\phi(\alpha)$ is differentiable as a function of α at $\alpha=1$. For this, we make use of the following lemma.

Lemma A.8 *Let C be a nonempty closed and convex set and $\mathbf{h}\notin C$. Then*

$$\lim_{\varepsilon\to 0}\frac{dist(\mathbf{h}+\varepsilon\mathbf{h},C)-dist(\mathbf{h},C)}{\varepsilon}=\langle\mathbf{h},\frac{\Pi(\mathbf{h},C)}{\|\Pi(\mathbf{h},C)\|_2}\rangle,$$

Proof: Let H be a hyperplane of $\mathscr{C}$ at $\text{Proj}(\mathbf{h},\mathscr{C})$ orthogonal to $\Pi(\mathbf{h},C)$. Using the second statement of Fact 2.2, H is a supporting hyperplane and $\mathbf{h}$ and C lie on different half planes induced by H (also see [23]). Also, observe that $\Pi(\mathbf{h},\mathscr{C})=\Pi(\mathbf{h},H)$ and $\text{Proj}(\mathbf{h},\mathscr{C})=\text{Proj}(\mathbf{h},H)$. Choose $\varepsilon>0$ sufficiently small such that $(1+\varepsilon)\mathbf{h}$ lies on the same half-plane as $\mathbf{h}$. We then have,

$$\|\Pi((1+\varepsilon)\mathbf{h},\mathscr{C})\|_2\ge\|\Pi((1+\varepsilon)\mathbf{h},H)\|_2=\|\Pi(\mathbf{h},\mathscr{C})\|_2+\left\langle\varepsilon\mathbf{h},\frac{\Pi(\mathbf{h},\mathscr{C})}{\|\Pi(\mathbf{h},\mathscr{C})\|_2}\right\rangle. \tag{A.23}$$

Denote the $n-1$ dimensional subspace that is orthogonal to $\Pi(\mathbf{h},H)$ and parallel to H by H_0. Decomposing $\varepsilon\mathbf{h}$ to its orthonormal components along $\Pi(\mathbf{h},H)$ and H_0, we have

$$\|\Pi((1+\varepsilon)\mathbf{h},C)\|_2^2\le\|(1+\varepsilon)\mathbf{h}-\text{Proj}(\mathbf{h},C)\|_2^2=\left(\|\Pi(\mathbf{h},C)\|_2+\left\langle\varepsilon\mathbf{h},\frac{\Pi(\mathbf{h},C)}{\|\Pi(\mathbf{h},C)\|_2}\right\rangle\right)^2+\varepsilon^2\|\text{Proj}(\mathbf{h},H_0)\|_2^2. \tag{A.24}$$

Take square roots in both sides of (A.24) and apply on the right hand side the useful inequality $\sqrt{a^2+b^2} \leq a+\frac{b^2}{2a}$, which is true for all $a,b \in \mathbb{R}^+$. Combine the result with the lower bound in (A.23) and let $\varepsilon \to 0$ to conclude the proof. ■

Since $\mathbf{h} \notin \mathscr{C}$, it follows from Lemma A.8, that $\text{dist}(\alpha\mathbf{h},\mathscr{C})$ is differentiable as a function of α at $\alpha = 1$, implying the same result for $\phi(\alpha)$. In fact, we have

$$\phi'(1) = C_{up}\text{dist}(\mathbf{h},\mathscr{C}) + C_{up}\frac{\langle \Pi(\mathbf{h},\mathscr{C}), \text{Proj}(\mathbf{h},\mathscr{C})\rangle}{\text{dist}(\mathbf{h},\mathscr{C})} - \sqrt{C_{up}^2+\sigma^2}\|\mathbf{g}\|_2 < 0,$$

where the negativity follows from assumption (3.45). To conclude the proof, we make use of the following simple lemma.

Lemma A.9 *Suppose $f : \mathbb{R} \to \mathbb{R}$ is a convex function, that is differentiable at $x_0 \in \mathbb{R}$ and $f'(x_0) < 0$. Then, $f(x) \geq f(x_0)$ for all $x \leq x_0$.*

Proof: By convexity of $f(\cdot)$, for all $x \leq x_0$:

$$\begin{aligned} f(x) &\geq f(x_0) + \underbrace{f'(x_0)}_{<0}\underbrace{(x-x_0)}_{\leq 0} \\ &\geq f(x_0) \end{aligned}$$

■

Applying Lemma A.9 for the convex function $\phi(\cdot)$ at $\alpha = 1$, gives that $\phi(\alpha) \geq \phi(1)$ for all $\alpha \in [0,1]$. Therefore, $\alpha^* = 1$.

A.4.2.3 Probabilistic Result

We consider the setting where m is sufficiently large and,

$$(1-\varepsilon_L)m \geq \max\left(\mathbf{D}(\mathscr{C}) + \mathbf{C}(\mathscr{C}), \mathbf{D}(\mathscr{C})\right), \quad \mathbf{D}(\mathscr{C}) \geq \varepsilon_L m \tag{A.25}$$

Choose $C_{up} = \sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}$ which would give $C_{up}^2 + \sigma^2 = \sigma^2\frac{m}{m-\mathbf{D}(\mathscr{C})}$. Hence, the assumption (3.45) in the second statement of Lemma 3.7 can be rewritten as,

$$\sqrt{m}\|\mathbf{g}\|_2\text{dist}(\mathbf{h},\mathscr{C}) > \sqrt{\mathbf{D}(\mathscr{C})}(\text{dist}(\mathbf{h},\mathscr{C})^2 + \text{corr}(\mathbf{h},\mathscr{C})). \tag{A.26}$$

The proof technique is as follows. We first show that (A.26) (and thus (3.45)) holds with high probability. Also, that $\mathbf{h} \notin \mathscr{C}$ with high probability. Then, as a last step we make use of the second statement of Lemma 3.7 to compute the lower bound on $\hat{\mathscr{U}}$.

• (3.45) *holds with high probability:*

Using standard concentration arguments (see Lemma A.2), we have

$$\sqrt{m}\|\mathbf{g}\|_2 \text{dist}(\mathbf{h},\mathscr{C}) \geq \sqrt{m}(\sqrt{m-1}-t)(\sqrt{\mathbf{D}(\mathscr{C})-1}-t)$$

with probability $1-4\exp\left(\frac{-t^2}{2}\right)$. Choose a sufficiently small constant $\delta > 0$ and set $t = \delta\sqrt{\mathbf{D}(\mathscr{C})}$ to ensure,

$$\sqrt{m}\|\mathbf{g}\|_2 \text{dist}(\mathbf{h},\mathscr{C}) \geq (1-\frac{\varepsilon_L}{2})m\sqrt{\mathbf{D}(\mathscr{C})} \tag{A.27}$$

with probability $1-\exp(-\mathscr{O}(m))$, where we used $(1-\varepsilon_L) \geq \mathbf{D}(\mathscr{C}) \geq \varepsilon_L m$. In particular, for sufficiently large $\mathbf{D}(\mathscr{C})$ we need $(1-\delta)^2 > 1-\frac{\varepsilon_L}{2}$.

Equation (A.27) establishes a high probability lower bound for the expression at the left hand side of (A.26). Next, we show that the expression at the right hand side of (A.26) is upper bounded with high probability by the same quantity.

Case 1: If $\mathscr{C}$ is a cone, $\text{corr}(\mathbf{h},\mathscr{C}) = 0$ and using Lemma A.3 $\text{dist}(\mathbf{h},\mathscr{C})^2 \leq \mathbf{D}(\mathscr{C}) + 2t\sqrt{\mathbf{D}(\mathscr{C})} + t^2 \leq (1-\varepsilon_L)m + 2t\sqrt{m} + t^2$ with probability $1-2\exp(-\frac{t^2}{2})$. Hence, we can choose $t = \varepsilon\sqrt{m}$ for a small constant $\varepsilon > 0$ to ensure, $\text{dist}(\mathbf{h},\mathscr{C})^2 < (1-\frac{\varepsilon_L}{2})m$ with probability $1-\exp(-\mathscr{O}(m))$. This gives (A.26) in combination with (A.27).

Case 2: Otherwise, from Lemma A.4, we have that $\mathbf{P}(\mathscr{C}) \leq 2(n+\mathbf{D}(\mathscr{C}))$ and from (A.25), $m \geq \mathbf{D}(\mathscr{C})$. Then, applying Lemma A.3, we have

$$\begin{aligned}
\text{dist}(\mathbf{h},\mathscr{C})^2 + \text{corr}(\mathbf{h},\mathscr{C}) &\leq \mathbf{D}(\mathscr{C}) + \mathbf{C}(\mathscr{C}) + 3t\underbrace{\sqrt{\mathbf{D}(\mathscr{C})}}_{\leq\sqrt{m}} + t\underbrace{\sqrt{\mathbf{P}(\mathscr{C})}}_{\leq\sqrt{2(n+m)}} + 2(t^2+1) \\
&\leq \mathbf{D}(\mathscr{C}) + \mathbf{C}(\mathscr{C}) + 3t\sqrt{m} + t\sqrt{2(n+m)} + 2(t^2+1) \\
&\leq (1-\varepsilon_L)m + 3t\sqrt{m} + t\sqrt{2(n+m)} + 2(t^2+1).
\end{aligned}$$

with probability $1-4\exp\left(\frac{-t^2}{2}\right)$. Therefore, with the same probability,

$$\sqrt{\mathbf{D}(\mathscr{C})}(\text{dist}(\mathbf{h},\mathscr{C})^2+\text{corr}(\mathbf{h},\mathscr{C})) \leq (1-\varepsilon_L)m\sqrt{\mathbf{D}(\mathscr{C})}+3t\sqrt{m}\sqrt{\mathbf{D}(\mathscr{C})} \\ +t\sqrt{2(n+m)}\sqrt{\mathbf{D}(\mathscr{C})}+2(t^2+1)\sqrt{\mathbf{D}(\mathscr{C})} \tag{A.28}$$

Comparing the right hand sides of inequalities A.27 and A.28 , we need to ensure that,

$$3t\sqrt{m}\sqrt{\mathbf{D}(\mathscr{C})}+t\sqrt{2(n+m)}\sqrt{\mathbf{D}(\mathscr{C})}+2(t^2+1)\sqrt{\mathbf{D}(\mathscr{C})} \leq \frac{\varepsilon_L}{2}m\sqrt{\mathbf{D}(\mathscr{C})} \iff \\ 3t\sqrt{m}+t\sqrt{2(n+m)}+2(t^2+1) \leq \frac{\varepsilon_L}{2}m. \tag{A.29}$$

Choose $t=\varepsilon\min\{\sqrt{m},\frac{m}{\sqrt{n}}\}$ for sufficiently small ε such that (A.29) and (A.26) then hold with probability $1-\exp\left(-\mathscr{O}\left(\min\{\frac{m^2}{n},m\}\right)\right)$.

Combining Case 1 and Case 2, (A.26) holds with probability $1-\exp\left(-\mathscr{O}\left(\gamma(m,n)\right)\right)$ where $\gamma(m,n)=m$ when $\mathscr{C}$ is cone and $\gamma(m,n)=\min\{\frac{m^2}{n},m\}$ otherwise.

• $\mathbf{h}\notin\mathscr{C}$ *with high probability:*

Apply Lemma A.2 on $\text{dist}(\mathbf{h},\mathscr{C})$ with $t=\varepsilon\sqrt{\mathbf{D}(\mathscr{C})}$ to show that $\text{dist}(\mathbf{h},\mathscr{C})$ is strictly positive. This proves that $\mathbf{h}\notin\mathscr{C}$, with probability $1-\exp(-\mathscr{O}\left(\mathbf{D}(\mathscr{C})\right))$=$1-\exp(-\mathscr{O}\left(m\right))$.

• *High probability lower bound for* $\hat{\mathscr{U}}$*:*

Thus far we have proved that assumptions $\mathbf{h}\notin\mathscr{C}$ and (3.45) of the second statement in Lemma 3.7 hold with the desired probability. Therefore, (3.46) holds with the same high probability, namely,

$$\hat{\mathscr{U}}(\mathbf{g},\mathbf{h})=\frac{\sigma}{\sqrt{m-\mathbf{D}(\mathscr{C})}}\left(\sqrt{m}\|\mathbf{g}\|_2-\sqrt{\mathbf{D}(\mathscr{C})}\text{dist}(\mathbf{h},\mathscr{C})\right) \tag{A.30}$$

We will use similar concentration arguments as above to upper bound the right hand side of (A.30). For any $t>0$:

$$\sqrt{m}\|\mathbf{g}\|_2 \leq m+t\sqrt{m}$$
$$\sqrt{\mathbf{D}(\mathscr{C})}\text{dist}(\mathbf{h},\mathscr{C}) \geq \sqrt{\mathbf{D}(\mathscr{C})}(\sqrt{\mathbf{D}(\mathscr{C})-1}-t)$$

with probability $1-4\exp(-\frac{t^2}{2})$. Thus,

$$\sqrt{m}\|\mathbf{g}\|_2-\sqrt{\mathbf{D}(\mathscr{C})}\text{dist}(\mathbf{h},\mathscr{C})\leq m-\mathbf{D}(\mathscr{C})+t(\sqrt{m}+\sqrt{\mathbf{D}(\mathscr{C})})+1. \tag{A.31}$$

For a given constant $\varepsilon>0$, substitute (A.31) in (A.30) and choose $t=\varepsilon'\sqrt{m}$ (for some sufficiently small constant $\varepsilon'>0$), to ensure that,

$$\hat{\mathscr{U}}(\mathbf{g},\mathbf{h})\leq(1+\varepsilon)\sigma\sqrt{m-\mathbf{D}(\mathbf{x}_0,\lambda)}$$

with probability $1-4\exp(\frac{-\varepsilon'^2m}{2})$. Combining this with the high probability events of all previous steps, we obtain the desired result.

A.4.3 Proof of Lemma 3.8

A.4.3.1 Scalarization

The reduction of $\mathscr{L}_{dev}(\mathbf{g},\mathbf{h})$ to an one-dimensional optimization problem follows identically the steps as in the proof for $\mathscr{L}(\mathbf{g},\mathbf{h})$ in Section A.4.1.1.

A.4.3.2 Deterministic Result

From the first statement of Lemma 3.8,

$$\mathscr{L}_{dev}(\mathbf{g},\mathbf{h})=\min_{\alpha\in S_{dev}}\left\{\underbrace{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2-\alpha\cdot\text{dist}(\mathbf{h},\mathscr{C})}_{:=L(\alpha)}\right\}, \tag{A.32}$$

where we have denoted the objective function as $L(\alpha)$ for notational convenience. It takes no much effort (see also statements 1 and 2 of Lemma A.10) to prove that $L(\cdot)$:

- is a strictly convex function,
- attains its minimum at

$$\alpha^*(\mathbf{g},\mathbf{h})=\frac{\sigma\cdot\text{dist}(\mathbf{h},\mathscr{C})}{\sqrt{\|\mathbf{g}\|_2^2-\text{dist}^2(\mathbf{h},\mathscr{C})}}.$$

The minimization of $L(\alpha)$ in (A.32) is restricted to the set S_{dev}. Also, by assumption (3.48), $\alpha^*(\mathbf{g},\mathbf{h})\notin S_{dev}$. Strict convexity implies then that the minimum of $L(\cdot)$ over $\alpha\in S_{dev}$ is attained at the boundary points of

the set S_{dev}, i.e. at $(1 \pm \delta_{dev})C_{dev}$ [23]. Thus, $\mathscr{L}_{dev}(\mathbf{g},\mathbf{h}) = L((1 \pm \delta_{dev})C_{dev})$, which completes the proof.

A.4.3.3 Probabilistic Result

Choose $C_{dev} = \sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}}$ and consider the regime where $(1-\varepsilon_L)m > \mathbf{D}(\mathscr{C}) > \varepsilon_L m$ for some constant $\varepsilon_L > 0$. $\delta_{dev} > 0$ is also a constant.

• **Mapping $\mathscr{L}_{dev}$ to Lemma A.10:** It is helpful for the purposes of the presentation to consider the function

$$L(x) := L(x;a,b) = \sqrt{x^2+\sigma^2}a - xb, \tag{A.33}$$

over $x \geq 0$, and a,b are positive parameters. Substituting a,b,x with $\|\mathbf{g}\|_2, \text{dist}(\mathbf{h},\mathscr{C}), \alpha$, we can map $L(x;a,b)$ to our function of interest,

$$L(\alpha;\|\mathbf{g}\|_2,\text{dist}(\mathbf{h},\mathscr{C})) = \sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\|_2 - \alpha\text{dist}(\mathbf{h},\mathscr{C}).$$

In Lemma A.10 we have analyzed useful properties of the function $L(x;a,b)$, which are of key importance for the purposes of this proof. This lemma focuses on perturbation analysis and investigates $L(x';a',b') - L(x;a,b)$ where x',a',b' are the perturbations from the fixed values x,a,b. In this sense, a',b' correspond to $\|\mathbf{g}\|_2, \text{dist}(\mathbf{h},\mathscr{C})$ which are probabilistic quantities and a,b correspond to $\sqrt{m}, \sqrt{\mathbf{D}(\mathscr{C})}$, i.e. the approximate means of the former ones.

In what follows, we refer continuously to statements of Lemma A.10 and use them to complete the proof of the "Probabilistic result" of Lemma 3.8. Let us denote the minimizer of $L(x;a,b)$ by $x^*(a,b)$. To see how the definitions above are relevant to our setup, it follows from the first statement of Lemma A.10 that,

$$L\left(x^*(\sqrt{m},\sqrt{\mathbf{D}(\mathscr{C})});\sqrt{m},\sqrt{\mathbf{D}(\mathscr{C})}\right) = \sigma\sqrt{m-\mathbf{D}(\mathscr{C})}, \tag{A.34}$$

and

$$x^*(\sqrt{m},\sqrt{\mathbf{D}(\mathscr{C})}) = \sigma\sqrt{\frac{\mathbf{D}(\mathscr{C})}{m-\mathbf{D}(\mathscr{C})}} = C_{dev}, \tag{A.35}$$

• **Verifying assumption (3.48):** Going back to the proof, we begin by proving that assumption (3.48) of the second statement of Lemma 3.8 is valid with high probability. Observe that from the definition of S_{dev}

and (A.35), assumption (3.48) can be equivalently written as

$$\left| \frac{x^*(\|\mathbf{g}\|_2, \mathrm{dist}(\mathbf{h}, \mathscr{C}))}{x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathscr{C})})} - 1 \right| \leq \delta_{dev}. \tag{A.36}$$

On the other hand, from the third statement of Lemma A.10 there exists sufficiently small constant $\varepsilon_1 > 0$ such that (A.36) is true for all $\mathbf{g}$ and $\mathbf{h}$ satisfying

$$|\|\mathbf{g}\|_2 - \sqrt{m}| \leq \varepsilon_1 \sqrt{m} \quad \text{and} \quad |\mathrm{dist}(\mathbf{h}, \mathscr{C}) - \sqrt{\mathbf{D}(\mathscr{C})}| \leq \varepsilon_1 \sqrt{m}. \tag{A.37}$$

Furthermore, for large enough $\mathbf{D}(\mathscr{C})$ and from basic concentration arguments (see Lemma A.2), $\mathbf{g}$ and $\mathbf{h}$ satisfy (A.37) with probability $1 - 2\exp(-\frac{\varepsilon_1^2 m}{2})$. This proves that assumption (3.48) holds with the same high probability.

• **Lower bounding $\mathscr{L}_{dev}$:** From the deterministic result of Lemma 3.8, once (3.48) is satisfied then

$$\mathscr{L}_{dev}(\mathbf{g}, \mathbf{h}) = L\left((1 \pm \delta_{dev}) C_{dev}; \|\mathbf{g}\|_2, \mathrm{dist}(\mathbf{h}, \mathscr{C})\right). \tag{A.38}$$

Thus, to prove (3.49) we will show that there exists $t > 0$ such that

$$L\left((1 \pm \delta_{dev}) C_{dev}; \|\mathbf{g}\|_2, \mathrm{dist}(\mathbf{h}, \mathscr{C})\right) \geq (1 + t)\sigma \sqrt{m - \mathbf{D}(\mathscr{C})}, \tag{A.39}$$

with high probability. Equivalently, using (A.34), it suffices to show that there exists a constant $t > 0$ such that

$$L\left((1 \pm \delta_{dev}) x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathscr{C})}); \|\mathbf{g}\|_2, \mathrm{dist}(\mathbf{h}, \mathscr{C})\right) - L\left(x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathscr{C})}); \sqrt{m}, \sqrt{\mathbf{D}(\mathscr{C})}\right) \geq t\sigma\sqrt{m}, \tag{A.40}$$

with high probability. Applying the sixth statement of Lemma A.10 with $\gamma \leftarrow \delta_{dev}$, for any constant $\delta_{dev} > 0$, there exists constants t, ε_2 such that (A.40) holds for all $\mathbf{g}$ and $\mathbf{h}$ satisfying

$$|\|\mathbf{g}\|_2 - \sqrt{m}| \leq \varepsilon_2 \sqrt{m} \quad \text{and} \quad |\mathrm{dist}(\mathbf{h}, \mathscr{C}) - \sqrt{\mathbf{D}(\mathscr{C})}| \leq \varepsilon_2 \sqrt{m},$$

which holds with probability $1 - 2\exp(-\frac{\varepsilon_2^2 m}{2})$ for sufficiently large $\mathbf{D}(\mathscr{C})$. Thus, (A.40) is true with the same high probability.

Union bounding over the events that (A.36) and (A.40) are true, we end up with the desired result. The

reason is that with high probability (A.38) and (A.40) hold, i.e.,

$$\mathscr{L}_{dev}(\mathbf{g},\mathbf{h}) = L((1\pm\delta_{dev})C_{dev};\|\mathbf{g}\|_2,\mathrm{dist}(\mathbf{h},\mathscr{C})) \geq L\left(x^*(\sqrt{m},\sqrt{\mathbf{D}(\mathscr{C})});\sqrt{m},\sqrt{\mathbf{D}(\mathscr{C})}\right)+t\sigma\sqrt{m}$$
$$= \sigma\sqrt{m-\mathbf{D}(\mathscr{C})}+t\sigma\sqrt{m}.$$

A.5 Deviation Analysis: Key Lemma

Lemma A.10 *Consider the following function over $x \geq 0$:*

$$L(x) := L(x;a,b) = \sqrt{x^2+\sigma^2}a - xb$$

where $\sigma > 0$ is constant and a,b are positive parameters satisfying $(1-\varepsilon)a > b > \varepsilon a$ for some constant $\varepsilon > 0$. Denote the minimizer of $L(x;a,b)$ by $x^*(a,b)$. Then,

1. *$x^*(a,b) = \frac{\sigma b}{\sqrt{a^2-b^2}}$ and $L(x^*(a,b);a,b) = \sigma\sqrt{a^2-b^2}$.*
2. *For fixed a and b, $L(x;a,b)$ is strictly convex in $x \geq 0$.*
3. *For any constant $\eta > 0$, there exists sufficiently small constant $\varepsilon_1 > 0$, such that*

$$\left|\frac{x^*(a',b')}{x^*(a,b)} - 1\right| \leq \eta,$$

   for all a',b' satisfying $|a'-a| < \varepsilon_1 a$ and $|b'-b| < \varepsilon_1 a$.

4. *There exists positive constant $\eta > 0$, such that, for sufficiently small constant $\varepsilon_1 > 0$,*

$$\left|L(x^*(a,b);a',b') - L(x^*(a,b),a,b)\right| \leq \eta\varepsilon_1\sigma a,$$

   for all a',b' satisfying $|a'-a| < \varepsilon_1 a$ and $|b'-b| < \varepsilon_1 a$.

5. *For any constant $\gamma > 0$, there exists a constant $\varepsilon_2 > 0$ such that for sufficiently small constant $\varepsilon_1 > 0$,*

$$L(x;a',b') - L(x^*(a,b);a',b') \geq \varepsilon_2\sigma a,$$

   for all x,a' and b' satisfying $|x-x^*(a,b)| > \gamma x^*(a,b)$, $|a'-a| < \varepsilon_1 a$ and $|b'-b| < \varepsilon_1 a$.

6. *For any constant $\gamma > 0$, there exists a constant $\varepsilon_2 > 0$ such that for sufficiently small constant $\varepsilon_1 > 0$,*

$$L(x;a',b') - L(x^*(a,b);a,b) \geq \varepsilon_2 \sigma a,$$

for all x, a' and b' satisfying $|x - x^*(a,b)| > \gamma x^*(a,b)$, $|a' - a| < \varepsilon_1 a$ and $|b' - b| < \varepsilon_1 a$.

7. *Given $c_{low} > 0$, consider the restricted optimization, $\min_{x \geq c_{low}} L(x;a,b)$. We have,*

$$\lim_{c_{low} \to \infty} \min_{x \geq c_{low}} L(x;a,b) \to \infty$$

Proof: First statement: The derivative (w.r.t. x) of $L(x;a,b)$ is:

$$L'(x;a,b) = \frac{ax}{\sqrt{x^2 + \sigma^2}} - b.$$

Setting this to 0, using strict convexity and solving for x, we obtain the first statement.

Second statement: The second derivative is,

$$L''(x;a,b) = \frac{a\sqrt{x^2+\sigma^2} - \frac{ax^2}{\sqrt{x^2+\sigma^2}}}{x^2+\sigma^2} = \frac{a\sigma^2}{(x^2+\sigma^2)^{3/2}} > 0,$$

for all $x \geq 0$. Consequently, f is strictly convex.

Third statement: We can write,

$$\left| x^*(a',b') - x^*(a,b) \right| = \sigma \left| \frac{b'}{\sqrt{a'^2 - b'^2}} - \frac{b}{\sqrt{a^2 - b^2}} \right|.$$

Observe that $x^*(a,b) = \frac{b}{\sqrt{a^2-b^2}}$ is decreasing in a and increasing in b as long as $a > b \geq 0$. Also, for sufficiently small constant ε_1, we have, $a', b' > 0$ for all $|a' - a| < \varepsilon_1 a, |b' - b| < \varepsilon_1 a$. Therefore,

$$\frac{b - \varepsilon_1 a}{\sqrt{(a+\varepsilon_1 a)^2 - (b - \varepsilon_1 a)^2}} \leq \frac{b'}{\sqrt{a'^2 - b'^2}} \leq \frac{b + \varepsilon_1 a}{\sqrt{(a-\varepsilon_1 a)^2 - (b + \varepsilon_1 a)^2}}.$$

Now, for any constant $\delta > 0$, we can choose ε_1 sufficiently small such that both $b - \varepsilon_1 a$ and $b + \varepsilon_1 a$ lie in the interval $(1 \pm \delta)b$. Similarly, $(a \pm \varepsilon_1 a)^2 - (b \mp \varepsilon_1 a)^2$ can be also chosen to lie in the interval $(1 \pm \delta)(a^2 - b^2)$. Combining, we obtain the desired result

$$\left| \frac{b'}{\sqrt{a'^2 - b'^2}} - \frac{b}{\sqrt{a^2 - b^2}} \right| < \eta(\delta) \frac{b}{\sqrt{a^2 - b^2}}.$$

Fourth statement: For $|a-a'|<\varepsilon_1 a$ and $|b-b'|<\varepsilon_1 a$, we have,

$$|L(x^*(a,b);a',b')-L(x^*(a,b);a,b)|=\frac{\sigma}{\sqrt{a^2-b^2}}|(aa'-bb')-(a^2-b^2)|\leq \varepsilon_1\sigma\frac{|a^2+ab|}{a^2-b^2}.$$

By assumption, $(1-\varepsilon)a>b>\varepsilon a$. Thus,

$$\varepsilon_1\sigma\frac{|a^2+ab|}{a^2-b^2}\leq \varepsilon_1\sigma\frac{2a^2}{2\varepsilon a^2}=\frac{\varepsilon_1\sigma}{\varepsilon}.$$

Choosing ε_1 sufficiently small, we conclude with the desired result.

Fifth statement: We will show the statement for a sufficiently small γ. Notice that, as γ gets larger, the set $|x-x^*(a,b)|\geq\gamma x^*(a,b)$ gets smaller hence, proof for small γ implies the proof for larger γ.

Using the Third Statement, choose ε_1 to ensure that $|x^*(a',b')-x^*(a,b)|<\gamma x^*(a,b)$ for all $|a'-a|<\varepsilon_1 a$ and $|b'-b|<\varepsilon_1 a$. For each such a',b', since $L(x,a',b')$ is a strictly convex function of x and the minimizer $x^*(a',b')$ lies between $(1\pm\gamma)x^*(a,b)$ we have,

$$L(x,a',b')\geq\min\{L((1-\gamma)x^*(a,b),a',b'),L((1+\gamma)x^*(a,b),a',b')\},$$

for all $|x-x^*(a,b)|>\gamma x^*(a,b)$. In summary, we simply need to characterize the increase in the function value at the points $(1\pm\gamma)x^*(a,b)$.

We have that,

$$L((1\pm\gamma)x^*(a,b);a',b')=\frac{\sigma}{\sqrt{a^2-b^2}}(\sqrt{a^2+(\pm2\gamma+\gamma^2)b^2}a'-(1\pm\gamma)bb'),\tag{A.41}$$

and

$$L(x^*(a,b);a',b')=\frac{\sigma}{\sqrt{a^2-b^2}}(aa'-bb').\tag{A.42}$$

In the following discussion, without loss of generality, we consider only the "$+\gamma$" case in (A.41) since the exact same argument works for the "$-\gamma$" case as well.

Subtracting (A.42) from (A.41) and discarding the constant in front, we will focus on the following

quantity,

$$\begin{aligned}\text{diff}(\gamma) &= (\sqrt{a^2+(2\gamma+\gamma^2)b^2}a' - (1+\gamma)bb') - (aa' - bb') \\ &= (\underbrace{\sqrt{a^2+(2\gamma+\gamma^2)b^2}}_{:=g(\gamma)} - a)a' - \gamma bb'. \end{aligned} \tag{A.43}$$

To find a lower bound for $g(\gamma)$, write

$$\begin{aligned} g(\gamma) &= \sqrt{a^2+(2\gamma+\gamma^2)b^2} \\ &= \sqrt{(a+\gamma\frac{b^2}{a})^2 + \gamma^2(b^2 - \frac{b^4}{a^2})} \\ &\geq (a+\gamma\frac{b^2}{a}) + \frac{\gamma^2(b^2-\frac{b^4}{a^2})}{4(a+\gamma\frac{b^2}{a})}, \end{aligned} \tag{A.44}$$

where we have assumed $\gamma \leq 1$ and used the fact that $(a+\gamma\frac{b^2}{a})^2 \geq a^2 \geq b^2 - \frac{b^4}{a^2}$. Equation (A.44) can be further lower bounded by,

$$g(\gamma) \geq (a+\gamma\frac{b^2}{a}) + \frac{\gamma^2(a^2b^2-b^4)}{8a^3}$$

Combining with (A.43) , we find that,

$$\text{diff}(\gamma) \geq \gamma(\frac{b^2}{a}a' - bb') + \gamma^2\frac{a^2b^2-b^4}{8a^3}a'. \tag{A.45}$$

Consider the second term on the right hand side of the inequality in (A.45). Choosing $\varepsilon_1 < 1/2$, we ensure, $a' \geq a/2$, and thus,

$$\gamma^2\frac{a^2b^2-b^4}{8a^3}a' \geq \gamma^2\frac{a^2b^2-b^4}{16a^2} \geq \gamma^2\frac{\varepsilon a^2b^2}{16a^2} = \gamma^2\varepsilon\frac{b^2}{16}. \tag{A.46}$$

Next, consider the other term in (A.45). We have,

$$\left(\frac{b^2}{a}a' - bb'\right) = \frac{b^2}{a}(a'-a) - b(b'-b) \geq -\left(\left|\frac{b^2}{a}(a'-a)\right| + |b(b'-b)|\right).$$

Choosing ε_1 sufficiently small (depending only on γ), we can ensure that,

$$\left|\frac{b^2}{a}(a'-a)\right| + |b(b'-b)| < \gamma\varepsilon\frac{b^2}{32}. \tag{A.47}$$

Combining (A.45), (A.46) and (A.47), we conclude that there exists sufficiently small constant $\varepsilon_1 > 0$ such that,

$$\text{diff}(\gamma) \geq \gamma^2 \varepsilon \frac{b^2}{32}.$$

Multiplying with $\frac{\sigma}{\sqrt{a^2-b^2}}$, we end up with the desired result since $\frac{b^2}{\sqrt{a^2-b^2}} \geq \frac{\varepsilon^2}{\sqrt{1-\varepsilon^2}} a$.

Sixth statement: The last statement can be deduced from the fourth and fifth statements. Given $\gamma > 0$, choose $\varepsilon_1 > 0$ sufficiently small to ensure,

$$L(x; a', b') - L(x^*(a,b), a', b') \geq \varepsilon_2 \sigma a$$

and

$$|L(x^*(a,b); a, b) - L(x^*(a,b), a', b')| \geq \eta \varepsilon_1 \sigma a$$

Using the triangle inequality,

$$\begin{aligned} L(x; a', b') - L(x^*(a,b), a, b) &\geq L(x; a', b') - L(x^*(a,b), a', b') - |L(x^*(a,b), a', b') - L(x^*(a,b), a, b)| \\ &\geq (\varepsilon_2 - \eta \varepsilon_1) \sigma a. \end{aligned} \tag{A.48}$$

Choosing ε_1 to further satisfy $\eta \varepsilon_1 < \frac{\varepsilon_2}{2}$, (A.48) is guaranteed to be larger than $\frac{\varepsilon_2}{2} \sigma a$ which gives the desired result.

Seventh statement: To show this, we may use $a > b$ and simply write,

$$L(x; a, b) \geq (a-b)x \implies \lim_{c_{low} \to \infty} \min_{x \geq c_{low}} L(x; a, b) \geq \lim_{c_{low} \to \infty} (a-b) c_{low} = \infty$$

■

A.6 Proof of Lemma 3.20

Proof of the Lemma requires some work. We prove the statements in the specific order that they appear.

Statement 1: We have

$$\begin{aligned} n = \mathbb{E}\left[\|\mathbf{h}\|_2^2\right] = \mathbb{E}\left[\|\text{Proj}_\lambda(\mathbf{h}) + \mathbf{h} - \text{Proj}_\lambda(\mathbf{h})\|_2^2\right] &= \mathbb{E}[\|\text{Proj}_\lambda(\mathbf{h})\|_2^2] + \mathbb{E}[\|\Pi_\lambda(\mathbf{h})\|_2^2] + 2\mathbb{E}[\langle \Pi_\lambda(\mathbf{h}), \text{Proj}_\lambda(\mathbf{h}) \rangle] \\ &= \mathbf{P}(\lambda \partial f(\mathbf{x}_0)) + \mathbf{D}(\lambda \partial f(\mathbf{x}_0)) + 2\mathbf{C}(\lambda \partial f(\mathbf{x}_0)). \end{aligned}$$

Statement 2: We have $\text{Proj}_0(\mathbf{h}) = \mathbf{0}$ and $\Pi_0(\mathbf{h}) = \mathbf{h}$, and the statement follows easily.

Statement 3: Let $r = \inf_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \|\mathbf{s}\|_2$. Then, for any $\lambda \geq 0$, $\|\text{Proj}_\lambda(\mathbf{v})\|_2 \geq \lambda\|\mathbf{s}\|_2$, which implies $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ $\geq \lambda^2\|\mathbf{s}\|_2^2$. Letting $\lambda \to \infty$, we find $\mathbf{P}(\lambda\partial f(\mathbf{x}_0)) \to \infty$.

Similarly, for any $\mathbf{h}$, application of the triangle inequality gives

$$\|\Pi_\lambda(\mathbf{h})\|_2 \geq \lambda r - \|\mathbf{h}\|_2 \implies \|\Pi_\lambda(\mathbf{h})\|_2^2 \geq \lambda^2 r^2 - 2\lambda r\|\mathbf{h}\|_2.$$

Let $\mathbf{h} \sim \mathcal{N}(0, I)$ and take expectations in both sides of the inequality above. Recalling that $\mathbb{E}[\|\mathbf{h}\|_2] \leq \sqrt{n}$, and letting $\lambda \to \infty$, we find $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \to \infty$.

Finally, since $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) + \mathbf{P}(\lambda\partial f(\mathbf{x}_0)) + 2\mathbf{C}(\lambda\partial f(\mathbf{x}_0)) = n$, $\mathbf{C}(\lambda\partial f(\mathbf{x}_0)) \to -\infty$ as $\lambda \to \infty$. This completes the proof.

Statement 4: Continuity of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ follows from Lemma $B.2$ in Amelunxen et al. [4]. We will now show continuity of $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ and continuity of $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ will follow from the fact that $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ is a continuous function of $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$.

Recall that $\text{Proj}_\lambda(\mathbf{v}) = \lambda\text{Proj}_1(\frac{\mathbf{v}}{\lambda})$. Also, given $\mathbf{v}_1, \mathbf{v}_2$, we have,

$$\|\text{Proj}_\lambda(\mathbf{v}_1) - \text{Proj}_\lambda(\mathbf{v}_2)\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2$$

Consequently, given $\lambda_1, \lambda_2 > 0$,

$$\begin{aligned}
\|\text{Proj}_{\lambda_1}(\mathbf{v}) - \text{Proj}_{\lambda_2}(\mathbf{v})\|_2 &= \|\lambda_1\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1}) - \lambda_2\text{Proj}_1(\frac{\mathbf{v}}{\lambda_2})\|_2 \\
&\leq |\lambda_1 - \lambda_2|\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \|\lambda_2(\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1}) - \text{Proj}_1(\frac{\mathbf{v}}{\lambda_2}))\|_2 \\
&\leq |\lambda_1 - \lambda_2|\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \lambda_2\|\mathbf{v}\|_2\frac{|\lambda_1 - \lambda_2|}{\lambda_1\lambda_2} \\
&= |\lambda_1 - \lambda_2|(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})
\end{aligned}$$

Hence, setting $\lambda_2 = \lambda_1 + \varepsilon$,

$$\|\text{Proj}_{\lambda_2}(\mathbf{v})\|_2^2 \leq [\|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2 + \varepsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})]^2$$

which implies,

$$\|\text{Proj}_{\lambda_2}(\mathbf{v})\|_2^2 - \|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2^2 \leq 2\varepsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})\|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2 + \varepsilon^2(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})$$

Similarly, using $\|\text{Proj}_{\lambda_2}(\mathbf{v})\|_2 \geq \|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2 - \varepsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})$, we find,

$$\|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2^2 - \|\text{Proj}_{\lambda_2}(\mathbf{v})\|_2^2 \leq 2\varepsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})\|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2\lambda_1)$$

Combining these, we always have,

$$|\|\text{Proj}_{\lambda_2}(\mathbf{v})\|_2^2 - \|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2^2| \leq 2\varepsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})\|\text{Proj}_{\lambda_1}(\mathbf{v})\|_2 + \varepsilon^2(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\|_2 + \frac{\|\mathbf{v}\|_2}{\lambda_1})$$

Now, letting $\mathbf{v} \sim \mathcal{N}(0,I)$ and taking the expectation of both sides and letting $\varepsilon \to 0$, we conclude with the continuity of $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ for $\lambda > 0$.

To show continuity at 0, observe that, for any $\lambda > 0$, we have, $\|\text{Proj}_\lambda(\mathbf{v})\|_2 \leq R\lambda$ where $R = \sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \|\mathbf{s}\|_2$. Hence,

$$|\mathbf{P}(\lambda\partial f(\mathbf{x}_0)) - \mathbf{P}_f(\mathbf{x}_0,0)| = \mathbf{P}(\lambda\partial f(\mathbf{x}_0)) \leq R^2\lambda^2$$

As $\lambda \to 0$, $\mathbf{P}(\lambda\partial f(\mathbf{x}_0)) = 0$.

Statement 5: For a proof see Lemma $B.2$ in [4].

Statement 6: Based on Lemma A.11, given vector $\mathbf{v}$, set $\mathscr{C}$ and scalar $1 \geq c > 0$, we have,

$$\frac{\|\text{Proj}(c\mathbf{v},\mathscr{C})\|_2}{c} \geq \|\text{Proj}(\mathbf{v},\mathscr{C})\|_2$$

Given $\lambda_1 > \lambda_2 > 0$, this gives,

$$\|\text{Proj}(\mathbf{v},\lambda_1\partial f(\mathbf{x}_0))\|_2 = \lambda_1\|\text{Proj}(\frac{\mathbf{v}}{\lambda_1},\partial f(\mathbf{x}_0))\|_2 \geq \lambda_1\frac{\lambda_2}{\lambda_1}\|\text{Proj}(\frac{\mathbf{v}}{\lambda_2},\partial f(\mathbf{x}_0))\|_2 = \|\text{Proj}(\mathbf{v},\lambda_2\partial f(\mathbf{x}_0))\|_2$$

Since this is true for all $\mathbf{v}$, choosing $\mathbf{v} \sim \mathcal{N}(0,I)$, we end up with $\mathbf{D}_f(\mathbf{x}_0,\lambda_1) \geq \mathbf{D}_f(\mathbf{x}_0,\lambda_2)$.

Finally, at 0 we have $\mathbf{D}_f(\mathbf{x}_0,0) = 0$ and by definition $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \geq 0$ which implies the increase at $\lambda = 0$. For the rest of the discussion, given three points A,B,C in $\mathbb{R}^n$, the angle induced by the lines AB and BC will be denoted by $A\hat{B}C$.

Lemma A.11 *Let $\mathscr{C}$ be a convex and closed set in $\mathbb{R}^n$. Let $\mathbf{z}$ and $0 < \alpha < 1$ be arbitrary, let $\mathbf{p}_1 = Proj(\mathbf{z},\mathscr{C})$,*

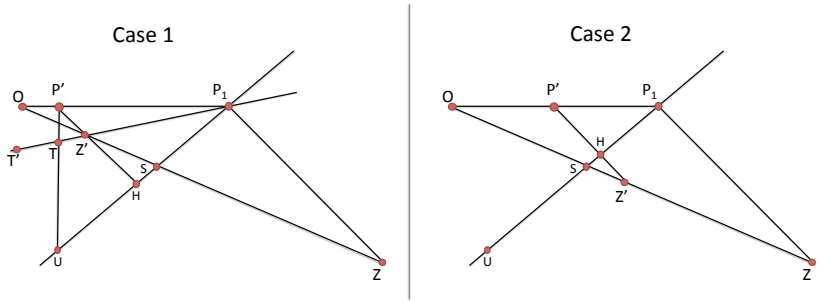


Figure A.1: Possible configurations of the points in Lemma A.11 when $Z\hat{P}_1O$ is wide angle.

$\mathbf{p}_2 = Proj(\alpha\mathbf{z}, \mathscr{C})$. *Then,*

$$\|\mathbf{p}_1\|_2 \leq \frac{\|\mathbf{p}_2\|_2}{\alpha}$$

Proof:

Denote the points whose coordinates are determined by $0, \mathbf{p}_1, \mathbf{p}_2, \mathbf{z}$ by O, P_1, P_2 and Z respectively. We start by reducing the problem to a two dimensional one. Obtain $\mathscr{C}'$ by projecting the set $\mathscr{C}$ to the $2D$ plane induced by the points Z, P_1 and O. Now, let $\mathbf{p}_2' = \text{Proj}(\alpha\mathbf{z}, \mathscr{C}')$. Due to the projection, we still have: $\|\mathbf{z} - \mathbf{p}_2'\|_2 \leq \|\mathbf{z} - \mathbf{p}_2\|_2$ and $\|\mathbf{p}_2'\|_2 \leq \|\mathbf{p}_2\|_2$. We wish to prove that $\|\mathbf{p}_2'\|_2 \geq \|\alpha\mathbf{p}_1\|_2$. Figures A.1 and A.2 will help us explain our approach.

Let the line UP_1 be perpendicular to ZP_1. Let $P'Z'$ be parallel to P_1Z_1. Observe that P' corresponds to $\alpha\mathbf{p}_1$. H is the intersection of $P'Z'$ and P_1U. Denote the point corresponding to $\mathbf{p}_2'$ by P_2'. Observe that P_2' satisfies the following:

- P_1 is the closest point to Z in $\mathscr{C}$ hence P_2' lies on the side of P_1U which doesn't include Z.
- P_2 is the closest point to Z'. Hence, $Z'\hat{P}_2P_1$ is not acute angle. Otherwise, we can draw a perpendicular to P_2P_1 from Z' and end up with a shorter distance. This would also imply that $Z'\hat{P}_2'P_1$ is not acute as well as $Z'P_1$ stays same but $|Z'P_2'| \leq |Z'P_2|$ and $|P_2'P_1| \leq |P_2P_1|$.

We will do the proof case by case.

When $Z\hat{P}_1O$ is wide angle: Assume $Z\hat{P}_1O$ is wide angle and UP_1 crosses ZO at S.

Based on these observations, we investigate the problem in two cases illustrated by Figure A.1.

Case 1 (S lies on $Z'Z$): Consider the lefthand side of Figure A.1. If P_2' lies on the triangle $P'P_1H$ then $O\hat{P}'P_2' > O\hat{P}'Z$ which implies $O\hat{P}'P_2'$ is wide angle and $|OP_2'| \geq |OP'|$. If P_2' lies on the region induced by $OP'Z'T'$ then $P_1\hat{P}_2'Z'$ is acute angle as $P_1\hat{Z}'P_2' > P_1\hat{Z}'O$ is wide, which contradicts with $P_1\hat{P}_2'Z'$ is not acute.

Finally, let U be chosen so that $P'U$ is perpendicular to OP_1. Then, if P_2' lies on the quadrilateral $UTZ'H$

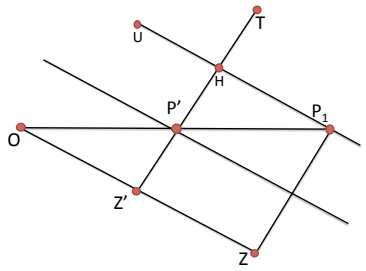


Figure A.2: Lemma A.11 when $Z\hat{P}_1O$ is acute or right angle.

then $|OP_2'| \geq |OP'|$ as $O\hat{P}'P_2'$ is wide or right angle. If it lies on the remaining region $T'TU$, then $Z'\hat{P}_2'P_1$ is acute. The reason is, $P_2'\hat{Z}'P_1$ is wide as follows:

$$P_2'\hat{Z}'P_1 \geq U\hat{Z}'P_1 > U\hat{T}P_1 > U\hat{P}'P_1 = \frac{\pi}{2}$$

Case 2 (S lies on OZ'): Consider the righthand side of Figure A.1. Due to location restrictions, P_2' lies on either $P_1P'H$ triangle or the region induced by $OP'HU$. If it lies on $P_1P'H$ then, $O\hat{P}'P_2' > O\hat{P}'H$ which implies $|OP_2'| \geq |OP'|$ as $O\hat{P}'P_2'$ is wide angle.

If P_2' lies on $OP'HU$ then, $P_1\hat{P}_2'Z' < P_1\hat{H}Z' = \frac{\pi}{2}$ hence $P_1\hat{P}_2'Z'$ is acute angle which cannot happen as it was discussed in the list of properties of P_2'.

When $Z\hat{P}_1O$ is right or acute angle: Consider Figure A.2. P_2' lies above UP_1. It cannot belong to the region induced by UHT as it would imply $Z'\hat{P}_2'P_1 < Z'\hat{H}P_1 \leq \frac{\pi}{2}$. Then, it belongs to the region induced by THP_1 which implies the desired result as $O\hat{P}'P_2'$ is at least right angle.

In all cases, we end up with $|OP_2'| \geq |OP'|$ which implies $\|\mathbf{p}_2\|_2 \geq \|\mathbf{p}_2'\|_2 \geq \alpha\|\mathbf{p}_1\|_2$ as desired.

■

Statement 7: For a proof see Lemma $B.2$ in [4].

Statement 8: From Statement 7, $\mathbf{C}(\lambda\partial f(\mathbf{x}_0)) = -\frac{\lambda}{2}\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda}$. Also from Statement 5, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is strictly convex. Thus, $\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda} \leq 0$ for all $\lambda \in [0, \lambda_{\text{best}}]$ which yields $\mathbf{C}(\lambda\partial f(\mathbf{x}_0)) \geq 0$ for all $\lambda \in [0, \lambda_{\text{best}}]$. Similarly, $\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda} \geq 0$ for all $\lambda \in [\lambda_{\text{best}}, \infty)$ which yields $\mathbf{C}(\lambda\partial f(\mathbf{x}_0)) \leq 0$ for all $\lambda \in [\lambda_{\text{best}}, \infty)$. Finally, λ_{best} minimizes $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$. Hence $\frac{d\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{d\lambda}|_{\lambda=\lambda_{\text{best}}} = 0$ which yields $\mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{best}}) = 0$.

Statement 9: We prove that for any $0 \leq \lambda_1 < \lambda_2 \leq \lambda_{\text{best}}$,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda_1) + \mathbf{C}_f(\mathbf{x}_0, \lambda_1) > \mathbf{D}_f(\mathbf{x}_0, \lambda_2) + \mathbf{C}_f(\mathbf{x}_0, \lambda_2). \tag{A.49}$$

From Statement 5, $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is strictly decreasing for $\lambda \in [0, \lambda_{\text{best}}]$. Thus,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda_1) > \mathbf{D}_f(\mathbf{x}_0, \lambda_2). \tag{A.50}$$

Furthermore, from Statement 6, $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ is an increasing function of λ. Thus,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda_1) + 2\mathbf{C}_f(\mathbf{x}_0, \lambda_1) \geq \mathbf{D}_f(\mathbf{x}_0, \lambda_2) + 2\mathbf{C}_f(\mathbf{x}_0, \lambda_2). \tag{A.51}$$

where we have used Statement 1. Combining (A.50) and (A.51), we conclude with (A.49), as desired.

A.7 Explicit formulas for well-known functions

A.7.1 ℓ_1 minimization

Let $\mathbf{x}_0 \in \mathbb{R}^n$ be a k sparse vector and let $\beta = \frac{k}{n}$. Then, we have the following when $f(\cdot) = \|\cdot\|_1$,

- $\frac{\mathbf{D}(\lambda\partial f(\mathbf{x}_0))}{n} = (1+\lambda^2)(1-(1-\beta)\text{erf}(\frac{\lambda}{\sqrt{2}})) - \sqrt{\frac{2}{\pi}}(1-\beta)\lambda\exp(-\frac{\lambda^2}{2})$
- $\frac{\mathbf{P}(\lambda\partial f(\mathbf{x}_0))}{n} = \beta\lambda^2 + (1-\beta)[\text{erf}(\frac{\lambda}{\sqrt{2}}) + \lambda^2\text{erfc}(\frac{\lambda}{\sqrt{2}}) - \sqrt{\frac{2}{\pi}}\lambda\exp(-\frac{\lambda^2}{2})]$
- $\frac{\mathbf{C}(\lambda\partial f(\mathbf{x}_0))}{n} = -\lambda^2\beta + (1-\beta)[\sqrt{\frac{2}{\pi}}\lambda\exp(-\frac{\lambda^2}{2}) - \lambda^2\text{erfc}(\frac{\lambda}{\sqrt{2}})]$

These are not difficult to obtain. For example, to find $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, pick $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$ and consider the vector $\Pi(\mathbf{g}, \lambda\partial f(\mathbf{x}_0))$. The distance vector to the subdifferential of the ℓ_1 norm takes the form of soft thresholding on the entries of $\mathbf{g}$. In particular,

$$(\Pi(\mathbf{g}, \lambda\partial f(\mathbf{x}_0)))_i = \begin{cases} \mathbf{g}(i) - \lambda\cdot\text{sgn}(\mathbf{x}_0(i)) & \text{if } \mathbf{x}_0(i) \neq 0, \\ \text{shrink}_\lambda(\mathbf{g}(i)) & \text{otherwise.} \end{cases}$$

where $\text{shrink}_\lambda(\mathbf{g}(i))$ is the soft thresholding operator defined as,

$$\text{shrink}_\lambda(x) = \begin{cases} x-\lambda & \text{if } \ x > \lambda, \\ 0 & \text{if } \ |x| \leq \lambda, \\ x+\lambda & \text{if } \ x < -\lambda. \end{cases}$$

Consequently, we obtain our formulas after taking the expectation of $\mathbf{g}(i) - \lambda\cdot\text{sgn}(\mathbf{x}_0(i))$ and $\text{shrink}_\lambda(\mathbf{g}(i))$. For more details on these formulas, the reader is referred to [72, 74, 83, 192] which calculate the phase

transitions of ℓ_1 minimization.

A.7.1.1 Closed form bound

We will now find a closed form bound on $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ for the same sparse signal $\mathbf{x}_0$. In particular, we will show that $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq (\lambda^2+2)k$ for $\lambda \geq \sqrt{2\log\frac{n}{k}}$. Following the above discussion and letting $\mathbf{g} \sim \mathcal{N}(0,\mathbf{I}_n)$, first observe that, $\mathbb{E}[(\mathbf{g}_i - \lambda\cdot\mathrm{sgn}(\mathbf{x}_0(i)))^2] = \lambda^2+1$

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) = \sum \mathbb{E}[(\mathbf{g}(i)-\lambda\cdot\mathrm{sgn}(\mathbf{x}_0(i)))^2] + (n-k)\,\mathbb{E}[\mathrm{shrink}_\lambda(\mathbf{g}(i))^2] \tag{A.52}$$

The sum on the left hand side is simply $(\lambda^2+1)k$. The interesting term is $\mathrm{shrink}_\lambda(\mathbf{g}(i))$. To calculate this, we will use the following lemma.

Lemma A.12 *Let x be a nonnegative random variable. Assume, there exists $c>0$ such that for all $t>0$,*

$$\mathbb{P}(x \geq c+t) \leq \exp(-\frac{t^2}{2})$$

For any $a \geq 0$, we have,

$$\mathbb{E}[\mathit{shrink}_{a+c}(x)^2] \leq \frac{2}{a^2+1}\exp(-\frac{a^2}{2}).$$

Proof: Let $Q(t) = \mathbb{P}(x \geq t)$.

$$\begin{aligned}
\mathbb{E}[\mathrm{shrink}_{a+c}(x)^2] &= \int_{a+c}^{\infty}(x-a-c)^2 d(-Q(x)) \\
&\leq -[Q(x)(x-a-c)^2]_{a+c}^{\infty} + \int_{a+c}^{\infty} Q(x)d(x-a-c)^2 = \int_{a+c}^{\infty} Q(x)d(x-a-c)^2 \qquad \text{(A.53)} \\
&\leq \int_{a+c}^{\infty} 2(x-a-c)Q(x)d(x-a-c) \leq 2\int_{a+c}^{\infty}(x-a-c)\exp(-\frac{(x-c)^2}{2})d(x-a-c) \\
&\leq 2\int_{a}^{\infty}(u-a)\exp(-\frac{u^2}{2})du \leq 2\exp(-\frac{a^2}{2}) - 2a\frac{a}{a^2+1}\exp(-\frac{a^2}{2}) = \frac{2}{a^2+1}\exp(-\frac{a^2}{2}) \qquad \text{(A.54)}
\end{aligned}$$

(A.53) follows from integration by parts and (A.54) follows from the standard result on Gaussian tail bound, $\int_a^\infty \exp(-\frac{u^2}{2})du \geq \frac{a}{a^2+1}\exp(-\frac{a^2}{2})$ ■

To calculate $\mathbb{E}[\mathrm{shrink}_\lambda(g)^2]$ for $g \sim \mathcal{N}(0,1)$ we make use of the standard fact about Gaussian distribution, $\mathbb{P}(|g|>t) \leq \exp(-\frac{t^2}{2})$. Applying the Lemma A.12 with $c=0$ and $a=\lambda$ yields, $\mathbb{E}[|\mathrm{shrink}_\lambda(g)|^2] \leq$

$\frac{2}{\lambda^2+1}\exp(-\frac{\lambda^2}{2})$. Combining this with (A.52), we find,

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq (\lambda^2+1)k + \frac{2n}{\lambda^2+1}\exp(-\frac{\lambda^2}{2})$$

For $\lambda \geq \sqrt{2\log\frac{n}{k}}$, $\exp(-\frac{\lambda^2}{2}) \leq \frac{k}{n}$. Hence, we obtain,

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq (\lambda^2+1)k + \frac{2k}{\lambda^2+1} \leq (\lambda^2+3)k$$

A.7.2 Nuclear norm minimization

Assume $\mathbf{X}_0$ is a $d\times d$ matrix of rank r and $\mathbf{x}_0$ is its vector representation where $n=d^2$ and we choose nuclear norm to exploit the structure. Denote the spectral norm of a matrix by $\|\cdot\|_2$. Assume $\mathbf{X}_0$ has skinny singular value decomposition $\mathbf{U}\Sigma\mathbf{V}^T$ where $\Sigma\in\mathbb{R}^{r\times r}$. Define the "support" subspace of $\mathbf{X}_0$ as,

$$S_{\mathbf{X}_0} = \{\mathbf{M}\in\mathbb{R}^{d\times d}\big|(\mathbf{I}-\mathbf{U}\mathbf{U}^T)\mathbf{M}(\mathbf{I}-\mathbf{V}\mathbf{V}^T)=0\}$$

The subdifferential of nuclear norm is given as,

$$\partial\|\mathbf{X}_0\|_\star = \{\mathbf{S}\in\mathbb{R}^{d\times d}\big|\text{Proj}(\mathbf{S},S_{\mathbf{X}_0})=\mathbf{U}\mathbf{V}^T,\text{ and }\|\text{Proj}(\mathbf{S},S_{\mathbf{X}_0}^\perp)\|\leq 1\}$$

Based on this, we wish to calculate $\text{dist}(\mathbf{G},\lambda\partial f(\mathbf{x}_0))$ when $\mathbf{G}$ has i.i.d. standard normal entries. As it has been discussed in [77, 163, 165], $\Pi(\mathbf{G},\lambda\partial f(\mathbf{x}_0))$ effectively behaves as singular value soft thresholding. In particular, we have,

$$\Pi(\mathbf{G},\lambda\partial f(\mathbf{x}_0)) = (\text{Proj}(\mathbf{G},S_{\mathbf{X}_0})-\lambda\mathbf{U}\mathbf{V}^T) + \sum_{i=1}^{n-r}\text{shrink}_\lambda(\sigma_{\mathbf{G},i})\mathbf{u}_{\mathbf{G},i}\mathbf{v}_{\mathbf{G},i}^T$$

where $\text{Proj}(\mathbf{G},S_{\mathbf{X}_0}^\perp)$ has singular value decomposition $\sum_{i=1}^{n-r}\sigma_{\mathbf{G},i}\mathbf{u}_{\mathbf{G},i}\mathbf{v}_{\mathbf{G},i}^T$.

Based on this behavior, $\text{dist}(\mathbf{G},\lambda\partial f(\mathbf{x}_0))$ has been analyzed in various works in the linear regime where $\frac{r}{d}$ is constant. This is done by using the fact that the singular value distribution of a $d\times d$ matrix approaches to the quarter circle law when singular values are normalized by $\sqrt{d}$

$$\psi(x) = \begin{cases}\frac{1}{\pi}\sqrt{4-x^2}\text{ if }0\leq x\leq 2\\ 0\text{ else}\end{cases}.$$

Based on ψ, define the quantities related to the moments of tail of ψ. Namely,

$$\Psi_i(x) = \int_x^\infty x^i \psi(x) dx$$

We can now give the following explicit formulas for the asymptotic behavior of $\partial \|\mathbf{X}_0\|_\star$ where $\frac{r}{d} = \beta$ is fixed. Define,

$$\upsilon = \frac{\lambda}{2\sqrt{1-\beta}}$$

- $\frac{\mathbf{D}_f(\mathbf{x}_0, \lambda\sqrt{d})}{n} = [2\beta - \beta^2 + \beta\lambda^2] + [(1-\beta)\lambda^2\Psi_0(\upsilon) + (1-\beta)^2\Psi_2(\upsilon) - 2(1-\beta)^{3/2}\lambda\Psi_1(\upsilon)]$
- $\frac{\mathbf{P}_f(\mathbf{x}_0, \lambda\sqrt{d})}{n} = \beta\lambda^2 + (1-\beta)\lambda^2\Psi_0(\upsilon) + (1-\beta)^2(1-\Psi_2(\upsilon))$
- $\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda\sqrt{d})}{n} = -\lambda^2\beta - (1-\beta)\lambda^2\Psi_0(\upsilon) + (1-\beta)^{3/2}\lambda\Psi_1(\upsilon)$

A.7.2.1 Closed form bounds

Our approach will exactly follow the proof of Proposition 3.11 in [50]. Given $\mathbf{G}$ with i.i.d. standard normal entries, the spectral norm of the off-support term $\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}^\perp)$ satisfies,

$$\mathbb{P}(\|\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}^\perp)\|_2 \geq 2\sqrt{d-r} + t) \leq \exp(-\frac{t^2}{2})$$

It follows that all singular values of $\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}^\perp)$ satisfies the same inequality as well. Consequently, for any singular value and for $\lambda \geq 2\sqrt{d-r}$, applying Lemma A.12, we may write,

$$\mathbb{E}[\text{shrink}_\lambda(\sigma_{\mathbf{G},i})^2] \leq \frac{2}{(\lambda - 2\sqrt{d-r})^2 + 1} \exp(-\frac{(\lambda - 2\sqrt{d-r})^2}{2}) \leq 2$$

It follows that,

$$\sum_{i=1}^{d-r} \mathbb{E}[\text{shrink}_\lambda(\sigma_{\mathbf{G},i})^2] \leq 2(d-r)$$

To estimate the in-support terms, we need to consider $\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}) - \lambda\mathbf{U}\mathbf{V}^T$. Since $\lambda\mathbf{U}\mathbf{V}^T$ and $\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0})$ are independent, we have,

$$\|\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}) - \lambda\mathbf{U}\mathbf{V}^T\|_F^2 = \lambda^2 r + |S_{\mathbf{X}_0}| = \lambda^2 r + 2dr - r^2$$

Combining, we find,

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq \lambda^2 r + 2dr - r^2 + 2d - 2r \leq (\lambda^2 + 2d)r + 2d$$

A.7.3 Block sparse signals

Let $n = t \times b$ and assume entries of $\mathbf{x}_0 \in \mathbb{R}^n$ can be partitioned into t blocks of size b so that only k of these t blocks are nonzero. To induce the structure, use the $\ell_{1,2}$ norm which sums up the ℓ_2 norms of the blocks, [91, 175, 191]. In particular, denoting the subvector corresponding to i'th block of $\mathbf{x}$ by $\mathbf{x}_i$

$$\|\mathbf{x}\|_{1,2} = \sum_{i=1}^{t} \|\mathbf{x}_i\|$$

To calculate $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)), \mathbf{C}(\lambda\partial f(\mathbf{x}_0)), \mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ with $f(\cdot) = \|\cdot\|_{1,2}$, pick $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$ and consider $\Pi(\mathbf{g}, \lambda\partial\|\mathbf{x}_0\|_{1,2})$ and $\text{Proj}(\mathbf{g}, \lambda\partial\|\mathbf{x}_0\|_{1,2})$. Similar to ℓ_1 norm and the nuclear norm, distance to subdifferential will correspond to a "soft-thresholding". In particular, $\Pi(\mathbf{g}, \lambda\partial\|\mathbf{x}_0\|_{1,2})$ has been studied in [175, 191] and is given as,

$$\Pi(\mathbf{g}, \lambda\partial\|\mathbf{x}_0\|_{1,2}) = \begin{cases} \mathbf{g}_i - \lambda\frac{\mathbf{x}_{0,i}}{\|\mathbf{x}_{0,i}\|} & \text{if} \quad \mathbf{x}_{0,i} \neq 0 \\ \text{vshrink}_\lambda(\mathbf{g}_i) & \text{else} \end{cases}$$

where the vector shrinkage vshrink_λ is defined as,

$$\text{vshrink}_\lambda(\mathbf{v}) = \begin{cases} \mathbf{v}(1 - \frac{\lambda}{\|\mathbf{v}\|}) & \text{if} \quad \|\mathbf{v}\| > \lambda \\ 0 \quad \text{if} \quad \|\mathbf{v}\| \leq \lambda \end{cases}$$

When $\mathbf{x}_{0,i} \neq 0$ and $\mathbf{g}_i$ is i.i.d. standard normal, $\mathbb{E}[\|\mathbf{g}_i - \lambda\frac{\mathbf{x}_{0,i}}{\|\mathbf{x}_{0,i}\|^2}\|^2] = \mathbb{E}[\|\mathbf{g}_i\|^2] + \lambda^2 = b + \lambda^2$. Calculation of $\text{vshrink}_\lambda(\mathbf{g}_i)$ and has to do with the tails of χ^2-distribution with b degrees of freedom (see Section 3 of [175]). Similar to previous section, define the tail function of a χ^2-distribution with b degrees of freedom as,

$$\Psi_i(x) = \int_x^\infty x^i \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1}\exp(-\frac{x}{2})dx$$

Then, $\mathbb{E}[\|\text{vshrink}_\lambda(\mathbf{g}_i)\|^2] = \Psi_1(\lambda^2) + \Psi_0(\lambda^2)\lambda^2 - 2\Psi_{\frac{1}{2}}(\lambda^2)\lambda$. Based on this, we calculate $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$, $\mathbf{P}(\lambda\partial f(\mathbf{x}_0))$ and $\mathbf{C}(\lambda\partial f(\mathbf{x}_0))$ as follows.

- $\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) = k(b + \lambda^2) + [\Psi_1(\lambda^2) + \Psi_0(\lambda^2)\lambda^2 - 2\Psi_{\frac{1}{2}}(\lambda^2)\lambda](t - k)$

- $\mathbf{P}(\lambda\partial f(\mathbf{x}_0)) = \lambda^2 k + [(\Psi_1(0) - \Psi_1(\lambda^2)) + \lambda^2\Psi_0(\lambda^2)](t-k)$
- $\mathbf{C}(\lambda\partial f(\mathbf{x}_0)) = -\lambda^2 k + [\lambda\Psi_{\frac{1}{2}}(\lambda^2) - \lambda^2\Psi_0(\lambda^2)](t-k)$

A.7.3.1 Closed form bound

Similar to Proposition 3 of [101], we will make use of the following bound for a x distributed with χ^2-distribution with b degrees of freedom.

$$\mathbb{P}(\sqrt{x} \geq \sqrt{b} + t) \leq \exp(-\frac{t^2}{2}) \quad \text{for all } t > 0 \tag{A.55}$$

Now, the total contribution of nonzero blocks to $\mathbf{D}(\lambda\partial f(\mathbf{x}_0))$ is simply $(\lambda^2+b)k$ as $\mathbb{E}[\|\mathbf{g}_i - \lambda\frac{\mathbf{x}_{0,i}}{\|\mathbf{x}_{0,i}\|}\|^2] = \lambda^2 + b$. For the remaining, we need to estimate $\mathbb{E}[\|\text{vshrink}_\lambda(\mathbf{g}_i)\|^2]$ for an i.i.d. standard normal $\mathbf{g}_i \in \mathbb{R}^d$. Using Lemma A.12, with $c = \sqrt{b}$ and $a = \lambda - \sqrt{b}$ and using the tail bound (A.55), we obtain,

$$\mathbb{E}[\|\text{vshrink}_\lambda(\mathbf{g}_i)\|^2] \leq \frac{2}{(\lambda-\sqrt{b})^2+1}\exp(-\frac{(\lambda-\sqrt{b})^2}{2})$$

Combining everything,

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq k(\lambda^2+b) + \frac{2t}{(\lambda-\sqrt{b})^2+1}\exp(-\frac{(\lambda-\sqrt{b})^2}{2})$$

Setting $\lambda \geq \sqrt{b} + \sqrt{2\log\frac{t}{k}}$, we ensure, $\exp(-\frac{(\lambda-\sqrt{b})^2}{2}) \leq \frac{k}{t}$, hence,

$$\mathbf{D}(\lambda\partial f(\mathbf{x}_0)) \leq k(\lambda^2+b) + \frac{2k}{(\lambda-\sqrt{b})^2+1} \leq k(\lambda^2+b+2)$$

A.8 Gaussian Width of the Widened Tangent Cone

The results in this appendix will be useful to show the stability of ℓ_2^2-LASSO for all $\tau > 0$. Recall the definition of Gaussian width from Chapter 2. The following lemma provides a Gaussian width characterization of "widening of a tangent cone".

Lemma A.13 *Assume $f(\cdot)$ is a convex function and $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$. Given $\varepsilon_0 > 0$, consider the ε_0-widened tangent cone defined as,*

$$\mathscr{T}_f(\mathbf{x}_0, \varepsilon_0) = Cl(\{\alpha\cdot\mathbf{w} \big| f(\mathbf{x}_0+\mathbf{w}) \leq f(\mathbf{x}_0) + \varepsilon_0\|\mathbf{w}\|_2,\ \alpha \geq 0\}) \tag{A.56}$$

Let $R_{min} = \min_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \|\mathbf{s}\|_2$ *and* $\mathscr{B}^{n-1}$ *be the unit* ℓ_2*-ball in* $\mathbb{R}^n$*. Then,*

$$\omega(\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}) \leq \omega(\mathscr{T}_f(\mathbf{x}_0)\cap\mathscr{B}^{n-1}) + \frac{\varepsilon_0\sqrt{n}}{R_{min}}$$

Proof: Let $\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)$. Write $\mathbf{w}=\mathbf{w}_1+\mathbf{w}_2$ via Moreau's decomposition theorem (Fact 2.1) where $\mathbf{w}_1\in\mathscr{T}_f(\mathbf{x}_0)$ and $\mathbf{w}_2\in\text{cone}(\partial f(\mathbf{x}_0))$ and $\mathbf{w}_1^T\mathbf{w}_2=0$. Here we used the fact that $\mathbf{x}_0$ is not a minimizer and $\mathscr{T}_f(\mathbf{x}_0)^* = \text{cone}(\partial f(\mathbf{x}_0))$. To find a bound on $\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)$ in terms of $\mathscr{T}_f(\mathbf{x}_0)$, our intention will be to find a reasonable bound on $\mathbf{w}_2$ and to argue $\mathbf{w}$ cannot be far away from its projection on the tangent cone.

To do this, we will make use of the followings.

- If $\mathbf{w}_2\neq 0$, since $\mathbf{w}_1^T\mathbf{w}_2=0$, $\max_{\mathbf{s}\in\partial f(\mathbf{x}_0)}\mathbf{w}_1^T\mathbf{s}=0$.
- Assume $\mathbf{w}_2\neq 0$. Then $\mathbf{w}_2=\alpha\mathbf{s}(\mathbf{w}_2)$ for some $\alpha>0$ and $\mathbf{s}(\mathbf{w}_2)\in\partial f(\mathbf{x}_0)$.

From convexity, for any $1>\varepsilon>0$, $\varepsilon\varepsilon_0\|\mathbf{w}\|_2 \geq f(\varepsilon\mathbf{w}+\mathbf{x}_0)-f(\mathbf{x}_0)$. Now, using Proposition 3.3 with $\delta\to 0$, we obtain,

$$\begin{aligned}\varepsilon_0\|\mathbf{w}\|_2 \geq \lim_{\varepsilon\to 0}\frac{f(\varepsilon\mathbf{w}+\mathbf{x}_0)-f(\mathbf{x}_0)}{\varepsilon} &= \sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)}\mathbf{w}^T\mathbf{s}\\ &\geq \mathbf{w}^T\mathbf{s}(\mathbf{w}_2) = \mathbf{w}_1^T\mathbf{s}(\mathbf{w}_2)+\mathbf{w}_2^T\mathbf{s}(\mathbf{w}_2)\\ &= \|\mathbf{w}_2\|_2\|\mathbf{s}(\mathbf{w}_2)\|_2 \geq \|\mathbf{w}_2\|_2 R_{min}\end{aligned}$$

This gives, $\frac{\|\mathbf{w}_2\|_2}{\|\mathbf{w}\|_2}\leq\frac{\varepsilon_0}{R_{min}}$. Equivalently, for a unit size $\mathbf{w}$, $\|\mathbf{w}_2\|_2\leq\frac{\varepsilon_0}{R_{min}}$.

What remains is to estimate the Gaussian width of $\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}$. Let $\mathbf{g}\sim\mathscr{N}(0,\mathbf{I}_n)$. $\mathbf{w}_1,\mathbf{w}_2$ still denote the projection of $\mathbf{w}$ onto $\mathscr{T}_f(\mathbf{x}_0)$ and $\text{cone}(\partial f(\mathbf{x}_0))$ respectively.

$$\begin{aligned}\omega(\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}) &= \mathbb{E}[\sup_{\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}}\mathbf{w}^T\mathbf{g}]\\ &\leq \mathbb{E}[\sup_{\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}}\mathbf{w}_1^T\mathbf{g}]+\mathbb{E}[\sup_{\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}}\mathbf{w}_2^T\mathbf{g}]\end{aligned}$$

Observe that, for $\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}$, $\|\mathbf{w}_2\|_2\leq\frac{\varepsilon_0}{R_{min}}$,

$$\mathbb{E}[\sup_{\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}}\mathbf{w}_2^T\mathbf{g}] \leq \mathbb{E}[\sup_{\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}}\|\mathbf{w}_2\|_2\|\mathbf{g}\|_2] \leq \frac{\varepsilon_0}{R_{min}}\mathbb{E}\|\mathbf{g}\|_2 \leq \frac{\varepsilon_0\sqrt{n}}{R_{min}}$$

For $\mathbf{w}_1$, we have $\mathbf{w}_1 \in \mathscr{T}_f(\mathbf{x}_0)$ and $\|\mathbf{w}_1\|_2 \leq \|\mathbf{w}\|_2 \leq 1$ which gives,

$$\mathbb{E}[\sup_{\mathbf{w}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}} \mathbf{w}_1^T\mathbf{g}] \leq \mathbb{E}[\sup_{\mathbf{w}'\in\mathscr{T}_f(\mathbf{x}_0)\cap\mathscr{B}^{n-1}} \mathbf{w}'^T\mathbf{g}] = \omega(\mathscr{T}_f(\mathbf{x}_0)\cap\mathscr{B}^{n-1})$$

Combining these individual bounds, we find,

$$\omega(\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}) \leq \omega(\mathscr{T}_f(\mathbf{x}_0)\cap\mathscr{B}^{n-1}) + \frac{\varepsilon_0\sqrt{n}}{R_{min}}$$

■

Lemma A.14 *Let $\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)$ denote the widened cone defined in (A.56) and consider the exact same setup in Lemma A.13. Fix $\varepsilon_1 > 0$. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have i.i.d. standard normal entries. Then, whenever,*

$$\gamma(m,f,\varepsilon_0,\varepsilon_1) := \sqrt{m-1} - \sqrt{\mathbf{D}(cone(\partial f(\mathbf{x}_0)))} - \frac{\varepsilon_0\sqrt{n}}{R_{min}} - \varepsilon_1 > 0$$

we have,

$$\mathbb{P}(\min_{\mathbf{v}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}} \|\mathbf{A}\mathbf{v}\|_2 \geq \varepsilon_1) \geq 1 - \exp(-\frac{1}{2}\gamma(m,f,\varepsilon_0,\varepsilon_1)^2)$$

Proof: Our proof will be based on Proposition 1.3. Pick $\mathscr{C} = \mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}$ in the above proposition. Combined with Lemma A.13 for any $t < (\sqrt{m-1} - \omega(\mathscr{T}_f(\mathbf{x}_0)) - \frac{\varepsilon_0\sqrt{n}}{R_{min}})$, we have,

$$\mathbb{P}(\min_{\mathbf{v}\in\mathscr{T}_f(\mathbf{x}_0,\varepsilon_0)\cap\mathscr{B}^{n-1}} \|\mathbf{A}\mathbf{v}\|_2 \geq \sqrt{m-1} - \omega(\mathscr{T}_f(\mathbf{x}_0)) - \frac{\varepsilon_0\sqrt{n}}{R_{min}} - t) \geq 1 - \exp(-\frac{t^2}{2})$$

Now, choose $t = \gamma(m,f,\varepsilon_0,\varepsilon_1)$ and use the fact that $\omega(\mathscr{T}_f(\mathbf{x}_0)\cap\mathscr{B}^{n-1})^2 \leq \mathbf{D}(\text{cone}(\partial f(\mathbf{x}_0)))$. ■

Appendix B

Further Proofs for Chapter 5

B.1 Properties of Cones

In this appendix, we state some results regarding cones which are used in the proof of general recovery.

Corollary B.1.1 *Let $\mathscr{C}$ be a closed convex cone and $\mathbf{a},\mathbf{b}$ be vectors satisfying $Proj_{\mathscr{C}}(\mathbf{a}-\mathbf{b})=0$. Then*

$$\|\mathbf{b}\|_2 \geq \|Proj_{\mathscr{C}}(\mathbf{a})\|_2.$$

Proof: Using the last statement of Fact 2.2, we have $\|\text{Proj}_{\mathscr{C}}(\mathbf{a})\|_2 = \|\text{Proj}_{\mathscr{C}}(\mathbf{a}) - \text{Proj}_{\mathscr{C}}(\mathbf{a}-\mathbf{b})\|_2 \leq \|\mathbf{b}\|_2$. ■

To proceed, we require the following result which is in similar spirit to Proposition 1.3.

Theorem B.1.1 (Escape through a mesh, [112]) *Let $\mathscr{D}$ be a subset of the unit sphere $\mathscr{S}^{n-1}$. Given m, let $d=\sqrt{n-m}-\frac{1}{4\sqrt{n-m}}$. Provided that $\omega(\mathscr{D})\leq d$ a random $m-$dimensional subspace which is uniformly drawn w.r.t. Haar measure will have no intersection with $\mathscr{D}$ with probability at least*

$$1-3.5\exp(-(d-\omega(\mathscr{D}))^2). \tag{B.1}$$

Theorem B.1.2 *Consider a random Gaussian map $\mathscr{G}:\mathbb{R}^n\to\mathbb{R}^m$ with i.i.d. entires and the corresponding adjoint operator $\mathscr{G}^*$. Let $\mathscr{C}$ be a closed and convex cone and recalling Definition 5.2.1, let*

$$\zeta(\mathscr{C}):=1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})},\ \gamma(\mathscr{C}):=2\sqrt{\frac{1+\sqrt{\bar{\mathbf{D}}(\mathscr{C})}}{1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})}}}.$$

where $\bar{\mathbf{D}}(\mathscr{C})=\frac{\mathbf{D}(\mathscr{C})}{n}$. Then, if $m\leq\frac{7\zeta(\mathscr{C})}{16}n$, with probability at least $1-6\exp(-(\frac{\zeta(\mathscr{C})}{4})^2n)$, for all $\mathbf{z}\in\mathbb{R}^n$ we

have

$$\|\mathscr{G}^*(\mathbf{z})\|_2 \leq \gamma(\mathscr{C})\|Proj_{\mathscr{C}}(\mathscr{G}^*(\mathbf{z}))\|_2. \tag{B.2}$$

Proof: For notational simplicity, let $\zeta = \zeta(\mathscr{C})$ and $\gamma = \gamma(\mathscr{C})$. Consider the set

$$\mathscr{D} = \left\{\mathbf{x} \in \mathscr{S}^{n-1} : \ \|\mathbf{x}\|_2 \geq \gamma\|\text{Proj}_{\mathscr{C}}(\mathbf{x})\|_2\right\}.$$

and we are going to show that with high probability, the range of $\mathscr{G}^*$ misses $\mathscr{D}$. Using Moreau's decomposition (Fact 2.1), for any $\mathbf{x} \in \mathscr{D}$, we may write

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{g}\rangle &= \langle \text{Proj}_{\mathscr{C}}(\mathbf{x}) + \text{Proj}_{\mathscr{C}^\circ}(\mathbf{x}), \text{Proj}_{\mathscr{C}}(\mathbf{g}) + \text{Proj}_{\mathscr{C}^\circ}(\mathbf{g})\rangle \\
&\leq \langle \text{Proj}_{\mathscr{C}}(\mathbf{x}), \text{Proj}_{\mathscr{C}}(\mathbf{g})\rangle + \langle \text{Proj}_{\mathscr{C}^\circ}(\mathbf{x}), \text{Proj}_{\mathscr{C}^\circ}(\mathbf{g})\rangle \qquad \text{(B.3)} \\
&\leq \|\text{Proj}_{\mathscr{C}}(\mathbf{x})\|_2\|\text{Proj}_{\mathscr{C}}(\mathbf{g})\|_2 + \|\text{Proj}_{\mathscr{C}^\circ}(\mathbf{x})\|_2\|\text{Proj}_{\mathscr{C}^\circ}(\mathbf{g})\|_2 \\
&\leq \gamma^{-1}\|\text{Proj}_{\mathscr{C}}(\mathbf{g})\|_2 + \|\text{Proj}_{\mathscr{C}^\circ}(\mathbf{g})\|_2
\end{aligned}$$

where in (B.3) we used the fact that elements of $\mathscr{C}$ and $\mathscr{C}^\circ$ have nonpositive inner products and $\|\text{Proj}_{\mathscr{C}}(\mathbf{x})\|_2 \leq \|\mathbf{x}\|_2$ is by Fact 2.2. Hence, from the definition of Gaussian width,

$$\begin{aligned}
\omega(\mathscr{D}) = \mathbb{E}\left[\sup_{\mathbf{x}\in\mathscr{D}} \langle \mathbf{x}, \mathbf{g}\rangle\right] &\leq \gamma^{-1}\,\mathbb{E}\left[\|\text{Proj}_{\mathscr{C}}(\mathbf{g})\|_2\right] + \mathbb{E}\left[\|\text{Proj}_{\mathscr{C}^\circ}(\mathbf{g})\|_2\right] \\
&\leq \sqrt{n}(\gamma^{-1}\sqrt{\bar{\mathbf{D}}(\mathscr{C}^\circ)} + \sqrt{\bar{\mathbf{D}}(\mathscr{C})}) \leq \frac{2-\zeta}{2}\sqrt{n}.
\end{aligned}$$

Where we used the fact that $\gamma \geq \frac{2\sqrt{\bar{\mathbf{D}}(\mathscr{C}^\circ)}}{1-\sqrt{\bar{\mathbf{D}}(\mathscr{C})}}$; which follows from $\bar{\mathbf{D}}(\mathscr{C}) + \bar{\mathbf{D}}(\mathscr{C}^\circ) = 1$ (recall Fact 2.1). Hence, whenever,

$$m \leq \frac{7\zeta}{16}n \leq (1 - (\frac{4-\zeta}{4})^2)n = m',$$

using the upper bound on $\omega(\mathscr{D})$, we have,

$$(\sqrt{n-m} - \omega(\mathscr{D}) - \frac{1}{4\sqrt{n-m}})^2 \geq (\sqrt{n-m} - \omega(\mathscr{D}))^2 - \frac{1}{2} \geq (\frac{\zeta}{4})^2 n - \frac{1}{2}. \tag{B.4}$$

Now, using Theorem B.1.1, the range space of $\mathscr{G}^*$ will miss the undesired set $\mathscr{D}$ with probability at least $1 - 3.5\exp(-(\frac{\zeta}{4})^2 n + \frac{1}{2}) \geq 1 - 6\exp(-(\frac{\zeta}{4})^2 n)$. ■

Lemma B.1.1 *Consider the cones $\mathbb{S}^d$ and $\mathbb{S}^d_+$ in the space $\mathbb{R}^{d\times d}$. Then, $\bar{\mathbf{D}}(\mathbb{S}^d) < \frac{1}{2}$ and $\bar{\mathbf{D}}(\mathbb{S}^d_+) < \frac{3}{4}$.*

Proof: Let $\mathbf{G}$ be a $d \times d$ matrix with i.i.d. standard normal entries. Set of symmetric matrices $\mathbb{S}^d$ is an $\frac{d(d+1)}{2}$ dimensional subspace of $\mathbb{R}^{d\times d}$. Hence, $\mathbb{E}\,\|\text{Proj}_{\mathbb{S}^d}(\mathbf{G})\|_F^2 = \frac{d(d+1)}{2}$ and $\mathbb{E}\,\|\text{Proj}_{(\mathbb{S}^d)^\circ}(\mathbf{G})\|_F^2 = \frac{d(d-1)}{2}$. Hence,

$$\bar{\mathbf{D}}(\mathbb{S}^d) = \frac{d(d-1)}{2d^2} < \frac{1}{2}.$$

To prove the second statement, observe that projection of a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ onto $\mathbb{S}^d_+$ is obtained by first projecting $\mathbf{A}$ onto $\mathbb{S}^d$ and then taking the matrix induced by the positive eigenvalues of $\text{Proj}_{\mathbb{S}^d}(\mathbf{A})$. Since, $\mathbf{G}$ and $-\mathbf{G}$ are identically distributed and $\mathbb{S}^d_+$ is a self dual cone, $\text{Proj}_{\mathbb{S}^d_+}(\mathbf{G})$ is identically distributed as $-\text{Proj}_{\mathbb{S}^d_-}(\mathbf{G})$ where $\mathbb{S}^d_- = (\mathbb{S}^d_+)^\circ$ stands for negative semidefinite matrices. Hence,

$$\mathbb{E}\,\|\text{Proj}_{\mathbb{S}^d_+}(\mathbf{G})\|_F^2 = \frac{\mathbb{E}\,\|\text{Proj}_{\mathbb{S}^d}(\mathbf{G})\|_F^2}{2} = \frac{d(d+1)}{4},\ \mathbb{E}\,\|\text{Proj}_{(\mathbb{S}^d_+)^\circ}(\mathbf{G})\|_F^2 = \frac{d(3d-1)}{4}.$$

Consequently, $\bar{\mathbf{D}}(\mathbb{S}^d_+) = \frac{3}{4} - \frac{1}{4d} < \frac{3}{4}$.

■

B.2 Norms in Sparse and Low-rank Model

B.2.1 Relevant notation for the proofs

Let $[k]$ denote the set $\{1,2,\ldots,k\}$. Let S_c, S_r denote the indexes of the nonzero columns and rows of $\mathbf{X}_0$ so that nonzero entries of $\mathbf{X}_0$ lies on $S_r \times S_c$ submatrix. $\mathscr{S}_c, \mathscr{S}_r$ denotes the k_1, k_2 dimensional subspaces of vectors whose nonzero entries lie on S_c and S_r respectively.

Let $\mathbf{X}_0$ have singular value decomposition $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ such that $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ and columns of $\mathbf{U}, \mathbf{V}$ lies on $\mathscr{S}_c, \mathscr{S}_r$ respectively.

B.2.2 Proof of Lemma 5.6.1

Proof: Observe that $T_c = \mathbb{R}^d \times \mathscr{S}_c$ and $T_r = \mathscr{S}_r \times \mathbb{R}^d$ hence $T_c \cap T_r$ is the set of matrices that lie on $S_r \times S_c$. Hence, $\mathbf{E}_\star = \mathbf{U}\mathbf{V}^T \in T_c \cap T_r$. Similarly, $\mathbf{E}_c$ and $\mathbf{E}_r$ are the matrices obtained by scaling columns and rows of $\mathbf{X}_0$ to have unit size. As a result, they also lie on $S_r \times S_c$ and $T_c \cap T_r$. $\mathbf{E}_\star \in T_\star$ by definition.

Next, we may write $\mathbf{E}_c = \mathbf{X}_0\mathbf{D}_c$ where $\mathbf{D}_c$ is the scaling nonnegative diagonal matrix. Consequently, $\mathbf{E}_c$ lies on the range space of $\mathbf{X}_0$ and belongs to $T_\star$. This follows from definition of $T_\star$ in Lemma 5.5.4 and the fact that $(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{E}_c = 0$.

In the exact same way, $\mathbf{E}_r = \mathbf{D}_r\mathbf{X}_0$ for some nonnegative diagonal $\mathbf{D}_r$ and lies on the range space of $\mathbf{X}^T$ and hence lies on $T_\star$. Consequently, $\mathbf{E}_\star, \mathbf{E}_c, \mathbf{E}_r$ lies on $T_c \cap T_r \cap T_\star$.

Now, consider

$$\langle \mathbf{E}_c, \mathbf{E}_\star \rangle = \left\langle \mathbf{X}_0\mathbf{D}_c, \mathbf{U}\mathbf{V}^T \right\rangle = \operatorname{tr}\left(\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}_c\right) = \operatorname{tr}\left(\mathbf{V}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}^c\right) \geq 0.$$

since both $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$ and $\mathbf{D}^c$ are positive semidefinite matrices. In the exact same way, we have $\langle \mathbf{E}_c, \mathbf{E}_\star \rangle \geq 0$. Finally,

$$\langle \mathbf{E}_c, \mathbf{E}_r \rangle = \langle \mathbf{X}_0\mathbf{D}_c, \mathbf{D}_r\mathbf{X}_0 \rangle = \operatorname{tr}\left(\mathbf{D}_c\mathbf{X}_0^T\mathbf{D}_r\mathbf{X}_0\right) \geq 0,$$

since both $\mathbf{D}_c$ and $\mathbf{X}_0^T\mathbf{D}_r\mathbf{X}_0$ are PSD matrices. Overall, the pairwise inner products of $\mathbf{E}_r, \mathbf{E}_c, \mathbf{E}_\star$ are nonnegative.

■

B.2.3 Results on the positive semidefinite constraint

Lemma B.2.1 *Assume* $\mathbf{X}, \mathbf{Y} \in \mathbb{S}_+^d$ *have eigenvalue decompositions* $\mathbf{X} = \sum_{i=1}^{rank(\mathbf{X})} \sigma_i \mathbf{u}_i\mathbf{u}_i^T$ *and* $\mathbf{Y} = \sum_{i=1}^{rank(\mathbf{Y})} c_i \mathbf{v}_i\mathbf{v}_i^T$. *Further, assume* $\langle \mathbf{Y}, \mathbf{X} \rangle = 0$. *Then,* $\mathbf{U}^T\mathbf{Y} = 0$ *where* $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_{rank(\mathbf{X})}]$.

Proof: Observe that,

$$\langle \mathbf{Y}, \mathbf{X} \rangle = \sum_{i=1}^{\operatorname{rank}(\mathbf{X})} \sum_{j=1}^{\operatorname{rank}(\mathbf{Y})} \sigma_i c_j |\mathbf{u}_i^T\mathbf{v}_j|^2.$$

Since $\sigma_i, c_j > 0$, right hand side is 0 if and only if $\mathbf{u}_i^T\mathbf{v}_j = 0$ for all i, j. Hence, the result follows. ■

Lemma B.2.2 *Assume* $\mathbf{X}_0 \in \mathbb{S}_+^d$ *so that in Section B.2.1,* $S_c = S_r$, $T_c = T_r$, $k_1 = k_2 = k$ *and* $\mathbf{U} = \mathbf{V}$. *Let* $\mathscr{R} = T_c \cap T_r \cap T_\star \cap \mathbb{S}^d$, $S_\star = T_\star \cap \mathbb{S}^d$, *and,*

$$\mathscr{Y} = \{\mathbf{Y} \big| \mathbf{Y} \in (\mathbb{S}_+^d)^*,\ \langle \mathbf{Y}, \mathbf{X}_0 \rangle = 0\},$$

Then, the following statements hold.

- $S_\star \subseteq span(\mathscr{Y})^\perp$. *Hence,* $\mathscr{R} \subseteq S_\star$ *and is orthogonal to* $\mathscr{Y}$.
- $\mathbf{E}_\star \in \mathscr{R}$, $\frac{\|Proj_{\mathscr{R}}(\mathbf{E}_c)\|_F}{\|\mathbf{E}_c\|_F} = \frac{\|Proj_{\mathscr{R}}(\mathbf{E}_r)\|_F}{\|\mathbf{E}_r\|_F} \geq \frac{1}{\sqrt{2}}$.

Proof: The dual of $\mathbb{S}_+^d$ with respect to $\mathbb{R}^{d\times d}$ is the set sum of $\mathbb{S}_+^d$ and Skew^d where Skew^d is the set of skew-symmetric matrices. Now, assume, $\mathbf{Y} \in \mathscr{Y}$ and $\mathbf{X} \in S_\star$. Then, $\langle \mathbf{Y}, \mathbf{X} \rangle = \left\langle \frac{\mathbf{Z}}{2}, \mathbf{X} \right\rangle$ where $\mathbf{Z} =$

$\mathbf{Y}+\mathbf{Y}^T \in \mathbb{S}^d_+$ and $\langle \mathbf{Z}, \mathbf{X}_0 \rangle = 0$. Since $\mathbf{X}_0$, $\mathbf{Z}$ are both PSD, applying Lemma B.2.1, we have $\mathbf{U}^T\mathbf{Z} = 0$ hence $(\mathbf{I}-\mathbf{U}\mathbf{U}^T)\mathbf{Z}(\mathbf{I}-\mathbf{U}\mathbf{U}^T) = \mathbf{Z}$ which means $\mathbf{Z} \in T_\star^\perp$. Hence, $\langle \mathbf{Z}, \mathbf{X} \rangle = \langle \mathbf{Y}, \mathbf{X} \rangle = 0$ as $\mathbf{X} \in S_\star \subset T_\star$. Hence, $\text{span}(\mathscr{Y}) \subseteq S_\star^\perp$.

For the second statement, let $T_\cap = T_\star \cap T_c \cap T_r$. Recalling Lemma 5.6.1, observe that $\mathbf{E}_\star \in T_\cap$. Since $\mathbf{E}_\star$ is also symmetric, $\mathbf{E}_\star \in \mathscr{R}$. Similarly, $\mathbf{E}_c, \mathbf{E}_r \in T_\cap$, $\langle \mathbf{E}_c, \mathbf{E}_r \rangle \geq 0$ and $\|\text{Proj}_{\mathscr{R}}(\mathbf{E}_c)\| = \|\frac{\mathbf{E}_c+\mathbf{E}_r}{2}\|_F \geq \frac{\|\mathbf{E}_c\|_F}{\sqrt{2}}$. Similar result is true for $\mathbf{E}_r$.

■

B.3 Results on non-convex recovery

Next two lemmas are standard results on sub-gaussian measurement operators.

Lemma B.3.1 (Properties of sub-gaussian mappings) *Assume* $\mathbf{X}$ *is an arbitrary matrix with unit Frobenius norm. A measurement operator* $\mathscr{A}(\cdot)$ *with i.i.d zero-mean isotropic subgaussian rows (see Section 5.3) satisfies the following:*

- $\mathbb{E}[\|\mathscr{A}(\mathbf{X})\|_2^2] = m$.
- *There exists an absolute constant* $c > 0$ *such that, for all* $1 \geq \varepsilon \geq 0$*, we have*

$$\mathbb{P}(|\|\mathscr{A}(\mathbf{X})\|_2^2 - m| \geq \varepsilon m) \leq 2\exp(-c\varepsilon^2 m).$$

Proof: Observe that, when $\|\mathbf{X}\|_F = 1$, entries of $\mathscr{A}(\mathbf{X})$ are zero-mean with unit variance. Hence, the first statement follows directly. For the second statement, we use the fact that square of a sub-gaussian random variable is sub-exponential and view $\|\mathscr{A}(\mathbf{X})\|_2^2$ as a sum of m i.i.d. subexponentials with unit mean. Then, result follows from Corollary 5.17 of [213]. ■

For the consequent lemmas, $\mathscr{S}^{d_1 \times d_2}$ denotes the unit Frobenius norm sphere in $\mathbb{R}^{d_1 \times d_2}$.

Lemma B.3.2 *Let* $\mathscr{D} \in \mathbb{R}^{d_1 \times d_2}$ *be an arbitrary cone and* $\mathscr{A}(\cdot) : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ *be a measurement operator with i.i.d zero-mean and isotropic sub-gaussian rows. Assume that the set* $\bar{\mathscr{D}} = \mathscr{S}^{d_1 \times d_2} \cap \mathscr{D}$ *has* ε*-covering number bounded above by* $\eta(\varepsilon)$*. Then, there exists constants* $c_1, c_2 > 0$ *such that whenever* $m \geq c_1 \log \eta(1/4)$*, with probability* $1 - 2\exp(-c_2 m)$*, we have*

$$\mathscr{D} \cap \mathit{Null}(\mathscr{A}) = \{0\}.$$

Proof: Let $\eta = \eta(\frac{1}{4})$, and $\{\mathbf{X}_i\}_{i=1}^{\eta}$ be a $\frac{1}{4}$-covering of $\bar{\mathscr{D}}$. With probability at least $1-2\eta\exp(-c\varepsilon^2 m)$, for all i, we have

$$(1-\varepsilon)m \leq \|\mathscr{A}(\mathbf{X}_i)\|_2^2 \leq (1+\varepsilon)m.$$

Now, let $\mathbf{X}_{\text{sup}} = \arg\sup_{\mathbf{X}\in\bar{\mathscr{D}}} \|\mathbf{A}(\mathbf{X})\|_2$. Choose $1 \leq a \leq \eta$ such that $\|\mathbf{X}_a - \mathbf{X}_{\text{sup}}\|_2 \leq 1/4$. Then:

$$\|\mathbf{A}(\mathbf{X}_{\text{sup}})\|_2 \leq \|\mathbf{A}(\mathbf{X}_a)\|_2 + \|\mathbf{A}(\mathbf{X}_{\text{sup}} - \mathbf{X}_a)\|_2 \leq (1+\varepsilon)m + \frac{1}{4}\|\mathbf{A}(\mathbf{X}_{\text{sup}})\|_2.$$

Hence, $\|\mathbf{A}(\mathbf{X}_{\text{sup}})\|_2 \leq \frac{4}{3}(1+\varepsilon)m$. Similarly, let $\mathbf{X}_{\text{inf}} = \arg\inf_{\mathbf{X}\in\bar{\mathscr{D}}} \|\mathbf{A}(\mathbf{X})\|_2$. Choose $1 \leq b \leq \eta$ satisfying $\|\mathbf{X}_b - \mathbf{X}_{\text{inf}}\| \leq 1/4$. Then,

$$\|\mathbf{A}(\mathbf{X}_{\text{inf}})\|_2 \geq \|\mathbf{A}(\mathbf{X}_b)\|_2 - \|\mathbf{A}(\mathbf{X}_{\text{inf}} - \mathbf{X}_b)\|_2 \geq (1-\varepsilon)m - \frac{1}{3}(1+\varepsilon)m.$$

This yields $\|\mathbf{A}(\mathbf{X}_{\text{inf}})\|_2 \geq \frac{2-4\varepsilon}{3}m$. Choosing $\varepsilon = 1/4$ whenever $m \geq \frac{32}{c}\log(\eta)$ with the desired probability, $\|\mathbf{A}(\mathbf{X}_{\text{inf}})\|_2 > 0$. Equivalently, $\bar{\mathscr{D}} \cap \text{Null}(\mathbf{A}) = \emptyset$. Since $\mathbf{A}(\cdot)$ is linear and $\mathscr{D}$ is a cone, the claim is proved. ■

The following lemma gives a covering number of the set of low rank matrices.

Lemma B.3.3 (Candes and Plan, [39]) *Let M be the set of matrices in $\mathbb{R}^{d_1\times d_2}$ with rank at most r. Then, for any $\varepsilon > 0$, there exists a covering of $\mathscr{S}^{d_1\times d_2} \cap M$ with size at most $(\frac{c_3}{\varepsilon})^{(d_1+d_2)r}$ where c_3 is an absolute constant. In particular, $\log(\eta(1/4))$ is upper bounded by $C^{(d_1+d_2)r}$ for some constant $C > 0$.*

Now, we use Lemma B.3.3 to find the covering number of the set of simultaneously low rank and sparse matrices.

B.3.1 Proof of Lemma 5.6.2

Proof: Assume M has $\frac{1}{4}$-covering number N. Then, using Lemma B.3.2, whenever $m \geq c_1 \log N$, (5.18) will hold. What remains is to find N. To do this, we cover each individual $s_1 \times s_2$ submatrix and then take the union of the covers. For a fixed submatrix, using Lemma B.3.3, $\frac{1}{4}$-covering number is given by $C^{(s_1+s_2)q}$. In total there are $\binom{d_1}{s_1} \times \binom{d_2}{s_2}$ distinct submatrices. Consequently, by using $\log\binom{d}{s} \approx s\log\frac{d}{s} + s$, we find

$$\log N \leq \log\left(\binom{d_1}{s_1} \times \binom{d_2}{s_2} C^{(s_1+s_2)q}\right) \leq s_1 \log\frac{d_1}{s_1} + s_1 + s_2\log\frac{d_2}{s_2} + s_2 + (s_1+s_2)q\log C,$$

and obtain the desired result. ■

Appendix C

Further Proofs for Chapter 6

C.1 On the success of the simple program

The theorems in Section 6.2.1 provide the conditions under which Program 6.1 succeeds or fails. In this section, we provide the proofs of the success results, i.e., the last statements of Theorems 6.1 and 6.2. The failure results will be the topic of Section C.2.

Notation: Before we proceed, we need some additional notation. $\mathbf{1}^n$ will denote a vector in $\mathbb{R}^n$ with all ones. Complement of a set S will be denoted by S^c. Let $\mathscr{R}_{i,j} = \mathscr{C}_i \times \mathscr{C}_j$ for $1 \leq i, j \leq K+1$. One can see that $\{\mathscr{R}_{i,j}\}$ divides $[n] \times [n]$ into $(K+1)^2$ disjoint regions similar to a grid which is illustrated in Figure C.1. Thus, $\mathscr{R}_{i,i}$ is the region induced by i'th cluster for any $i \leq K$.

Let $\mathscr{A} \subseteq [n] \times [n]$ be the set of nonzero coordinates of $\mathbf{A}$. Then the sets,

1. $\mathscr{A} \cap \mathscr{R}$ corresponds to the edges inside the clusters.
2. $\mathscr{A}^c \cap \mathscr{R}$ corresponds to the missing edges inside the clusters.
3. $\mathscr{A} \cap \mathscr{R}^c$ corresponds to the set of edges outside the clusters, which should be ideally not present.

Let c and d be positive integers. Consider a matrix, $\mathbf{X} \in \mathbb{R}^{c \times d}$. Let β be a subset of $[c] \times [d]$. Then, let $\mathbf{X}_\beta$ denote the matrix induced by the entries of $\mathbf{X}$ on β i.e.,

$$(\mathbf{X}_\beta)_{i,j} = \begin{cases} \mathbf{X}_{i,j} & \text{if } (i,j) \in \beta \\ 0 & \text{otherwise .} \end{cases}$$

In other words, $\mathbf{X}_\beta$ is a matrix whose entries match those of $\mathbf{X}$ in the positions $(i,j) \in \beta$ and zero otherwise. For example, $\mathbb{1}_{\mathscr{A}}^{n \times n} = \mathbf{A}$. Given a matrix $\mathbf{A}$, sum$(\mathbf{A})$ will denote the sum of all entries of $\mathbf{A}$. Finally, we

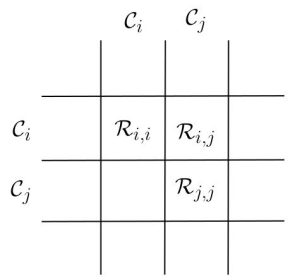


Figure C.1: Illustration of $\{\mathscr{R}_{i,j}\}$ dividing $[n] \times [n]$ into disjoint regions similar to a grid.

introduce the following parameter which will be useful for the subsequent analysis. This parameter can be seen as a measure of distinctness of the "worst" cluster from the "background noise". Here, by background noise we mean the edges over $\mathscr{R}^c$. Given $q, \{p_i\}_{i=1}^K$, let,

$$\begin{aligned} \mathbf{D}_{\mathscr{A}} &= \frac{1}{2}\min\{1-2q, \{2p_i - 1 - \frac{1}{\lambda n_i}{}_{i=1}^{K}\} \\ &= \frac{1}{2}\min\{1-2q, \frac{\mathbf{ED}_i - \lambda^{-1}}{n_i}\} \end{aligned} \tag{C.1}$$

For our proofs, we will make use of the following Big O notation. $f(n) = \Omega(n)$ will mean there exists a positive constant c such that for sufficiently large n, $f(n) \geq cn$. $f(n) = O(n)$ will mean there exists a positive constant c such that for sufficiently large n, $f(n) \leq cn$.

Observe that the success condition of Theorem 6.1 is a special case of that of Theorem 6.2. Considering Theorem 6.1, suppose $\mathbf{ED}_{\min} \geq (1+\varepsilon)\Lambda_{succ}^{-1}$ and $\lambda = (1-\delta)\Lambda_{succ}$ where $\delta > 0$ is to be determined. Choose δ so that $1-\delta = (1+\varepsilon)^{-1/2}$. Now, considering Theorem 6.2, we already have, $\lambda \leq (1-\delta)\Lambda_{succ}$ and we also satisfy the second requirement as we have $\mathbf{ED}_{\min} \geq (1+\varepsilon)\Lambda_{succ}^{-1} = (1+\varepsilon)(1-\delta)\lambda^{-1} = \sqrt{1+\varepsilon}\lambda^{-1}$. Consequently, we will only prove Theorem 6.2 and we will assume that there exists a constant $\varepsilon > 0$ such that,

$$\begin{aligned} &\lambda \leq (1-\varepsilon)\Lambda_{succ} \\ &\mathbf{ED}_{\min} \geq (1+\varepsilon)\lambda^{-1} \end{aligned} \tag{C.2}$$

This implies that $\mathbf{D}_{\mathscr{A}}$ is lower bounded by a positive constant. The reason is $p_{min} > 1/2$ hence $2p_i - 1 > 0$

and we additionally have that $2p_i - 1 \geq (1+\varepsilon)\frac{1}{\lambda n_i}$. Together, these ensure, $2p_i - 1 - \frac{1}{\lambda n_i} \geq \frac{\varepsilon}{1+\varepsilon}(2p_i - 1)$.

C.1.1 Conditions for Success of the Simple Program

In order to show that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal solution to the program (6.1), we need to prove that the objective function strictly increases for any perturbation, i.e.,

$$(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda \, \|\mathbf{S}^0 + \mathbf{E}^S\|_1) - (\|\mathbf{L}^0\|_\star + \lambda \, \|\mathbf{S}^0\|_1) > 0, \tag{C.3}$$

for all feasible perturbations $(\mathbf{E}^L, \mathbf{E}^S)$.

For the following discussion, we will use a slightly abused notation where we denote a subgradient of a norm $\|\cdot\|_*$ at the point $\mathbf{x}$ by $\partial\|\mathbf{x}\|_*$. In the standard notation, $\partial\|\mathbf{x}\|_*$ denotes the set of all subgradients, i.e., the subdifferential.

We can lower bound the LHS of the equation (C.3) using the subgradients as follows,

$$\begin{aligned}(\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda \, \|\mathbf{S}^0 + \mathbf{E}^S\|_1) - \big(\|\mathbf{L}^0\|_\star + \lambda \, \|\mathbf{S}^0\|_1\big) \\ \geq \langle \partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda \langle \partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle,\end{aligned} \tag{C.4}$$

where $\partial\|\mathbf{L}^0\|_\star$ and $\partial\|\mathbf{S}^0\|_1$ are subgradients of nuclear norm and ℓ_1-norm respectively at the points $\big(\mathbf{L}^0, \mathbf{S}^0\big)$.

To make use of (C.4), it is crucial to choose good subgradients. Our efforts will now focus on construction of such subgradients.

C.1.1.1 Subgradient construction

Write $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = \text{diag}\{n_1, n_2, \ldots, n_K\}$ and $\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_K] \in \mathbb{R}^{n\times K}$, with

$$\mathbf{u}_{l,i} = \begin{cases} \frac{1}{\sqrt{n_l}} & \text{if } i \in \mathscr{C}_l \\ 0 & \text{otherwise.} \end{cases}$$

Then the subgradient $\partial\|\mathbf{L}^0\|_\star$ is of the form $\mathbf{U}\mathbf{U}^T + \mathbf{W}$ such that $\mathbf{W} \in \mathscr{M}_U := \{\mathbf{X} : \mathbf{X}\mathbf{U} = \mathbf{U}^T\mathbf{X} = 0, \|\mathbf{X}\| \leq 1\}$. The subgradient $\partial\|\mathbf{S}^0\|_1$ is of the form $\text{sign}(\mathbf{S}^0) + \mathbf{Q}$ where $\mathbf{Q}_{i,j} = 0$ if $\mathbf{S}^0_{i,j} \neq 0$ and $\|\mathbf{Q}\|_\infty \leq 1$. We note that since $\mathbf{L} + \mathbf{S} = \mathbf{A}$, $\mathbf{E}^L = -\mathbf{E}^S$. Note that $\text{sign}(\mathbf{S}^0) = \mathbb{1}^{n\times n}_{\mathscr{A}\cap\mathscr{R}^c} - \mathbb{1}^{n\times n}_{\mathscr{A}^c\cap\mathscr{R}}$. Choosing $\mathbf{Q} = \mathbb{1}^{n\times n}_{\mathscr{A}\cap\mathscr{R}} - \mathbb{1}^{n\times n}_{\mathscr{A}^c\cap\mathscr{R}^c}$, we get,

$$
\begin{aligned}
&\|\mathbf{L}^0+\mathbf{E}^L\|_\star+\lambda\ \|\mathbf{S}^0+\mathbf{E}^S\|_1-(\|\mathbf{L}^0\|_\star+\lambda\ \|\mathbf{S}^0\|_1)\\
&\quad\geq\langle\partial\|\mathbf{L}^0\|_\star,\mathbf{E}^L\rangle+\lambda\langle\partial\|\mathbf{S}^0\|_1,\mathbf{E}^S\rangle\\
&\quad=\langle\mathbf{U}\mathbf{U}^T+\mathbf{W},\mathbf{E}^L\rangle+\lambda\langle\mathrm{sign}(\mathbf{S}^0)+\mathbf{Q},\mathbf{E}^S\rangle\\
&\quad=\underbrace{\sum_{i=1}^{K}\frac{1}{n_i}\mathrm{sum}(\mathbf{E}_{R_{i,i}})+\lambda\left(\mathrm{sum}(\mathbf{E}^L_{\mathscr{A}^c})-\mathrm{sum}(\mathbf{E}^L_{\mathscr{A}})\right)}_{:=g(\mathbf{E}^L)}\\
&\qquad\qquad+\left\langle\mathbf{W},\mathbf{E}^L\right\rangle. \qquad\text{(C.5)}
\end{aligned}
$$

Define,

$$
g(\mathbf{E}^L):=\sum_{i=1}^{K}\frac{1}{n_i}\mathrm{sum}(\mathbf{E}^L_{\mathscr{R}_{i,i}})+\lambda\left(\mathrm{sum}(\mathbf{E}^L_{\mathscr{A}^c})-\mathrm{sum}(\mathbf{E}^L_{\mathscr{A}})\right). \qquad\text{(C.6)}
$$

Also, define $f\left(\mathbf{E}^L,\mathbf{W}\right):=g\left(\mathbf{E}^L\right)+\left\langle\mathbf{W},\mathbf{E}^L\right\rangle$. Our aim is to show that for all feasible perturbations $\mathbf{E}^L$, there exists $\mathbf{W}$ such that,

$$
f\left(\mathbf{E}^L,\mathbf{W}\right)=g(\mathbf{E}^L)+\left\langle\mathbf{W},\mathbf{E}^L\right\rangle>0. \qquad\text{(C.7)}
$$

Note that $g(\mathbf{E}^L)$ does not depend on $\mathbf{W}$.

Lemma C.1 *Given $\mathbf{E}^L$, assume there exists $\mathbf{W}\in\mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\|<1$ such that $f(\mathbf{E}^L,\mathbf{W})\geq 0$. Then at least one of the followings holds:*

- *There exists $\mathbf{W}^*\in\mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}^*\|\leq 1$ and $f(\mathbf{E}^L,\mathbf{W}^*)>0$.*
- *For all $\mathbf{W}\in\mathscr{M}_{\mathbf{U}}$, $\left\langle\mathbf{E}^L,\mathbf{W}\right\rangle=0$.*

Proof: Let $c=1-\|\mathbf{W}\|$. Assume $\left\langle\mathbf{E}^L,\mathbf{W}'\right\rangle\neq 0$ for some $\mathbf{W}'\in\mathscr{M}_{\mathbf{U}}$. If $\left\langle\mathbf{E}^L,\mathbf{W}'\right\rangle>0$, choose $\mathbf{W}^*=\mathbf{W}+c\mathbf{W}'$. Otherwise, choose $\mathbf{W}^*=\mathbf{W}-c\mathbf{W}'$. Since $\|\mathbf{W}'\|\leq 1$, we have, $\|\mathbf{W}^*\|\leq 1$ and $\mathbf{W}^*\in\mathscr{M}_{\mathbf{U}}$. Consequently,

$$
\begin{aligned}
f(\mathbf{E}^L,\mathbf{W}^*)&=f(\mathbf{E}^L,\mathbf{W})+|\left\langle\mathbf{E}^L,c\mathbf{W}'\right\rangle|\\
&>f(\mathbf{E}^L,\mathbf{W})\geq 0 \qquad\text{(C.8)}
\end{aligned}
$$

■

Notice that, for all $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$, $\langle \mathbf{E}^L, \mathbf{W} \rangle = 0$ is equivalent to $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$ which is the orthogonal complement of $\mathscr{M}_{\mathbf{U}}$ in $\mathbb{R}^{n\times n}$. $\mathscr{M}_{\mathbf{U}}^{\perp}$ has the following characterization:

$$\mathscr{M}_{\mathbf{U}}^{\perp} = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X} = \mathbf{U}\mathbf{M}^T + \mathbf{N}\mathbf{U}^T \\ \text{for some } \mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times K}\}. \tag{C.9}$$

Now we have broken down our aim into two steps.

1. Construct $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$, such that $f(\mathbf{E}^L, \mathbf{W}) \geq 0$ for all feasible perturbations $\mathbf{E}^L$.
2. For all non-zero feasible $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$, show that $g(\mathbf{E}^L) > 0$.

As a first step, in Section C.1.2, we will argue that, under certain conditions, there exists a $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$ such that with high probability, $f(\mathbf{E}^L, \mathbf{W}) \geq 0$ for all feasible $\mathbf{E}^L$. This $\mathbf{W}$ is called the dual certificate. Secondly, in Section C.1.3, we will show that, under certain conditions, for all $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$ with high probability, $g(\mathbf{E}^L) > 0$. Finally, combining these two arguments, and using Lemma C.1 we will conclude that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal with high probability.

C.1.2 Showing existence of the dual certificate

Recall that

$$\begin{aligned} f(\mathbf{E}^L, \mathbf{W}) &= \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{E}^L_{\mathscr{R}_{i,i}}) + \langle \mathbf{E}^L, \mathbf{W} \rangle \\ &\quad + \lambda \left(\text{sum}\left(\mathbf{E}^L_{\mathscr{A}^c}\right) - \text{sum}\left(\mathbf{E}^L_{\mathscr{A}}\right)\right) \end{aligned}$$

$\mathbf{W}$ will be constructed from the candidate $\mathbf{W}_0$, which is given as follows.

C.1.2.1 Candidate $\mathbf{W}_0$

Based on Program 6.1, we propose the following,

$$\mathbf{W}_0 = \sum_{i=1}^{K} c_i \mathbb{1}^{n\times n}_{\mathscr{R}_{i,i}} + c\,\mathbb{1}^{n\times n}_{\mathscr{R}^c} + \lambda \left(\mathbb{1}^{n\times n}_{\mathscr{A}} - \mathbb{1}^{n\times n}_{\mathscr{A}^c} \right),$$

where $\{c_i\}_{i=1}^K, c$ are real numbers to be determined.

We now have to find a bound on the spectral norm of $\mathbf{W}_0$. Note that $\mathbf{W}_0$ is a random matrix where randomness is due to $\mathscr{A}$. In order to ensure a small spectral norm, we will set its expectation to 0, i.e., we will choose $c, \{c_i\}'s$ to ensure that $\mathbb{E}[\mathbf{W}_0] = 0$.

Following from the Stochastic Block Model 6.1, the expectation of an entry of $\mathbf{W}_0$ on $\mathscr{R}_{i,i}$ (region corresponding to cluster i) and $\mathscr{R}^c$ (region outside the clusters) is $c_i + \lambda(2p_i - 1)$ and $c + \lambda(2q-1)$ respectively. Hence, we set,

$$c_i = -\lambda(2p_i - 1) \quad \text{and} \quad c = -\lambda(2q-1),$$

With these choices, the candidate $\mathbf{W}_0$ and $f(\mathbf{E}^L, \mathbf{W}_0)$ take the following forms,

$$\begin{aligned} \mathbf{W}_0 &= 2\lambda \left[\sum_{i=1}^{K} (1-p_i)\ \mathbb{1}^{n\times n}_{\mathscr{R}_{i,i}\cap\mathscr{A}} - p_i\ \mathbb{1}^{n\times n}_{\mathscr{R}_{i,i}\cap\mathscr{A}^c} \right] \\ &\quad +2\lambda \left[(1-q)\ \mathbb{1}^{n\times n}_{\mathscr{R}^c\cap\mathscr{A}} - q\ \mathbb{1}^{n\times n}_{\mathscr{R}^c\cap\mathscr{A}^c} \right] \end{aligned} \tag{C.10}$$

$$\begin{aligned} f(\mathbf{E}^L, \mathbf{W}_0) &= \lambda \left[(1-2q)\ \mathrm{sum}(\mathbf{E}^L_{\mathscr{R}^c}) \right] \\ &\quad -\lambda \left[\sum_{i=1}^{K} \left(2p_i - 1 - \frac{1}{\lambda n_i} \right)\ \mathrm{sum}(\mathbf{E}^L_{\mathscr{R}_{i,i}}) \right] \end{aligned} \tag{C.11}$$

From $\mathbf{L}^0$ and (6.2), it follows that,

$$\begin{aligned} &\mathbf{E}^L_{\mathscr{R}^c} \text{ is (entrywise) nonnegative.} \\ &\mathbf{E}^L_{\mathscr{R}} \text{ is (entrywise) nonpositive.} \end{aligned} \tag{C.12}$$

Thus, $\mathrm{sum}(\mathbf{E}^L_{\mathscr{R}^c}) \le 0$ and $\mathrm{sum}(\mathbf{E}^L_{\mathscr{R}_{i,i}}) \ge 0$. When $\lambda(2p_i - 1) - \frac{1}{n_i} \ge 0$ and $\lambda(2q-1) \le 0$; we will have $f(\mathbf{E}^L, \mathbf{W}_0) \ge 0$ for all feasible $\mathbf{E}^L$. This indeed holds due to the assumptions of Theorem 6.1 (see (C.1)), as we assumed $2p_i - 1 > \frac{1}{\lambda n_i}$ for $i = 1, 2\cdots, K$ and $1 > 2q$.

We will now proceed to find a tight bound on the spectral norm of $\mathbf{W}^0$. Let us define the zero-mean

Bernoulli distribution $\text{Bern}_0(\alpha)$ as follows. $X \sim \text{Bern}_0(\alpha)$ if,

$$X = \begin{cases} 1-\alpha & w.p. \quad \alpha \\ -\alpha & w.p. \quad 1-\alpha \end{cases}.$$

Theorem C.1 *Assume* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *obeys the stochastic block model* (6.1) *and let* $\mathbf{M} \in \mathbb{R}^{n\times n}$. *Let entries of* $\mathbf{M}$ *be as follows.*

$$\mathbf{M}_{i,j} \sim \begin{cases} Bern_0(p_k) & if \quad (i,j) \in \mathscr{R}_{k,k} \\ Bern_0(q) & if \quad (i,j) \in \mathscr{R}^c. \end{cases}$$

Then, for a constant ε' *(to be determined) each of the following holds with probability* $1-\exp(-\Omega(n))$.

- $\|\mathbf{M}\| \leq (1+\varepsilon')\sqrt{n}$.
- $\|\mathbf{M}\| \leq 2\sqrt{q(1-q)}\sqrt{n}$
  $+\max_{i\leq K} 2\sqrt{q(1-q)+p_i(1-p_i)}\sqrt{n_i}+\varepsilon'\sqrt{n}$.
- *Assume* $\max_{1\leq i\leq K} n_i = o(n)$. *Then, for sufficiently large n,*

$$\|\mathbf{M}\| \leq (2\sqrt{q(1-q)}+\varepsilon')\sqrt{n}.$$

Proof: The entries of $\mathbf{M}$ are i.i.d. with maximum variance of $1/4$. Hence, the first statement follows directly from [216]. For the second statement, let,

$$\mathbf{M}_1(i,j) = \begin{cases} \mathbf{M}(i,j) & \text{if } i,j \in \mathbb{R}^c \\ \text{Bern}_0(q) \text{ else} \end{cases}.$$

Also let $\mathbf{M}_2 = \mathbf{M} - \mathbf{M}_1$. Observe that, $\mathbf{M}_1$ has i.i.d. $\text{Bern}_0(q)$ entries. From standard results on random matrix theory, it follows that, with the desired probability

$$\|\mathbf{M}_1\| \leq (2\sqrt{q(1-q)}+\varepsilon')\sqrt{n}.$$

For $\mathbf{M}_2$, first observe that over $\mathscr{R}_{i,i}$ $\mathbf{M}_2$ has i.i.d. entries with variance $q(1-q)+p_i(1-p_i)$. This similarly gives,

$$\|\mathbf{M}_{2,\mathscr{R}_{i,i}}\| \leq 2\sqrt{q(1-q)+p_i(1-p_i)}\sqrt{n_i}+\varepsilon'\sqrt{n}.$$

Now, observing, $\|\mathbf{M}_2\| = \sup_{i\le K} \|\mathbf{M}_{2,\mathscr{R}_{i,i}}\|$ and using a union bound over $i \le K$ we have,

$$\|\mathbf{M}_2\| \le \max_{i\le K} 2\sqrt{q(1-q)+p_i(1-p_i)}\sqrt{n_i} + \varepsilon'\sqrt{n}.$$

Finally, we use the triangle inequality $\|\mathbf{M}\| \le \|\mathbf{M}_1\| + \|\mathbf{M}_2\|$ to conclude. ■

The following lemma gives a bound on $\|\mathbf{W}_0\|$.

Lemma C.2 *Recall that, $\mathbf{W}_0$ is a random matrix; where randomness is on the stochastic block model $\mathscr{A}$ and it is given by,*

$$\begin{aligned}\mathbf{W}_0 = 2\lambda \sum_{i=1}^{K} \left[(1-p_i)\mathbb{1}^{n\times n}_{\mathscr{A}\cap\mathscr{R}_{i,i}} - p_i \mathbb{1}^{n\times n}_{\mathscr{A}^c\cap\mathscr{R}_{i,i}}\right] \\ + 2\lambda\left[(1-q)\mathbb{1}^{n\times n}_{\mathscr{A}\cap\mathscr{R}^c} - q\mathbb{1}^{n\times n}_{\mathscr{A}^c\cap\mathscr{R}^c}\right]\end{aligned} \tag{C.13}$$

Then, for any $\varepsilon' > 0$, with probability $1-\exp(-\Omega(n))$, we have

$$\begin{aligned}\|\mathbf{W}_0\| &\le 4\lambda\sqrt{q(1-q)}\sqrt{n} \\ &\quad + \max_{i\le K} 4\lambda\sqrt{q(1-q)+p_i(1-p_i)}\sqrt{n_i} + \varepsilon'\lambda\sqrt{n} \\ &\le \lambda\Lambda_{succ}^{-1} + \varepsilon'\lambda\sqrt{n}\end{aligned}$$

Further, if $\max_{1\le i\le K} n_i = o(n)$. Then, for sufficiently large n, with the same probability,

$$\|\mathbf{W}_0\| \le 4\lambda\sqrt{q(1-q)n} + \varepsilon'\lambda\sqrt{n}.$$

Proof: $\frac{1}{2\lambda}\mathbf{W}_0$ is a random matrix whose entries are i.i.d. and distributed as $\text{Bern}_0(p_i)$ on $\mathscr{R}_{i,i}$ and $\text{Bern}_0(q)$ on $\mathscr{R}^c$. Consequently, using Theorem C.1 and recalling the definition of Λ_{succ} we obtain the result. ■

Lemma C.2 verifies that asymptotically with high probability we can make $\|\mathbf{W}_0\| < 1$ as long as λ is sufficiently small. However, $\mathbf{W}_0$ itself is not sufficient for construction of the desired $\mathbf{W}$, since we do not have any guarantee that $\mathbf{W}_0 \in \mathscr{M}_{\mathbf{U}}$. In order to achieve this, we will *correct* $\mathbf{W}_0$ by projecting it onto $\mathscr{M}_{\mathbf{U}}$. Following lemma suggests that $\mathbf{W}_0$ does not change much by such a correction.

C.1.2.2 Correcting the candidate $\mathbf{W}_0$

Lemma C.3 *$\mathbf{W}_0$ is as described previously in (C.13). Let $\mathbf{W}^H$ be the projection of $\mathbf{W}_0$ on $\mathscr{M}_{\mathbf{U}}$. Then*

- $\|\mathbf{W}^H\| \leq \|\mathbf{W}_0\|$
- *For any $\varepsilon'' > 0$ (constant to be determined), with probability $1 - 6n^2 \exp(-2\varepsilon''^2 n_{min})$ we have*

$$\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty \leq 3\lambda\varepsilon''.$$

Proof: Choose arbitrary vectors $\{\mathbf{u}_i\}_{i=K+1}^n$ to make $\{\mathbf{u}_i\}_{i=1}^n$ an orthonormal basis in $\mathbb{R}^n$. Call $\mathbf{U}_2 = [\mathbf{u}_{K+1} \ \dots \ \mathbf{u}_n]$ and $\mathbf{P} = \mathbf{U}\mathbf{U}^T$, $\mathbf{P}_2 = \mathbf{U}_2\mathbf{U}_2^T$. Now notice that for any matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, $\mathbf{P}_2\mathbf{X}\mathbf{P}_2$ is in $\mathscr{M}_{\mathbf{U}}$ since $\mathbf{U}^T\mathbf{U}_2 = 0$. Let $\mathbf{I}$ denote the identity matrix. Then,

$$\begin{aligned}\mathbf{X} - \mathbf{P}_2\mathbf{X}\mathbf{P}_2 &= \mathbf{X} - (\mathbf{I}-\mathbf{P})\mathbf{X}(\mathbf{I}-\mathbf{P}) \\ &= \mathbf{P}\mathbf{X} + \mathbf{X}\mathbf{P} - \mathbf{P}\mathbf{X}\mathbf{P} \in \mathscr{M}_{\mathbf{U}}^{\perp}\end{aligned} \tag{C.14}$$

Hence, $\mathbf{P}_2\mathbf{X}\mathbf{P}_2$ is the orthogonal projection on $\mathscr{M}_{\mathbf{U}}$. Clearly,

$$\|\mathbf{W}^H\| = \|\mathbf{P}_2\mathbf{W}_0\mathbf{P}_2\| \leq \|\mathbf{P}_2\|^2\|\mathbf{W}_0\| \leq \|\mathbf{W}_0\|$$

For analysis of $\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty$ we can consider terms on the right hand side of (C.14) separately as we have:

$$\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty \leq \|\mathbf{P}\mathbf{W}_0\|_\infty + \|\mathbf{W}_0\mathbf{P}\|_\infty + \|\mathbf{P}\mathbf{W}_0\mathbf{P}\|_\infty.$$

Clearly $\mathbf{P} = \sum_{i=1}^K \frac{1}{n_i}\mathbb{1}_{\mathbb{R}_{i,i}}^{n\times n}$. Then, each entry of $\frac{1}{\lambda}\mathbf{P}\mathbf{W}_0$ is either a summation of n_i i.i.d. $\text{Bern}_0(p_i)$ or $\text{Bern}_0(q)$ random variables scaled by n_i^{-1} for some $i \leq K$ or 0. Hence any $c,d \in [n]$ and $\varepsilon'' > 0$

$$\mathbb{P}[|(\mathbf{P}\mathbf{W}_0)_{c,d}| \geq \lambda\varepsilon''] \leq 2\exp(-2\varepsilon''^2 n_{min}).$$

Same (or better) bounds holds for entries of $\mathbf{W}_0\mathbf{P}$ and $\mathbf{P}\mathbf{W}_0\mathbf{P}$. Then a union bound over all entries of the three matrices will give with probability $1 - 6n^2\exp(-2\varepsilon''^2 n_{min})$, we have $\|\mathbf{W}_0 - \mathbf{W}^H\|_\infty \leq 3\lambda\varepsilon''$. ■

Recall that $\gamma_{\text{succ}} := \max_{1\leq i\leq K} 4\sqrt{(q(1-q)+p_i(1-p_i))n_i}$, and $\Lambda_{\text{succ}} := \frac{1}{4\sqrt{q(1-q)n}+\gamma_{\text{succ}}}$.

We can summarize our discussion so far in the following lemma,

Lemma C.4 $\mathbf{W}_0$ *is as described previously in* (C.10). *Choose* $\mathbf{W}$ *to be projection of* $\mathbf{W}_0$ *on* $\mathscr{M}_{\mathbf{U}}$. *Also suppose* $\lambda \leq (1-\delta)\Lambda_{succ}$. *Then, with probability* $1-6n^2\exp(-\Omega(n_{min}))-4\exp(-\Omega(n))$ *we have,*

- $\|\mathbf{W}\| < 1$
- *For all feasible* $\mathbf{E}^L$, $f(\mathbf{E}^L, \mathbf{W}) \geq 0$.

Proof: To begin with, observe that Λ_{succ}^{-1} is $\Omega(\sqrt{n})$. Since $\lambda \leq \Lambda_{succ}$, $\lambda\sqrt{n} = \mathscr{O}(1)$. Consequently, using $\lambda\Lambda_{succ}^{-1} < 1$ and applying Lemma C.2, and choosing a sufficiently small $\varepsilon' > 0$, we conclude with,

$$\|\mathbf{W}\| \leq \|\mathbf{W}_0\| < 1,$$

with probability $1-\exp(-\Omega(n))$ where the constant in the exponent depends on the constant $\varepsilon' > 0$.

Next, from Lemma C.3 with probability $1-6n^2\exp(-\frac{2}{9}\varepsilon''^2 n_{min})$ we have $\|\mathbf{W}_0 - \mathbf{W}\|_\infty \leq \lambda\varepsilon''$. Then based on (C.11) for all $\mathbf{E}^L$, we have that,

$$\begin{aligned}
f(\mathbf{E}^L, \mathbf{W}) &= f(\mathbf{E}^L, \mathbf{W}_0) - \langle \mathbf{W}_0 - \mathbf{W}, \mathbf{E}^L \rangle \\
&\geq f(\mathbf{E}^L, \mathbf{W}_0) - \lambda\varepsilon'' \left(\text{sum}(\mathbf{E}^L_{\mathscr{R}}) - \text{sum}(\mathbf{E}^L_{\mathscr{R}^c})\right) \\
&= \lambda\left[(1-2q-\varepsilon'')\text{sum}(\mathbf{E}^L_{\mathscr{R}^c})\right] \\
&\quad -\lambda\sum_{i=1}^{K}\left[(2p_i - 1 - \frac{1}{\lambda n_i} - \varepsilon'')\text{sum}(\mathbf{E}^L_{\mathscr{R}_{i,i}})\right] \\
&\geq 0
\end{aligned}$$

where we chose ε'' to be a sufficiently small constant. In particular, we set $\varepsilon'' < \mathbf{D}_{\mathscr{A}}$, i.e., set $\varepsilon'' < 1-2q$ and $\varepsilon'' < 2p_i - 1 - \frac{1}{\lambda n_i}$ for all $i \leq K$.

Hence, by using a union bound $\mathbf{W}$ satisfies both of the desired conditions. ■

Summary so far: Combining the last lemma with Lemma C.1, with high probability, either there exists a dual vector $\mathbf{W}^*$ which ensures $f(\mathbf{E}^L, \mathbf{W}^*) > 0$ or $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$. If former, we are done. Hence, we need to focus on the latter case and show that for all perturbations $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$, the objective will strictly increase at $(\mathbf{L}^0, \mathbf{S}^0)$ with high probability.

C.1.3 Solving for $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$ case

Recall that,

$$g\left(\mathbf{E}^L\right) = \sum_{i=1}^{K}\frac{1}{n_i}\text{sum}(\mathbf{E}_{R_{i,i}}) + \lambda\left(\text{sum}(\mathbf{E}^L_{\mathscr{A}^c}) - \text{sum}(\mathbf{E}^L_{\mathscr{A}})\right)$$

Let us define,

$$g_1(\mathbf{X}) := \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{X}_{\mathscr{R}_{i,i}}),$$

$$g_2(\mathbf{X}) := \text{sum}(\mathbf{X}_{\mathscr{A}^c}) - \text{sum}(\mathbf{X}_{\mathscr{A}}),$$

so that, $g(\mathbf{X}) = g_1(\mathbf{X}) + \lambda g_2(\mathbf{X})$. Also let $\mathbf{V} = [\mathbf{v}_1 \ \ldots \ \mathbf{v}_K]$ where $\mathbf{v}_i = \sqrt{n_i}\mathbf{u}_i$. Thus, $\mathbf{V}$ is basically obtained by, normalizing columns of $\mathbf{U}$ to make its nonzero entries 1. Assume $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$. Then, by definition of $\mathscr{M}_{\mathbf{U}}^{\perp}$, we can write,

$$\mathbf{E}^L = \mathbf{V}\mathbf{M}^T + \mathbf{N}\mathbf{V}^T.$$

Let $\mathbf{m}_i, \mathbf{n}_i$ denote i'th columns of $\mathbf{M}, \mathbf{N}$ respectively. From $\mathbf{L}^0$ and (6.2) it follows that

$$\mathbf{E}^L_{\mathscr{R}^c} \text{ is (entrywise) nonnegative}$$

$$\mathbf{E}^L_{\mathscr{R}} \text{ is (entrywise) nonpositive}$$

Now, we list some simple observations regarding structure of $\mathbf{E}^L$. We can write

$$\mathbf{E}^L = \sum_{i=1}^{K} (\mathbf{v}_i \mathbf{m}_i^T + \mathbf{n}_i \mathbf{v}_i^T) = \sum_{i=1}^{K+1} \sum_{j=1}^{K+1} \mathbf{E}^L_{\mathscr{R}_{i,j}}. \tag{C.15}$$

Notice that only two components : $\mathbf{v}_i \mathbf{m}_i^T$ and $\mathbf{n}_j \mathbf{v}_j^T$, contribute to the term $\mathbf{E}^L_{\mathscr{R}_{i,j}}$.

Let $\{a_{i,j}\}_{j=1}^{n_i}$ be an (arbitrary) indexing of elements of $\mathscr{C}_i$ i.e. $\mathscr{C}_i = \{a_{i,1}, \ldots, a_{i,n_i}\}$. For a vector $\mathbf{z} \in \mathbb{R}^n$, let $\mathbf{z}^i \in \mathbb{R}^{n_i}$ denote the vector induced by entries of $\mathbf{z}$ in $\mathscr{C}_i$. Basically, for any $1 \le j \le n_i$, $\mathbf{z}^i_j = \mathbf{z}_{a_{i,j}}$. Also, let $\mathbf{E}^{i,j} \in \mathbb{R}^{n_i \times n_j}$ which is $\mathbf{E}^L$ induced by entries on $\mathscr{R}_{i,j}$.

In other words,

$$\mathbf{E}^{i,j}_{c,d} = \mathbf{E}^L_{a_{i,c}, a_{j,d}} \quad \text{for all } (i,j) \in \mathscr{C}_i \times \mathscr{C}_j \text{ and}$$
$$\text{all } 1 \le c \le n_i, \ 1 \le d \le n_j$$

Basically, $\mathbf{E}^{i,j}$ is same as $\mathbf{E}^L_{\mathscr{R}_{i,j}}$ when we get rid of trivial zero rows and zero columns. Then

$$\mathbf{E}^{i,j} = \mathbf{1}^{n_i} \mathbf{m}_i^{j\,T} + \mathbf{n}_j^i \mathbf{1}^{n_j\,T}. \tag{C.16}$$

Clearly, given $\{\mathbf{E}^{i,j}\}_{1 \le i,j \le n}$, $\mathbf{E}^L$ is uniquely determined. Now, assume we fix $\text{sum}(\mathbf{E}^{i,j})$ for all i, j and

we would like to find the *worst* $\mathbf{E}^L$ subject to these constraints. Variables in such an optimization are $\mathbf{m}_i, \mathbf{n}_i$. Basically we are interested in,

$$\min g(\mathbf{E}^L) \tag{C.17}$$

subject to

$$\text{sum}(\mathbf{E}^{i,j}) = c_{i,j} \text{ for all } i,j$$

$$\mathbf{E}^{i,j} \begin{cases} \text{nonnegative if } i \neq j \\ \text{nonpositive if } i = j \end{cases} \tag{C.18}$$

where $\{c_{i,j}\}$ are constants. Constraint (C.18) follows from (C.12).

Remark: For the special case of $i = j = K+1$, notice that $\mathbf{E}^{i,j} = 0$.

In (C.17), $g_1(\mathbf{E}^L)$ is fixed and is equal to $\sum_{i=1}^{K} \frac{1}{n_i} c_{i,i}$. Consequently, we just need to do the optimization with the objective $g_2(\mathbf{E}^L) = \text{sum}(\mathbf{E}^L_{\mathscr{A}^c}) - \text{sum}(\mathbf{E}^L_{\mathscr{A}})$.

Let $\beta_{i,j} \subseteq [n_i] \times [n_j]$ be a set of coordinates defined as follows. For any $(c,d) \in [n_i] \times [n_j]$

$$(c,d) \in \beta_{i,j} \text{ iff } (a_{i,c}, a_{j,d}) \in \mathscr{A}.$$

For $(i_1, j_1) \neq (i_2, j_2)$, $(\mathbf{m}_{i_1}^{j_1}, \mathbf{n}_{j_1}^{i_1})$ and $(\mathbf{m}_{i_2}^{j_2}, \mathbf{n}_{j_2}^{i_2})$ are independent variables. Consequently, due to (C.16), we can partition problem (C.17) into the following smaller disjoint problems.

$$\min_{\mathbf{m}_i^j, \mathbf{n}_j^i} \text{sum}(\mathbf{E}^{i,j}_{\beta^c_{i,j}}) - \text{sum}(\mathbf{E}^{i,j}_{\beta_{i,j}}) \tag{C.19}$$

subject to

$$\text{sum}(\mathbf{E}^{i,j}) = c_{i,j}$$

$$\mathbf{E}^{i,j} \text{ is } \begin{cases} \text{nonnegative if } i \neq j \\ \text{nonpositive if } i = j \end{cases}$$

Then, we can solve these problems locally (for each i,j) to finally obtain,

$$g_2(\mathbf{E}^{L,*}) = \sum_{i,j} \text{sum}(\mathbf{E}^{i,j,*}_{\beta^c_{i,j}}) - \sum_{i,j} \text{sum}(\mathbf{E}^{i,j,*}_{\beta_{i,j}}),$$

to find the overall result of problem (C.17), where $*$ denotes the optimal solutions in problems (C.17) and (C.19). The following lemma will be useful for analysis of these local optimizations.

Lemma C.5 *Let* $\mathbf{a} \in \mathbb{R}^c$, $\mathbf{b} \in \mathbb{R}^d$ *and* $X = \mathbf{1}^c \mathbf{b}^T + \mathbf{a}\mathbf{1}^{dT}$ *be variables and* $C_0 \geq 0$ *be a constant. Also let* $\beta \subseteq [c] \times [d]$. *Consider the following optimization problem*

$$\begin{aligned} &\min_{\mathbf{a},\mathbf{b}} \ sum(\mathbf{X}_{\beta^c}) - sum(\mathbf{X}_{\beta}) \\ &subject\ to \\ &\mathbf{X}_{i,j} \geq 0 \ \ for\ all\ i,j \\ &sum(\mathbf{X}) = C_0 \end{aligned}$$

For this problem there exists a (entrywise) nonnegative minimizer $(\mathbf{a}^0, \mathbf{b}^0)$.

Proof: Let x_i denotes i'th entry of vector $\mathbf{x}$. Assume $(\mathbf{a}^*, \mathbf{b}^*)$ is a minimizer. Without loss of generality assume $b_1^* = \min_{i,j}\{\mathbf{a}_i^*, \mathbf{b}_j^*\}$. If $b_1^* \geq 0$ we are done. Otherwise, since $\mathbf{X}_{i,j} \geq 0$ we have $a_i^* \geq -b_1^*$ for all $i \leq c$. Then set $\mathbf{a}^0 = \mathbf{a}^* + \mathbf{1}^c b_1^*$ and $\mathbf{b}^0 = \mathbf{b}^* - \mathbf{1}^d b_1^*$. Clearly, $(\mathbf{a}^0, \mathbf{b}^0)$ is nonnegative. On the other hand, we have:

$$\mathbf{X}^* = \mathbf{1}^c \mathbf{b}^{*T} + \mathbf{a}^* \mathbf{1}^{dT} = \mathbf{1}^c \mathbf{b}^{0T} + \mathbf{a}^0 \mathbf{1}^{dT} = \mathbf{X}^0,$$

which implies,

$$\begin{aligned} \text{sum}(\mathbf{X}_\beta^*) - \text{sum}(\mathbf{X}_{\beta^c}^*) &= \text{sum}(\mathbf{X}_\beta^0) - \text{sum}(\mathbf{X}_{\beta^c}^0) \\ &= \text{optimal value} \end{aligned}$$

■

Lemma C.6 *A direct consequence of Lemma C.5 is the fact that in the local optimizations (C.19), Without loss of generality, we can assume* $(\mathbf{m}_i^j, \mathbf{n}_j^i)$ *entrywise nonnegative whenever* $i \neq j$ *and entrywise nonpositive when* $i = j$. *This follows from the structure of* $\mathbf{E}^{i,j}$ *given in (C.16) and (C.12).*

The following lemma will help us characterize the relationship between $\text{sum}(\mathbf{E}^{i,j})$ and $\text{sum}(\mathbf{E}^{i,j}_{\beta^c_{i,j}})$.

Lemma C.7 *Let* β *be a random set generated by choosing elements of* $[c] \times [d]$ *indecently with probability* $0 \leq r \leq 1$. *Then for any* $\varepsilon' > 0$ *with probability* $1 - d\exp(-2\varepsilon'^2 c)$ *for all nonzero and entrywise nonnegative*

$\mathbf{a} \in \mathbb{R}^d$ *we'll have:*

$$sum(\mathbf{X}_\beta) > (r-\varepsilon')sum(\mathbf{X}) \tag{C.20}$$

where $\mathbf{X} = \mathbf{1}^c\mathbf{a}^T$. *Similarly, with the same probability, for all such* $\mathbf{a}$, *we'll have* $sum(\mathbf{X}_\beta) < (r+\varepsilon')sum(\mathbf{X})$

Proof: We'll only prove the first statement (C.20) as the proofs are identical. For each $i \leq d$, a_i occurs exactly c times in $\mathbf{X}$ as i'th column of X is $\mathbf{1}^c a_i$. By using a Chernoff bound, we can estimate the number of coordinates of i'th column which are element of β (call this number C_i) as we can view this number as a sum of c i.i.d. Bernoulli(r) random variables. Then

$$\mathbb{P}(C_i \leq c(r-\varepsilon')) \leq \exp(-2\varepsilon'^2 c).$$

Now, we can use a union bound over all columns to make sure for all i, $C_i > c(r-\varepsilon')$

$$\mathbb{P}(C_i > c(r-\varepsilon') \text{ for all } i \leq d) \geq 1 - d\exp(-2\varepsilon'^2 c).$$

On the other hand if each $C_i > c(r-\varepsilon')$ then for any nonnegative $\mathbf{a} \neq 0$,

$$\begin{aligned}
\text{sum}(\mathbf{X}_\beta) &= \sum_{(i,j)\in\beta} \mathbf{X}_{i,j} = \sum_{i=}^{d} C_i a_i \\
&> c(r-\varepsilon')\sum_{i=1}^{d} a_i \\
&= (r-\varepsilon')\text{sum}(\mathbf{X})
\end{aligned}$$

■

Using Lemma C.7, we can calculate a lower bound for $g(\mathbf{E}^L)$ with high probability as long as the cluster sizes are sufficiently large. Due to (C.15) and the linearity of $g(\mathbf{E}^L)$, we can focus on contributions due to specific clusters i.e. $\mathbf{v}_i\mathbf{m}_i^T + \mathbf{n}_i\mathbf{v}_i^T$ for the i'th cluster. We additionally know the simple structure of $\mathbf{m}_i, \mathbf{n}_i$ from Lemma C.6. In particular, subvectors $\mathbf{m}_i^i$ and $\mathbf{n}_i^i$ of $\mathbf{m}_i, \mathbf{n}_i$ can be assumed to be nonpositive and rest of the entries are nonnegative.

Lemma C.8 *Assume,* $l \leq K$, $\mathbf{D}_{\mathscr{A}} > 0$ *and (without loss of generality)* $\mathbf{m}_l$ *has the structure described in Lemma C.6. Then, with probability* $1 - n\exp(-2\mathbf{D}_{\mathscr{A}}^2(n_l-1))$, *we have* $g(\mathbf{v}_l\mathbf{m}_l^T) \geq 0$ *for all* $\mathbf{m}_l$. *Also, if* $\mathbf{m}_l \neq 0$ *then inequality is strict.*

Proof: Recall that $\mathbf{m}_l$ satisfies $\mathbf{m}_l^i$ is nonpositive/nonnegative when $i = l/i \neq l$ for all i. Call $\mathbf{X}^i = \mathbf{1}^{n_l}\mathbf{m}_l^{iT}$. We can write

$$g(\mathbf{v}_l\mathbf{m}_l^T) = \frac{1}{n_l}\text{sum}(\mathbf{X}^l) + \sum_{i=1}^{K}\lambda h(\mathbf{X}^i, \beta_{l,i}^c)$$

where $h(\mathbf{X}^i, \beta_{l,i}^c) = \text{sum}(\mathbf{X}^i_{\beta_{l,i}^c}) - \text{sum}(\mathbf{X}^i_{\beta_{l,i}})$. Now assume $i \neq l$. Using Lemma C.7 and the fact that $\beta_{l,i}$ is a randomly generated subset (with parameter q), with probability $1 - n_i\exp(-2\varepsilon'^2 n_l)$, for all $\mathbf{X}^i$, we have,

$$\begin{aligned} h(\mathbf{X}^i, \beta_{l,i}^c) &\geq (1 - q - \varepsilon')\text{sum}(\mathbf{X}^i) - (q + \varepsilon')\text{sum}(\mathbf{X}^i) \\ &= (1 - 2q - 2\varepsilon')\text{sum}(\mathbf{X}^i) \end{aligned}$$

where inequality is strict if $X^i \neq 0$. Similarly, when $i = l$ with probability $1 - n_l\exp(-2\varepsilon'^2(n_l - 1))$, we have,

$$\begin{aligned} \frac{1}{\lambda n_l}\text{sum}(\mathbf{X}^l) + h(\mathbf{X}^l, \beta_{l,l}^c) \geq & \\ \left(1 - p_l + \varepsilon' + \frac{1}{\lambda n_l}\right)\text{sum}(\mathbf{X}^l) &- \left(p_l - \varepsilon'\right)\text{sum}(\mathbf{X}^l) \\ = -\left(2p_l - 1 - \frac{1}{\lambda n_l} - 2\varepsilon'\right)&\text{sum}(\mathbf{X}^l) \end{aligned} \tag{C.21}$$

Choosing $\varepsilon' = \frac{\mathbf{D}_{\mathscr{A}}}{2}$ and using the facts that $1 - 2q - 2\mathbf{D}_{\mathscr{A}} \geq 0$, $2p_l - 1 - \frac{1}{\lambda n_l} - 2\mathbf{D}_{\mathscr{A}} \geq 0$ and using a union bound, with probability $1 - n\exp(-2\mathbf{D}^2_{\mathscr{A}}(n_l - 1))$, we have $g(\mathbf{v}_l\mathbf{m}_l^T) \geq 0$ and the inequality is strict when $\mathbf{m}_l \neq 0$ as at least one of the $\mathbf{X}^i$'s will be nonzero. ■

The following lemma immediately follows from Lemma C.8 and summarizes the main result of the section.

Lemma C.9 *Let $\mathbf{D}_{\mathscr{A}}$ be as defined in (C.1) and assume $\mathbf{D}_{\mathscr{A}} > 0$. Then with probability $1 - 2nK\exp(-2\mathbf{D}^2_{\mathscr{A}}(n_{min} - 1))$ we have $g(\mathbf{E}^L) > 0$ for all nonzero feasible $\mathbf{E}^L \in \mathscr{M}_U^{\perp}$.*

For the proof, we basically use the fact that $\mathbf{E}^L$ is linear superposition of $\mathbf{v}_l\mathbf{m}_l^T$'s and $\mathbf{n}_l\mathbf{v}_l^T$'s and if $\mathbf{E}^L \neq 0$, due to Lemmas C.6 and C.8 at least one of the $\mathbf{v}_l\mathbf{m}_l^T$ (or $\mathbf{n}_l\mathbf{v}_l^T$) terms are nonzero and has a strictly positive contribution to $g(\mathbf{E}^L)$.

C.1.4 The Final Step

Lemma C.10 *Let $p_{min} > \frac{1}{2} > q$ and $\mathbf{G}$ be a random graph generated according to Model 6.1 with cluster sizes $\{n_i\}_{i=1}^K$. If $\lambda \leq (1 - \varepsilon)\Lambda_{succ}$ and $\mathbf{ED}_{\min} = \min_{1 \leq i \leq n}(2p_i - 1)n_i \geq (1 + \varepsilon)\frac{1}{\lambda}$, then $\left(\mathbf{L}^0, \mathbf{S}^0\right)$ is the unique*

optimal solution to Program 6.1 with probability $1-\exp(-\Omega(n))-6n^2\exp(-\Omega(n_{min}))$.

Proof: Based on Lemma C.4 and Lemma C.9, with probability $1-cn^2\exp(-C\min\{1-2q, 2p_{min}-1\}^2 n_{min})$,

- There exists $\mathbf{W}\in\mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\|<1$ such that for all feasible $\mathbf{E}^L$, $f(\mathbf{E}^L,\mathbf{W})\geq 0$.
- For all nonzero $\mathbf{E}^L\in\mathscr{M}_{\mathbf{U}}^{\perp}$ we have $g(\mathbf{E}^L)>0$.

Consequently based on Lemma C.1, $(\mathbf{L}^0,\mathbf{S}^0)$ is the unique optimal of Problem 6.1. ■

C.2 On the failure of the simple program

This section will provide the proofs of the failure results, i.e., the initial statements of Theorems 6.1 and 6.2. Let us start by arguing that, failure result of Theorem 6.2 implies failure result of Theorem 6.1. To see this, assume Theorem 6.2 holds and $\mathbf{ED}_{\min}\leq(1-\varepsilon)\Lambda_{fail}^{-1}$. Let ε' be a constant to be determined. If $\lambda\geq(1+\varepsilon')\Lambda_{fail}$ or $\mathbf{ED}_{\min}\leq(1-\varepsilon')\lambda^{-1}$, due to Theorem 6.2, Program 6.1 would fail and we can conclude. Suppose, these are not the case, i.e., $\lambda\leq(1+\varepsilon')\Lambda_{fail}$ and $\mathbf{ED}_{\min}\geq(1-\varepsilon')\lambda^{-1}$. These would imply, $\mathbf{ED}_{\min}\geq\frac{1-\varepsilon'}{1+\varepsilon'}\Lambda_{fail}^{-1}$. We can end up with a contradiction by choosing ε' small enough to ensure $\frac{1-\varepsilon'}{1+\varepsilon'}>1-\varepsilon$. Consequntly, we will only prove Theorem 6.2.

Lemma C.11 *Let* $p_{min}>\frac{1}{2}>q$ *and* $\mathbf{G}$ *be a random graph generated according to the Model 6.1 with cluster sizes* $\{n_i\}_{i=1}^K$.

1. *If* $\min_i\{n_i(2p_i-1)\}\leq(1-\varepsilon)\frac{1}{\lambda}$, *then* $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ *is not an optimal solution to the Program 6.1 with probability at least* $1-K\exp\left(-\Omega(n_{\min}^2)\right)$.
2. *If* $\lambda\geq(1+\varepsilon)\sqrt{\frac{n}{q\left(n^2-\sum_{i=1}^K n_i^2\right)}}$, *then* $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ *is not an optimal solution to the Program 6.1 with high probability.*

Proof:

Proof of the first statement: Choose ε' to be a constant satisfying $2p_i-1+\varepsilon'<\frac{1}{\lambda n_i}$ for some $1\leq i\leq K$. This is indeed possible if the assumption of the Statement 1 of Lemma C.11 holds. Lagrange for the Problem 6.1 can be written as follows,

$$\mathscr{L}(\mathbf{L},\mathbf{S};\mathbf{M},\mathbf{N})=\|\mathbf{L}\|_\star+\lambda\|\mathbf{S}\|_1+\mathrm{trace}(\mathbf{M}(\mathbf{L}-\mathbb{1}^{n\times n}))$$
$$-\mathrm{trace}(\mathbf{NL}). \tag{C.22}$$

where $\mathbf{M}$ and $\mathbf{N}$ are dual variables corresponding to the inequality constraints (6.2).

For $\mathbf{L}^0$ to be an optimal solution to (6.1), it has to satisfy the KKT conditions. Therefore, the subgradient of (C.22) at $\mathbf{L}^0$ has to be 0, i.e.,

$$\partial\|\mathbf{L}^0\|_\star + \lambda\ \partial\|\mathbf{A}-\mathbf{L}^0\|_1 + \mathbf{M}^0 - \mathbf{N}^0 = 0. \tag{C.23}$$

where $\mathbf{M}^0$ and $\mathbf{N}^0$ are optimal dual variables.

Also, by complementary slackness,

$$\text{trace}(\mathbf{M}^0(\mathbf{L}^0 - \mathbb{1}^{n\times n})) = 0, \tag{C.24}$$

and

$$\text{trace}(\mathbf{N}^0\mathbf{L}^0) = 0. \tag{C.25}$$

From (6.7), (C.24), and (C.25), we have $(\mathbf{M}^0)_{\mathscr{R}} \geq 0$, $(\mathbf{M}^0)_{\mathscr{R}^c} = 0$, $(\mathbf{N}^0)_{\mathscr{R}} = 0$ and $(\mathbf{N}^0)_{\mathscr{R}^c} \geq 0$. Hence $(\mathbf{M}^0 - \mathbf{N}^0)_{\mathscr{R}} \geq 0$ and $(\mathbf{M}^0 - \mathbf{N}^0)_{\mathscr{R}^c} \leq 0$.

Recall, $\mathbf{L}^0 = \mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{U} = [\mathbf{u}_1\ \ldots\ \mathbf{u}_K] \in \mathbb{R}^{n\times K}$,

$$\mathbf{u}_{l,i} = \begin{cases} \frac{1}{\sqrt{k_l}} & \text{if } i \in \mathscr{C}_l \\ 0 & \text{else.} \end{cases}$$

Also, recall that the subgradient $\partial\|\mathbf{L}^0\|_\star$ is of the form $\mathbf{U}\mathbf{U}^T + \mathbf{W}$ such that $\mathbf{W} \in \{\mathbf{X} : \mathbf{X}\mathbf{U} = \mathbf{U}^T\mathbf{X} = 0, \|\mathbf{X}\| \leq 1\}$. The subgradient $\partial\|\mathbf{S}^0\|_1$ is of the form $\text{sign}(\mathbf{S}^0) + \mathbf{Q}$ where $\mathbf{Q}_{i,j} = 0$ if $\mathbf{S}_{i,j} \neq 0$ and $\|\mathbf{Q}\|_\infty \leq 1$.

From (C.23), we have,

$$\mathbf{U}\mathbf{U}^T + \mathbf{W} - \lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right) + (\mathbf{M}^0 - \mathbf{N}^0) = 0. \tag{C.26}$$

Consider the sum of the entires corresponding $\mathscr{R}_{i,i}$, i.e.,

$$\underbrace{\text{sum}\left(\mathbf{L}^0_{\mathscr{R}_{i,i}}\right)}_{n_i} - \text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0) + \mathbf{Q}\right)_{\mathscr{R}_{i,i}}\right) + \underbrace{\text{sum}\left((\mathbf{M}^0 - \mathbf{N}^0)_{\mathscr{R}_{i,i}}\right)}_{\geq 0} = 0. \tag{C.27}$$

By Bernstein's inequality and using $\|\mathbf{Q}\|_\infty \leq 1$, with probability $1-\exp\left(-\Omega(n_i^2)\right)$ we have,

$$\text{sum}\left(\text{sign}(\mathbf{S}^0)\right) \leq -n_i^2(1-p_i-\frac{\varepsilon'}{2}) \tag{C.28}$$

$$\text{sum}\left(\mathbf{Q}\right) \leq n_i^2(p_i+\frac{\varepsilon'}{2}). \tag{C.29}$$

Thus, $-\text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0)+\mathbf{Q}\right)_{\mathscr{R}_{i,i}}\right) \geq \lambda n_i^2(1-2p_i-\varepsilon')$ and hence,

$$\underbrace{\text{sum}\left(\mathbf{L}^0_{\mathscr{R}_{i,i}}\right)}_{n_i} - \text{sum}\left(\lambda\left(\text{sign}(\mathbf{S}^0)+\mathbf{Q}\right)_{\mathscr{R}_{i,i}}\right) + \underbrace{\text{sum}\left(\left(\mathbf{M}^0-\mathbf{N}^0\right)_{\mathscr{R}_{i,i}}\right)}_{\geq 0} \geq n_i + \lambda n_i^2(1-2p_i-\varepsilon').$$

Now, choose $i = \arg\min_{1\leq j\leq K} n_i(2p_i-1)$. From the initial choice of ε', we have that $n_i+\lambda n_i^2(1-2p_i-\varepsilon') > 0$. Consequently, the equation (C.23) does not hold and hence $\mathbf{L}^0$ cannot be an optimal solution to the Program 6.1.

Proof of the second statement: Let ε' be a constant to be determined. Notice that $\left(\mathbf{U}\mathbf{U}^T\right)_{\mathscr{R}^c} = 0$ and the entries of $-\left(\text{sign}(\mathbf{S}^0)+\mathbf{Q}\right)$ and $\mathbf{M}^0-\mathbf{N}^0$ over $\mathscr{R}^c \cap \mathscr{A}$ are nonpositive. Hence from (C.26),

$$\begin{aligned}\|\mathbf{W}\|_F^2 &\geq \|\left(\mathbf{U}\mathbf{U}^T+\mathbf{W}\right)_{\mathscr{R}^c\cap\mathscr{A}}\|_F^2 \\ &\geq \|\lambda\left(\text{sign}(\mathbf{S}^0)+\mathbf{Q}\right)_{\mathscr{R}^c\cap\mathscr{A}}\|_F^2.\end{aligned} \tag{C.30}$$

Recall that $\mathbf{S}^0_{\mathscr{R}^c\cap\mathscr{A}} \neq 0$ and hence $\mathbf{Q}_{\mathscr{R}^c\cap\mathscr{A}} = 0$. Further, recall that by Model 6.1, each entry of $\mathbf{A}$ over $\mathscr{R}^c$ is non-zero with probability q. Hence with probability at least $1-\exp\left(-\Omega(|\mathscr{R}^c|)\right)$, $|\mathscr{R}^c\cap\mathscr{A}| \geq (q-\varepsilon')(n^2-\sum_{i=1}^K n_i^2)$. Thus from (C.30) we have,

$$\|\mathbf{W}\|_F^2 \geq \lambda^2(q-\varepsilon')(n^2-\sum_{i=1}^K n_i^2), \tag{C.31}$$

Recall that $\|\mathbf{W}\| \leq 1$ should hold true for $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ to be an optimal solution to the Program 6.1. Using the standard inequality $n\|\mathbf{W}\|^2 \geq \|\mathbf{W}\|_F^2$ and the equation (C.31), we find,

$$\|\mathbf{W}\| \geq \lambda\sqrt{\frac{(q-\varepsilon')\left(n^2-\sum_{i=1}^K n_i^2\right)}{n}}.$$

So, if $\lambda\sqrt{q(1-\varepsilon')\left(n^2-\sum_{i=1}^K n_i^2\right)/n} > 1$ then, $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ cannot be an optimal solution to Program 6.1. This is indeed the case with the choice $(1-\varepsilon')^{-1/2} < (1+\varepsilon)$. This gives us the Statement 2 of Lemma C.11.

■

C.3 Proof of Theorem 6.3

This section will show that, the optimal solution of Problem 6.3 is the pair $(\mathbf{L}^0,\mathbf{S}^0)$ under reasonable conditions, where,

$$\mathbf{L}^0 = \mathbf{1}^{n\times n}_{\mathscr{R}},\ \mathbf{S}^0 = \mathbf{1}^{n\times n}_{\mathscr{R}\cap\mathscr{A}^c} \tag{C.32}$$

Also denote the true optimal pair by $(\mathbf{L}^*,\mathbf{S}^*)$. Let $1 \geq p_{min} > q > 0$. $\mathbf{G}$ be a random graph generated according to the stochastic block model 6.1 with cluster sizes $\{n_i\}_{i=1}^K$. Theorem 6.3 is based on the following lemma:

Lemma C.12 *If $\lambda < \tilde{\Lambda}_{succ}$ and $\tilde{\mathbf{ED}}_{\min} > \frac{1}{\lambda}$, then $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ is the unique optimal solution to Program 6.3 with high probability.*

Given $q,\{p_i\}_{i=1}^K$, define the following parameter which will be useful for the subsequent analysis. This parameter can be seen as a measure of distinctness of the "worst" cluster from the "background noise". Here, by background noise we mean the edges over $\mathscr{R}^c$.

$$\begin{aligned}\tilde{\mathbf{ED}}_{\mathscr{A}} &= \frac{1}{2}\min_{1\leq i\leq K}\left\{(p_i-q)-\frac{1}{\lambda n_i}\right\} \\ &= \frac{1}{2}\min_{1\leq i\leq K}\frac{\tilde{\mathbf{ED}}_i-\lambda^{-1}}{n_i}\end{aligned} \tag{C.33}$$

C.3.1 Perturbation Analysis

Our aim is to show that $\left(\mathbf{L}^0,\mathbf{S}^0\right)$ defined in (C.32) is unique optimal solution to Problem 6.3.

Lemma C.13 *Let $(\mathbf{E}^L,\mathbf{E}^S)$ be a feasible perturbation. Then, the objective will increase by at least,*

$$f(\mathbf{E}^L,\mathbf{W}) = \sum_{i=1}^K \frac{1}{n_i} sum(\mathbf{E}^L_{\mathscr{R}_{i,i}}) + \langle \mathbf{E}^L,\mathbf{W}\rangle + \lambda\, sum(\mathbf{E}^L_{\mathscr{A}^c}) \tag{C.34}$$

for any $\mathbf{W}\in\mathscr{M}_{\mathbf{U}}$, $\|\mathbf{W}\|\leq 1$.

Proof: From constraint (6.5), we have $\mathbf{L}_{i,j} = \mathbf{S}_{i,j}$ whenever $\mathbf{A}_{i,j} = 0$. Entries of $\mathbf{S}$ over $\mathscr{A}$ are not constrained by the Improved Program hence, they will be equal to 0. Since, if they are not zero, setting them to be zero will strictly decrease the objective $\|\mathbf{S}\|_1$ without effecting the feasibility of the solution. Hence,

$$\mathbf{S}^* = \mathbf{L}^*_{\mathscr{A}^c}.$$

Recall that,

$$\begin{aligned}\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda\ \|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda\ \|\mathbf{S}^0\|_1) &\geq \langle \partial\|\mathbf{L}^0\|_\star, \mathbf{E}^L\rangle + \lambda\langle \partial\|\mathbf{S}^0\|_1, \mathbf{E}^S\rangle \\ &= \langle \mathbf{U}\mathbf{U}^T + \mathbf{W}, \mathbf{E}^L\rangle + \lambda\langle \operatorname{sign}(\mathbf{S}^0) + \mathbf{Q}, \mathbf{E}^S\rangle\end{aligned}$$

Using $\operatorname{sign}(\mathbf{S}^0) = \mathbb{1}^{n\times n}_{\mathscr{A}^c\cap\mathscr{R}}$, and choosing $\mathbf{Q} = \mathbb{1}^{n\times n}_{\mathscr{A}^c - (\mathscr{A}^c\cap\mathscr{R})}$, we get,

$$\begin{aligned}\|\mathbf{L}^0 + \mathbf{E}^L\|_\star + \lambda\ \|\mathbf{S}^0 + \mathbf{E}^S\|_1 - (\|\mathbf{L}^0\|_\star + \lambda\ \|\mathbf{S}^0\|_1) &\geq \langle \mathbf{W}, \mathbf{E}^L\rangle \\ &\quad + \underbrace{\sum_{i=1}^{K} \frac{1}{n_i}\operatorname{sum}(\mathbf{E}^L_{R_{i,i}}) + \lambda\left(\operatorname{sum}(\mathbf{E}^L_{\mathscr{A}^c})\right)}_{:=g(\mathbf{E}^L)} \qquad \text{(C.35)}\end{aligned}$$

for any $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$. ■

From this point onward, for simplicity we will ignore the superscript L on $\mathbf{E}^L$ and just use $\mathbf{E}$. Define,

$$g(\mathbf{E}) := \sum_{i=1}^{K} \frac{1}{n_i}\operatorname{sum}(\mathbf{E}_{\mathscr{R}_{i,i}}) + \lambda\operatorname{sum}(\mathbf{E}_{\mathscr{A}^c})). \qquad \text{(C.36)}$$

Also, define $f(\mathbf{E}, \mathbf{W}) := g(\mathbf{E}) + \langle \mathbf{W}, \mathbf{E}\rangle$. Our aim is to show that for all feasible perturbations $\mathbf{E}$, there exists $\mathbf{W}$ such that,

$$f(\mathbf{E}, \mathbf{W}) = g(\mathbf{E}) + \langle \mathbf{W}, \mathbf{E}\rangle > 0. \qquad \text{(C.37)}$$

Note that $g(\mathbf{E})$ does not depend on $\mathbf{W}$.

We can directly use Lemma C.1. So, as in the previous section, we have broken down our aim into two steps.

1. Construct $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$, such that $f(\mathbf{E}, \mathbf{W}) \geq 0$ for all feasible perturbations $\mathbf{E}$.
2. For all non-zero feasible $\mathbf{E} \in \mathscr{M}_{\mathbf{U}}^{\perp}$, show that $g(\mathbf{E}) > 0$.

As a first step, in Section C.3.2, we will argue that, under certain conditions, there exists a $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$ such that with high probability, $f(\mathbf{E}, \mathbf{W}) \geq 0$ for all feasible $\mathbf{E}$. Recall that such a $\mathbf{W}$ is called the dual certificate. Secondly, in Section C.3.3, we will show that, under certain conditions, for all $\mathbf{E} \in \mathscr{M}_{\mathbf{U}}^{\perp}$ with high probability, $g(\mathbf{E}) > 0$. Finally, combining these two arguments, and using Lemma C.1 we will conclude that $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal with high probability.

C.3.2 Showing existence of the dual certificate

Recall that

$$f(\mathbf{E}, \mathbf{W}) = \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{E}_{\mathscr{R}_{i,i}}) + \langle \mathbf{E}, \mathbf{W} \rangle + \lambda \, \text{sum}\,(\mathbf{E}_{\mathscr{A}^c})$$

$\mathbf{W}$ will be constructed from the candidate $\mathbf{W}_0$, which is given as follows.

C.3.2.1 Candidate $\mathbf{W}_0$

Based on Program 6.3, we propose the following,

$$\mathbf{W}_0 = \sum_{i=1}^{K} c_i \mathbb{1}^{n \times n}_{\mathscr{R}_{i,i}} + c \, \mathbb{1}^{n \times n} - \lambda \, \mathbb{1}^{n \times n}_{\mathscr{A}^c},$$

where $\{c_i\}_{i=1}^{K}, c$ are real numbers to be determined.

$$f(\mathbf{E}, \mathbf{W}^0) = \sum_{i=1}^{K} (\frac{1}{n_i} + c_i) \text{ sum}(\mathbf{E}_{\mathscr{R}_{i,i}}) + c \text{ sum}(\mathbf{E})$$

Note that $\mathbf{W}_0$ is a random matrix where randomness is due to $\mathbf{A}$. In order to ensure a small spectral norm, we will set its expectation to 0, i.e., we will choose $c, \{c_i\}'s$ to ensure that $\mathbb{E}[\mathbf{W}_0] = 0$.

Following from the Stochastic Block Model 6.1, the expectation of an entry of $\mathbf{W}_0$ on $\mathscr{R}_{i,i}$ (region corresponding to cluster i) and $\mathscr{R}^c$ (region outside the clusters) is $c_i + \lambda(p_i - q)$ and $c + \lambda(q-1)$ respectively. Hence, we set,

$$c_i = -\lambda(p_i - q) \quad \text{and} \quad c = \lambda(1-q),$$

With these choices, the candidate $\mathbf{W}_0$ and $f(\mathbf{E}, \mathbf{W}_0)$ take the following forms,

$$\begin{aligned} \mathbf{W}_0 \;\; = \;\; & \lambda \left[\sum_{i=1}^{K} -p_i \, \mathbb{1}^{n \times n}_{\mathscr{R}_{i,i} \cap \mathscr{A}^c} + (1-p_i) \mathbb{1}^{n \times n}_{\mathscr{R}_{i,i} \cap \mathscr{A}} \right] \\ & + \lambda \left[-q \mathbb{1}^{n \times n}_{\mathscr{R}^c \cap \mathscr{A}^c} + (1-q) \mathbb{1}^{n \times n}_{\mathscr{R}^c \cap \mathscr{A}} \right] \end{aligned} \tag{C.38}$$

$$f(\mathbf{E},\mathbf{W}_0) = \lambda\left[(1-q)\ \mathrm{sum}(\mathbf{E})\right] - \lambda\left[\sum_{i=1}^{K}\left((p_i-q)-\frac{1}{\lambda n_i}\right)\ \mathrm{sum}(\mathbf{E}_{\mathscr{R}_{i,i}})\right]$$

From $\mathbf{L}^0$ and the constraint $1 \geq \mathbf{L}_{i,j} \geq 0$, it follows that,

$$\begin{aligned}&\mathbf{E}_{\mathscr{R}^c} \text{ is (entrywise) nonnegative.}\\ &\mathbf{E}_{\mathscr{R}} \text{ is (entrywise) nonpositive.}\end{aligned} \tag{C.39}$$

Thus, $\mathrm{sum}(\mathbf{E}_{\mathscr{R}^c}) \leq 0$ and $\mathrm{sum}(\mathbf{E}_{\mathscr{R}_{i,i}}) \geq 0$. When $\lambda(p_i-q)-\frac{1}{n_i} \geq 0$ and $\lambda(1-q) \geq 0$; we will have $f(\mathbf{E},\mathbf{W}_0) \geq 0$ for all feasible $\mathbf{E}$. This indeed holds due to the assumptions of Theorem 6.3 (see (C.33)), as we assumed $p_i - q > \frac{1}{\lambda n_i}$ for $i = 1,2\cdots,K$ and $1 > q$.

Using the same technique as in Theorem C.1, we can bound the spectral norm of $\mathbf{W}^0$ as follows

Lemma C.14 *Recall that, $\mathbf{W}_0$ is a random matrix; where randomness is on the stochastic block model $\mathbf{A}$ and it is given by,*

$$\begin{aligned}\mathbf{W}_0 \quad = \quad & \lambda\left[\sum_{i=1}^{K} -p_i\ \mathbb{1}^{n\times n}_{\mathscr{R}_{i,i}\cap\mathscr{A}^c} + (1-p)\mathbb{1}^{n\times n}_{\mathscr{R}_{i,i}\cap\mathscr{A}}\right]\\ & +\lambda\left[-q\mathbb{1}^{n\times n}_{\mathscr{R}^c\cap\mathscr{A}^c} + (1-q)\mathbb{1}^{n\times n}_{\mathscr{R}^c\cap\mathscr{A}}\right]\end{aligned}$$

Then, for any $\varepsilon' > 0$, with probability $1-\exp(-\Omega(n))$, we have

$$\|\frac{1}{\lambda}\mathbf{W}_0\| \leq 2\sqrt{n}\sqrt{(1-q)q} + \max_{1\leq i\leq K} 2\sqrt{n_i}\sqrt{(1-p_i)p_i+(1-q)q} + \varepsilon'\sqrt{n}$$

Further, if $\max_{1\leq i\leq K} n_i = o(n)$. Then, for sufficiently large n, with the same probability,

$$\|\mathbf{W}_0\| \leq 2\lambda\sqrt{n}\sqrt{(1-q)q} + \varepsilon'\lambda\sqrt{n}.$$

Lemma C.14 verifies that asymptotically with high probability we can make $\|\mathbf{W}_0\| < 1$ as long as λ is sufficiently small. However, $\mathbf{W}_0$ itself is not sufficient for construction of the desired $\mathbf{W}$, since we do not have any guarantee that $\mathbf{W}_0 \in \mathscr{M}_{\mathbf{U}}$. In order to achieve this, we will *correct* $\mathbf{W}_0$ by projecting it onto $\mathscr{M}_{\mathbf{U}}$. Lemma C.3 can be used to here.

Recall that, $\tilde{\gamma}_{\mathrm{succ}} := 2\max_{1\leq i\leq K}\sqrt{n_i}\sqrt{(1-p_i)p_i+(1-q)q}$ and $\tilde{\Lambda}^{-1}_{\mathrm{succ}} := 2\sqrt{n}\sqrt{q(1-q)} + \tilde{\gamma}_{\mathrm{succ}}$.

We can summarize our discussion so far in the following lemma,

Lemma C.15 *$\mathbf{W}_0$ is as described previously in (C.38). Choose $\mathbf{W}$ to be projection of $\mathbf{W}_0$ on $\mathscr{M}_{\mathbf{U}}$. Also suppose $\lambda \leq (1-\delta)\tilde{\Lambda}_{succ}$. Then, with probability $1-6n^2\exp(-\Omega(n_{min}))-4\exp(-\Omega(n))$ we have,*

- $\|\mathbf{W}\| < 1$
- *For all feasible* $\mathbf{E}$, $f(\mathbf{E},\mathbf{W}) \geq 0$.

Proof: To begin with, observe that $\tilde{\Lambda}_{\text{succ}}^{-1}$ is $\Omega(\sqrt{n})$. Since $\lambda \leq \tilde{\Lambda}_{\text{succ}}$, $\lambda\sqrt{n} = \mathscr{O}(1)$. Consequently, using $\lambda\tilde{\Lambda}_{\text{succ}}^{-1} < 1$ and applying Lemma C.14, and choosing a sufficiently small $\varepsilon' > 0$, we conclude with,

$$\|\mathbf{W}\| \leq \|\mathbf{W}_0\| < 1$$

with probability $1-\exp(-\Omega(n))$ where the constant in the exponent depends on the constant $\varepsilon' > 0$.

Next, from Lemma C.3 with probability $1-6n^2\exp(-\frac{2}{9}\varepsilon''^2 n_{min})$ we have $\|\mathbf{W}_0 - \mathbf{W}\|_\infty \leq \lambda\varepsilon''$. Then based on (C.39) for all $\mathbf{E}$, we have that,

$$\begin{aligned}
f(\mathbf{E},\mathbf{W}) &= f(\mathbf{E},\mathbf{W}_0) - \langle \mathbf{W}_0 - \mathbf{W}, \mathbf{E}\rangle \\
&\geq f(\mathbf{E},\mathbf{W}_0) - \lambda\varepsilon''\left(\text{sum}(\mathbf{E}_{\mathscr{R}}) - \text{sum}(\mathbf{E}_{\mathscr{R}^c})\right) \\
&= \lambda\left[((1-q)-\varepsilon'')\text{sum}(\mathbf{E}_{\mathscr{R}^c})\right] \\
&\quad -\lambda\sum_{i=1}^{K}\left[\left((p_i-q)-\frac{1}{\lambda n_i}-\varepsilon''\right)\text{sum}(\mathbf{E}_{\mathscr{R}_{i,i}})\right] \\
&\geq 0
\end{aligned}$$

where we chose ε'' to be a sufficiently small constant. In particular, we set $\varepsilon'' < \tilde{\mathbf{E}}\mathbf{D}_{\mathscr{A}}$, i.e., set $\varepsilon'' < 1-q$ and $\varepsilon'' < (p_i-q)-\frac{1}{\lambda n_i}$ for all $1 \leq i \leq K$.

Hence, by using a union bound $\mathbf{W}$ satisfies both of the desired conditions. ■

Summary so far: Combining the last lemma with Lemma C.1, with high probability, either there exists a dual vector $\mathbf{W}^*$ which ensures $f(\mathbf{E},\mathbf{W}^*) > 0$ or $\mathbf{E} \in \mathscr{M}_{\mathbf{U}}^{\perp}$. If former, we are done. Hence, we need to focus on the latter case and show that for all perturbations $\mathbf{E} \in \mathscr{M}_{\mathbf{U}}^{\perp}$, the objective will strictly increase at $(\mathbf{L}^0,\mathbf{S}^0)$ with high probability.

C.3.3 Solving for $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$ case

Recall that,

$$g(\mathbf{E}) = \sum_{i=1}^{K}\frac{1}{n_i}\text{sum}(\mathbf{E}_{R_{i,i}}) + \lambda\,\text{sum}(\mathbf{E}_{\mathscr{A}^c})$$

Let us define,

$$g_1(\mathbf{X}) := \sum_{i=1}^{K} \frac{1}{n_i} \text{sum}(\mathbf{X}_{\mathscr{R}_{i,i}}),$$

$$g_2(\mathbf{X}) := \text{sum}(\mathbf{X}_{\mathscr{A}^c}),$$

so that, $g(\mathbf{X}) = g_1(\mathbf{X}) + \lambda g_2(\mathbf{X})$. Also let $\mathbf{V} = [\mathbf{v}_1 \ \dots \ \mathbf{v}_K]$ where $\mathbf{v}_i = \sqrt{n_i}\mathbf{u}_i$. Thus, $\mathbf{V}$ is basically obtained by, normalizing columns of $\mathbf{U}$ to make its nonzero entries 1. Assume $\mathbf{E} \in \mathscr{M}_{\mathbf{U}}^{\perp}$. Then, by definition of $\mathscr{M}_{\mathbf{U}}^{\perp}$, we can write,

$$\mathbf{E} = \mathbf{V}\mathbf{M}^T + \mathbf{N}\mathbf{V}^T.$$

Let $\mathbf{m}_i, \mathbf{n}_i$ denote i'th columns of $\mathbf{M}, \mathbf{N}$ respectively.

Again as in previous section C.1.3, we consider optimization problem C.17. Since $g_1(\mathbf{E})$ is fixed, we just need to optimize over $g_2(\mathbf{E})$. This optimization can be reduced to local optimizations C.19. Since $\mathbf{L}^0 = \mathbb{1}_{\mathscr{R}}^{n \times n}$ and the condition (6.4),

$$\mathbf{E}_{\mathscr{R}^c} \text{ is (entrywise) nonnegative}$$

$$\mathbf{E}_{\mathscr{R}} \text{ is (entrywise) nonpositive}$$

We can make use of Lemma C.6 and assume $\mathbf{m}_l^{\mathscr{C}_i}$ is nonpositive/nonnegative when $i = l / i \neq l$ for all i. Hence using Lemma C.7 we lower bound $g(\mathbf{v}_l \mathbf{m}_l^T)$ as described in the following section.

C.3.3.1 Lower bounding $g(\mathbf{E})$

Lemma C.16 *Assume,* $1 \leq l \leq K$, $\mathbf{E}\tilde{\mathbf{D}}_{\mathscr{A}} > 0$. *Then, with probability* $1 - n\exp(-2\mathbf{E}\tilde{\mathbf{D}}_{\mathscr{A}}^2(n_l - 1))$, *we have* $g(\mathbf{v}_l \mathbf{m}_l^T) \geq \lambda(1 - q - \mathbf{E}\tilde{\mathbf{D}}_{\mathscr{A}}) sum(\mathbf{v}_l \mathbf{m}_l^T)$ *for all* $\mathbf{m}_l$. *Also, if* $\mathbf{m}_l \neq 0$ *then inequality is strict.*

Proof: Recall that $\mathbf{m}_l$ satisfies $\mathbf{m}_l^{\mathscr{C}_i}$ is nonpositive/nonnegative when $i = l / i \neq l$ for all i. Define $\mathbf{X}^i := \mathbf{1}^{n_l} \mathbf{m}_l^{\mathscr{C}_i T}$. We can write

$$g(\mathbf{v}_l \mathbf{m}_l^T) = \frac{1}{n_l} \text{sum}(\mathbf{X}^l) + \sum_{i=1}^{K} \lambda \text{sum}(\mathbf{X}^i_{\beta^c_{l,i}})$$

Now assume $i \neq l$. Using Lemma C.7 and the fact that $\beta_{l,i}$ is a randomly generated subset (with parameter q), with probability $1 - n_i \exp(-2\varepsilon'^2 n_l)$, for all $\mathbf{X}^i$, we have,

$$\text{sum}(\mathbf{X}^i_{\beta^c_{l,i}}) \geq ((1-q)-\varepsilon')\text{sum}(\mathbf{X}^i) \tag{C.40}$$

where inequality is strict if $X^i \neq 0$. Similarly, when $i = l$ with probability $1 - n_l \exp(-2\varepsilon'^2(n_l - 1))$, we have,

$$\frac{1}{\lambda n_l}\text{sum}(\mathbf{X}^l) + \text{sum}(\mathbf{X}^l_{\beta^c_{l,l}}) \geq \left(\frac{1}{\lambda n_l} + (1-p_l) + \varepsilon'\right)\text{sum}(\mathbf{X}^l)$$

Together,

$$\begin{aligned} g(\mathbf{v}_l\mathbf{m}_l^T) &\geq \lambda \sum_{i\neq l}((1-q)-\varepsilon')\text{sum}(\mathbf{X}^i) + \left(\frac{1}{\lambda n_l} + (1-p_l) + \varepsilon'\right)\text{sum}(\mathbf{X}^l) \\ &\geq \lambda((1-q)-\varepsilon')\sum_{i=1}^{K}\text{sum}(\mathbf{X}^i) = \lambda((1-q)-\varepsilon')\text{sum}(\mathbf{v}_l\mathbf{m}_l^T) \end{aligned}$$

Choosing $\varepsilon' = \tilde{\mathbf{ED}}_{\mathscr{A}}$ and using the facts that $(1-q) - 2\tilde{\mathbf{ED}}_{\mathscr{A}} \geq 0$, $(p_l - q) - \frac{1}{\lambda n_l} - 2\tilde{\mathbf{ED}}_{\mathscr{A}} \geq 0$ and using a union bound, with probability $1 - n\exp(-2\tilde{\mathbf{ED}}^2_{\mathscr{A}}(n_l - 1))$, we have $g(\mathbf{v}_l\mathbf{m}_l^T) \geq 0$ and the inequality is strict when $\mathbf{m}_l \neq 0$ as at least one of the $\mathbf{X}^i$'s will be nonzero. ■

The following lemma immediately follows from Lemma C.16 by summing up over all $\mathbf{v}_l\mathbf{m}_l^T$ and $\mathbf{n}_l\mathbf{v}_l^T$ terms and using $\text{sum}(\mathbf{E}^L) \geq 0$. It summarizes the main result of the section.

Lemma C.17 *Let $\tilde{\mathbf{ED}}_{\mathscr{A}}$ be as defined in (C.1) and assume $\tilde{\mathbf{ED}}_{\mathscr{A}} > 0$. Then with probability $1 - 2nK\exp(-2\tilde{\mathbf{ED}}^2_{\mathscr{A}}(n_{min} - 1))$ we have $g(\mathbf{E}^L) > 0$ for all nonzero feasible $\mathbf{E}^L \in \mathscr{M}_U^{\perp}$.*

C.3.4 The Final Step

Lemma C.18 *Let $p_{min} > q$ and $\mathbf{G}$ be a random graph generated according to Model 6.1 with cluster sizes $\{n_i\}_{i=1}^K$. If $\lambda \leq (1-\varepsilon)\tilde{\Lambda}_{succ}$ and $\mathbf{ED}_{\min} = \min_{1\leq i\leq n} r(p_i - q)n_i \geq (1+\varepsilon)\frac{1}{\lambda}$, then $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal solution to Program 6.1 with probability $1 - \exp(-\Omega(n)) - 6n^2\exp(-\Omega(n_{min}))$.*

Proof: Based on Lemma C.15 and Lemma C.17,

with probability $1 - cn^2\exp(-C(p_{min} - q)^2 n_{min})$,

- There exists $\mathbf{W} \in \mathscr{M}_{\mathbf{U}}$ with $\|\mathbf{W}\| < 1$ such that for all feasible $\mathbf{E}^L$, $f(\mathbf{E}^L, \mathbf{W}) \geq 0$.
- For all nonzero $\mathbf{E}^L \in \mathscr{M}_{\mathbf{U}}^{\perp}$ we have $g(\mathbf{E}^L) > 0$.

Consequently based on Lemma C.1, $(\mathbf{L}^0, \mathbf{S}^0)$ is the unique optimal of Problem 6.3. ■