The Squared-Error of Generalized LASSO: A Precise Analysis

Samet Oymak, Christos Thrampoulidis and Babak Hassibi*

Department of Electrical Engineering
Caltech, Pasadena – 91125
soymak@caltech.edu, cthrampo@caltech.edu, hassibi@caltech.edu

Abstract

We consider the problem of estimating an unknown signal $\mathbf{x}_0$ from noisy linear observations $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$. In many practical instances of this problem, $\mathbf{x}_0$ has a certain structure that can be captured by a structure inducing function $f(\cdot)$. For example, ℓ_1 norm can be used to encourage a sparse solution. To estimate $\mathbf{x}_0$ with the aid of a convex $f(\cdot)$, we consider three variations of the widely used LASSO estimator and provide sharp characterizations of their performances. Our study falls under a generic framework, where the entries of the measurement matrix $\mathbf{A}$ and the noise vector $\mathbf{z}$ have zero-mean normal distributions with variances 1 and σ^2, respectively. For the LASSO estimator $\mathbf{x}^*$, we ask: "What is the precise estimation error as a function of the noise level σ, the number of observations m and the structure of the signal?". In particular, we attempt to calculate the Normalized Square Error (NSE) defined as $\frac{\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{\sigma^2}$. We show that, the structure of the signal $\mathbf{x}_0$ and choice of the function $f(\cdot)$ enter the error formulae through the summary parameters $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ and $\mathbf{D}_f(\mathbf{x}_0, \lambda)$, which are defined as the "Gaussian squared-distances" to the subdifferential cone and to the λ-scaled subdifferential of f at $\mathbf{x}_0$, respectively. The first estimator assumes a-priori knowledge of $f(\mathbf{x}_0)$ and is given by $\arg\min_{\mathbf{x}} \{\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$ subject to $f(\mathbf{x}) \le f(\mathbf{x}_0)\}$. We prove that its worst case NSE is achieved when $\sigma \to 0$ and concentrates around $\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}$. Secondly, we consider $\arg\min_{\mathbf{x}} \{\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda f(\mathbf{x})\}$, for some penalty parameter $\lambda \ge 0$. This time, the NSE formula depends on the choice of λ and is given by $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}$ over a range of λ. The last estimator is $\arg\min_{\mathbf{x}} \left\{\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \sigma\tau f(\mathbf{x})\right\}$. We establish a mapping between this and the second estimator and propose a formula for its NSE. As useful side results, we find explicit formulae for the optimal estimation performance and the optimal penalty parameters λ_{best} and τ_{best}. Finally, for a number of important structured signal classes, we translate our abstract formulae to closed-form upper bounds on the NSE.

Keywords: *convex optimization, generalized LASSO, structured sparsity, Gaussian processes, statistical estimation, duality, model fitting, linear inverse, first order approximation, noisy compressed sensing, random noise*

*This work was supported in part by the National Science Foundation under grants CCF-0729203, CNS-0932428 and CIF-1018927, by the Office of Naval Research under the MURI grant N00014-08-1-0747, and by a grant from Qualcomm Inc.

Contents

1. Introduction

1.1. The Generalized LASSO Problem

Recovering a structured signal $\mathbf{x}_0 \in \mathbb{R}^n$ from a vector of limited and noisy linear observations $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z} \in \mathbb{R}^m$, is a problem of fundamental importance encountered in several disciplines including machine learning, signal processing, network inference and many more [1–4]. A typical approach for estimating the structured signal $\mathbf{x}_0$ from the measurement vector $\mathbf{y}$, is picking some proper structure inducing function $f(\cdot)$ and solving the following problem

$$\mathbf{x}^*_{LASSO} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}) \right\}, \tag{1.1}$$

for some nonnegative penalty parameter λ.

In case $\mathbf{x}_0$ is a sparse vector, the associated structure inducing function is the ℓ_1 norm, i.e. $f(\mathbf{x}) = \|\mathbf{x}\|_1$. The resulting ℓ_1-penalized quadratic program in (1.1) is known as the LASSO in the statistics literature. LASSO was originally introduced in [5] and has since then been subject of great interest as a natural and powerful approach to do noise robust compressed sensing (CS), [5–15]. There are also closely related algorithms such as SOCP variations and the Dantzig selector [18, 19]. Of course, applications of (1.1) are not limited to sparse recovery; they extend to various problems including the recovery of block sparse signals [20, 21], the matrix completion problem [22, 23] and the total variation minimization [24]. In each application, $f(\cdot)$ is chosen in accordance to the structure of $\mathbf{x}_0$. See [25] for additional examples and a principled approach to constructing such penalty functions. In this work, we consider arbitrary convex penalty functions $f(\cdot)$ and we commonly refer to this generic formulation in (1.1) as the "Generalized LASSO" or simply "LASSO" problem.

1.2. 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.

1.2.1 Noiseless compressed sensing

In the noiseless CS problem one wishes to recover $\mathbf{x}_0$ from the random linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}_0$. A common approach is solving the following convex optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{1.2}$$

A critical performance criteria for the problem (1.2) concerns the minimum number of measurements needed to guarantee successful recovery of $\mathbf{x}_0$ [25–31]. Here, success means that $\mathbf{x}_0$ is the unique minimizer of (1.2), with high probability, over the realizations of the random matrix $\mathbf{A}$.

1.2.2 Proximal denoising

The proximal denoising problem tries to estimate $\mathbf{x}_0$ from noisy but uncompressed observations $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, where we write $\mathbf{I}_k$ for the identity matrix of size $k \times k$, $k \in \mathbb{Z}^+$. In particular, it solves,

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\sigma f(\mathbf{x}) \right\}. \tag{1.3}$$

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:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|_2^2 \quad \text{subject to} \quad f(\mathbf{x}) \leq f(\mathbf{x}_0). \tag{1.4}$$

The natural question to be posed in both cases is how well can one estimate $\mathbf{x}_0$ via (1.3) (or (1.4)) [40–44]? The minimizer $\mathbf{x}^*$ of (1.3) (or (1.4)) 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 $\frac{\mathbb{E}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{\sigma^2}$.

1.2.3 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 work: *proving precise bounds for the NSE of the Generalized LASSO problem.*

In the specific case of ℓ_1-penalization in (1.1), researchers have considered other performance criteria additional to the NSE [10–12]. As an example, we mention the support recovery criteria [10], which measures how well (1.1) 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 performance and is, thus, the sole focus in this work. In the relevant literature, researchers have dealt with the analysis of the NSE of (1.1) under several settings (see Section 1.4). Yet, *we still lack a general theory that would yield precise bounds for the squared-error of* (1.1) *for arbitrary convex regularizer* $f(\cdot)$*. This paper aims 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, [14, 15, 25, 31, 32, 41].

1.3. 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,

$$\mathbf{x}^*_c(\mathbf{A}, \mathbf{z}) = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \quad \text{subject to} \quad f(\mathbf{x}) \leq f(\mathbf{x}_0). \tag{1.5}$$

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

$$\mathbf{x}^*_{\ell_2}(\lambda, \mathbf{A}, \mathbf{z}) = \arg\min_{\mathbf{x}} \{ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda f(\mathbf{x}) \}. \tag{1.6}$$

- ⋆ ℓ_2^2**-LASSO**: the original form given in (1.1) :

$$\mathbf{x}^*_{\ell_2^2}(\tau, \mathbf{A}, \mathbf{z}) = \arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \sigma\tau f(\mathbf{x}) \right\}. \tag{1.7}$$

C-LASSO in (1.5) 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 (1.5)–(1.7), when clear from context.

1.4. 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 the paper. Table 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. [25] Amelunxen et al. [31]	Donoho and Tanner, [28] Stojnic, [26]
Proximal denoising	Donoho et al. [44] Oymak and Hassibi [41]	Donoho [40]
LASSO	Present paper	Bayati and Montanari, [14], [15] Stojnic, [37]

Table 1: Relevant Literature.

The works closest in spirit to our results include [14–16, 37], 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 [14, 15], 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 [16], Maleki et al. proposes Complex AMP, and characterizes the performance of LASSO for sparse signals with complex entries. In [37], Stojnic's approach relies on results on Gaussian processes [72,73] to derive sharp bounds for the *worst case NSE* of the ℓ_1-constrained LASSO problem in (1.5). Our approach in this work builds on the framework proposed by Stojnic, but extends the results in multiple directions as noted in the next section.

1.5. Contributions

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

- *generalizes* the results of [37] 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.

1.6. 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,

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

$\mathbf{A}^T\mathbf{A}$ 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 [70]). Hence,

$$\frac{\mathbb{E}\|\mathbf{x}^* - \mathbf{x}_0\|^2}{\sigma^2} = \frac{n}{m-n-1}. \tag{1.8}$$

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 (1.8) with a summary parameter $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ or $\mathbf{D}_f(\mathbf{x}_0, \lambda)$* . 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 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 (1.5). A more elaborate discussion, including estimation via ℓ_2-LASSO and ℓ_2^2-LASSO, can be found in Section 4.4. 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,

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \lesssim \frac{2k(\log \frac{n}{k} + 1)}{m - 2k(\log \frac{n}{k} + 1)}. \tag{1.9}$$

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) [50,51]. Hence, we observe $\mathbf{y} = \mathbf{A}\cdot\text{vec}(\mathbf{X}_0) + \mathbf{z}$ and solve,

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

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

$$\frac{\|\mathbf{X}_c^* - \mathbf{X}_0\|_F^2}{\sigma^2} \lesssim \frac{6dr}{m - 6dr}. \tag{1.10}$$

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, [46–49]. 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,

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|_2^2}{\sigma^2} \lesssim \frac{4k(\log \frac{t}{k} + b)}{m - 4k(\log \frac{t}{k} + b)}. \tag{1.11}$$

Note how (1.9)-(1.11) are similar in nature to (1.8).

2. 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. 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. Throughout, we use boldface lowercase letters to denote vectors and boldface capital letters to denote matrices. Also, to simplify the notation the ℓ_2-norm will be denoted as $\|\cdot\|$ from now on.

2.1. First-Order Approximation

Recall the ℓ_2-LASSO problem introduced in (1.6):

$$\mathbf{x}_{\ell_2}^* = \arg\min_{\mathbf{x}} \{\ \|\mathbf{y} - \mathbf{A}\mathbf{x}\| + \lambda f(\mathbf{x})\ \}. \tag{2.1}$$

A key idea behind our approach is using the linearization of the convex structure inducing function $f(\cdot)$ around the vector of interest $\mathbf{x}_0$ [77,86]:

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}_0) + \sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T(\mathbf{x} - \mathbf{x}_0). \tag{2.2}$$

$\partial f(\mathbf{x}_0)$ denotes the subdifferential of $f(\cdot)$ at $\mathbf{x}_0$ and is always a compact and convex set [86]. 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 $\|\mathbf{x}-\mathbf{x}_0\|$ is sufficiently small, then $\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$. We substitute $f(\cdot)$ in (2.1) 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 $\mathbf{w} = \mathbf{x} - \mathbf{x}_0$:

$$\hat{\mathbf{w}}_{\ell_2}(\lambda, \sigma, \mathbf{A}, \mathbf{v}) = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + \sup_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w} \right\}. \tag{2.3}$$

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 (2.3) and $\mathbf{w}^*_{\ell_2} = \mathbf{x}^*_{\ell_2} - \mathbf{x}_0$ for the optimal solution of the original problem in (2.1)[1]. Also, denote the optimal cost achieved in (2.2) by $\hat{\mathbf{w}}_{\ell_2}$, as $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$.

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 $\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\| \to 0$ 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.

2.2. Importance of $\sigma \to 0$

In this work, 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 (1.3) in [41, 44, 57] and, closer to the present paper, for the LASSO problem with ℓ_1 penalization (see Donoho et al. [62]).

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.

2.3. Gordon's Lemma

Perhaps the most important technical ingredient of the analysis presented in this work is a lemma proved by Gordon in [72]. Gordon's Lemma establishes a very useful (probabilistic) inequality for Gaussian processes.

Lemma 2.1 (Gordon [72]). *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 $\mathcal{S} \subset \mathbb{R}^n$ be an arbitrary set and $\psi : \mathcal{S} \to \mathbb{R}$ be an arbitrary function. Then, for any $c \in \mathbb{R}$,*

$$\mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}} \{\|\mathbf{G}\mathbf{x}\| + \|\mathbf{x}\|g - \psi(\mathbf{x})\} \geq c \right) \geq \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{S}} \left\{ \|\mathbf{x}\|\|\mathbf{g}\| - \mathbf{h}^T\mathbf{x} - \psi(\mathbf{x}) \right\} \geq c \right). \tag{2.4}$$

It is worth mentioning that the "escape through a mesh" lemma, which has been the backbone of the approach introduced by Stojnic [26] (and subsequently refined in [25]) for computing an asymptotic upper bound to the minimum number of measurements required in the Noiseless CS problem, is a corollary of Lemma 2.1

For the purposes of our analysis, we require a slight modification of this lemma. To avoid technicalities at this stage, we defer its precise statement to Section 5.3. Here, it suffices to observe that the original Gordon's Lemma 2.1 is (almost) directly applicable to the LASSO problem in (2.3). First, write $\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| = \max_{\|\mathbf{a}\|=1} \mathbf{a}^T[\mathbf{A}, -\mathbf{v}] \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix}$ and

[1] We follow this conventions throughout the paper: use the symbol "$\hat{}$" over variables that are associated with the approximated problems. To distinguish, use the symbol " * " for the variables associated with the original problem .

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 (2.4) takes the format of the LASSO problem in (2.3), except for the "distracting" factor $\|\mathbf{x}\|_g$. A simple argument shows that this term can be discarded without affecting the essence of the probabilistic statement of Lemma 2.1. Details being postponed to the later sections (cf. Section 5), Corollary 2.1 below summarizes the result of applying Gordon's Lemma to the LASSO problem.

Corollary 2.1 (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:*

$$\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})=\min_{\mathbf{w}}\left\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\|\mathbf{g}\|-\mathbf{h}^T\mathbf{w}+\max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)}\mathbf{s}^T\mathbf{w}\right\}. \tag{2.5}$$

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

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

Corollary 2.1 establishes a probabilistic connection between the LASSO problem and the minimization (2.5). 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, (2.5) 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 $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$. Due to its central role in our analysis, we often refer to problem (2.5) as "*key optimization*" or "*lower key optimization*". The term "lower" is attributed to the fact that analysis of (2.5) results in a probabilistic lower bound for the optimal cost of the LASSO problem.

2.4. Analyzing the Key Optimization

2.4.1 Deterministic Analysis

First, we perform the deterministic analysis of $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$ for fixed $\mathbf{g}\in\mathbb{R}^m$ and $\mathbf{h}\in\mathbb{R}^n$. In particular, we reduce the optimization in (2.5) to a *scalar* optimization. To see this, perform the optimization over a fixed ℓ_2-norm of $\mathbf{w}$ to equivalently write

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

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

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

This reduces (2.5) 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 2.2 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 $\mathcal{C}\in\mathbb{R}^n$ as

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

Lemma 2.2 (Deterministic Result)**.** *Let $\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})$ be a minimizer of the problem in (2.5). If $\|\mathbf{g}\|>dist(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$, then,*

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

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

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

2.4.2 Probabilistic Analysis

Of interest is making probabilistic statements about $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$ and the norm of its minimizer $\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|$. Lemma 2.2 provided closed form deterministic solutions for both of them, which only involve the quantities $\|\mathbf{g}\|^2$ and $\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$. For $\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_m)$ and $\mathbf{h}\sim\mathcal{N}(0,\mathbf{I}_n)$, standard results on Gaussian concentration show that, these quantities concentrate nicely around their means $\mathbb{E}\left[\|\mathbf{g}\|^2\right]=m$ and $\mathbb{E}\left[\text{dist}^2(\mathbf{h},\lambda\partial f(\mathbf{x}_0))\right]=:\mathbf{D}_f(\mathbf{x}_0,\lambda)$, respectively. Combining these arguments with Lemma 2.2, we conclude with Lemma 2.3 below.

Lemma 2.3 (Probabilistic Result)**.** *Assume that* $(1-\epsilon_L)m\geq\mathbf{D}_f(\mathbf{x}_0,\lambda)\geq\epsilon_L m$ *for some constant* $\epsilon_L>0$*. Define*[2]*,*

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

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

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

Remark: In Lemma 2.3, the condition "$(1-\epsilon_L)m\geq\mathbf{D}_f(\mathbf{x}_0,\lambda)$" ensures that $\|\mathbf{g}\|>\text{dist}(\mathbf{h},\lambda\partial f(\mathbf{x}_0))$ (cf. Lemma 2.2) with high probability over the realizations of $\mathbf{g}$ and $\mathbf{h}$.

2.5. Connecting back to the LASSO: The "Predictive Power of Gordon's Lemma"

Let us recap the last few steps of our approach. Application of Gordon's Lemma to the approximated LASSO problem in (2.3) introduced the simpler lower key optimization (2.5). Without much effort, we found in Lemma 2.3 that its cost $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$ and the normalized squared norm of its minimizer $\frac{\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|^2}{\sigma^2}$ concentrate around $\sigma\eta$ and γ, respectively. This brings the following question:

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

Application of Gordon's Lemma as performed in Corollary 2.1 when combined with Lemma 2.3, provide a preliminary answer to this question: $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ is lower bounded by $\sigma\eta$ with overwhelming probability. Formally,

Lemma 2.4 (Lower Bound)**.** *Assume* $(1-\epsilon_L)m\geq\mathbf{D}_f(\mathbf{x}_0,\lambda)\geq\epsilon_L m$ *for some constant* $\epsilon_L>0$ *and* m *is sufficiently large. Then, for any* $\epsilon>0$*, there exists a constant* $c>0$ *such that, with probability* $1-\exp(-cm)$*,*

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

But is that all? A major part of our technical analysis in the remainder of this work involves showing that the connection between the LASSO problem and the simple optimization (2.5) is much *deeper* than Lemma 2.4 predicts. In short, under certain conditions on λ and m (similar in nature to those involved in the assumption of Lemma 2.4), we prove that the followings are true:

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

In some sense, $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$ "predicts" $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})$ and $\|\hat{\mathbf{w}}(\mathbf{g},\mathbf{h})\|$ "predicts" $\|\hat{\mathbf{w}}_{\ell_2}(\mathbf{A},\mathbf{v})\|$, which attributes Gordon's Lemma (or more precisely to the lower key optimization) a "predictive power". This power is not necessarily restricted to the two examples above. In Section 10, 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 11 we rely on this predictive power of Gordon's Lemma 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 [37] 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.

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

2.6. Synopsis of the Technical Framework

We highlight the main steps of the technical framework.

1. Apply Gordon's Lemma to $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$ to find a *high-probability lower bound* for it. (cf. Lemma 2.4)
2. Apply Gordon's Lemma to the *dual* of $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$ to find a *high-probability upper bound* for it.
3. Both lower and upper bounds can be made arbitrarily close to $\sigma\eta$. Hence, $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$ concentrates with high probability around $\sigma\eta$ as well.
4. Assume $\frac{\|\hat{\mathbf{w}}_{\ell_2}\|^2}{\sigma^2}$ deviates from γ. A third application of Gordon's Lemma shows that such a deviation would result in a *significant increase* in the optimal cost, namely $\hat{\mathcal{F}}_{\ell_2}(\mathbf{A}, \mathbf{v})$ would be significantly larger than $\sigma\eta$.
5. From the previous step, conclude that $\frac{\|\hat{\mathbf{w}}_{\ell_2}\|^2}{\sigma^2}$ concentrates with high probability around γ.

2.7. Gaussian Squared Distance and Related Quantities

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

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) := \mathbb{E}\left[\text{dist}^2(\mathbf{h}, \lambda\partial f(\mathbf{x}_0))\right], \tag{2.6}$$

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 $\mathcal{C} \in \mathbb{R}^n$, denote its conic hull by $\text{cone}(\mathcal{C})$. Also, denote its polar cone by $\mathcal{C}^\circ$, which is the closed and convex set $\{\mathbf{u} \in \mathbb{R}^n \big| \mathbf{u}^T\mathbf{v} \leq 0 \text{ for all } \mathbf{v} \in \mathcal{C}\}$.

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

$$\mathbf{C}_f(\mathbf{x}_0, \lambda) := \mathbb{E}\left[(\mathbf{h} - \text{Proj}(\mathbf{h}, \lambda\partial f(\mathbf{x}_0)))^T \text{Proj}(\mathbf{h}, \lambda\partial f(\mathbf{x}_0))\right], \tag{2.7}$$

$$\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) := \mathbb{E}\left[\text{dist}^2(\mathbf{h}, \text{cone}(\partial f(\mathbf{x}_0)))\right]. \tag{2.8}$$

From the previous discussion, it has become clear how $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ appears in the analysis of the NSE of the ℓ_2-LASSO. $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ replaces $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ 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 (2.3) by replacing $\lambda\partial f(\mathbf{x}_0)$ with $\text{cone}(\partial f(\mathbf{x}_0))$. While $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ is the only quantity that appears in the analysis of the C-LASSO, the analysis of the ℓ_2-LASSO requires considering not only $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ but also $\mathbf{C}_f(\mathbf{x}_0, \lambda)$. $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ appears in the analysis during the second step of the framework described in Section 2.6. In fact, $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ is closely related to $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ as the following lemma shows.

Lemma 2.5 ([31]). *Suppose $\partial f(\mathbf{x}_0)$ is nonempty and does not contain the origin. Then,*

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

As a last remark, the quantities $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ and $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ 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 [31][3] that the noiseless compressed sensing problem (1.2) exhibits a transition from "failure" to "success" around $m \approx \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$. Also, [41, 43, 44] shows that $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ are equal to the worst case normalized mean-squared-error of the proximal denoisers (1.3) and (1.4) respectively. It is known that under mild assumptions, $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ relates to $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ as follows [31, 32, 41],

$$\min_{\lambda \geq 0} \mathbf{D}_f(\mathbf{x}_0, \lambda) \approx \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+). \tag{2.9}$$

[3] The authors in [31] coined the term "statistical dimension" of a cone $\mathcal{K}$ to denote the expected squared distance of a gaussian vector to its polar cone $\mathcal{K}^\circ$. In that terminology, $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ is the statistical dimension of the $(\text{cone}(\partial f(\mathbf{x}_0)))^\circ$, or equivalently (see Lemma 7.2) of the descent cone of $f(\cdot)$ at $\mathbf{x}_0$.

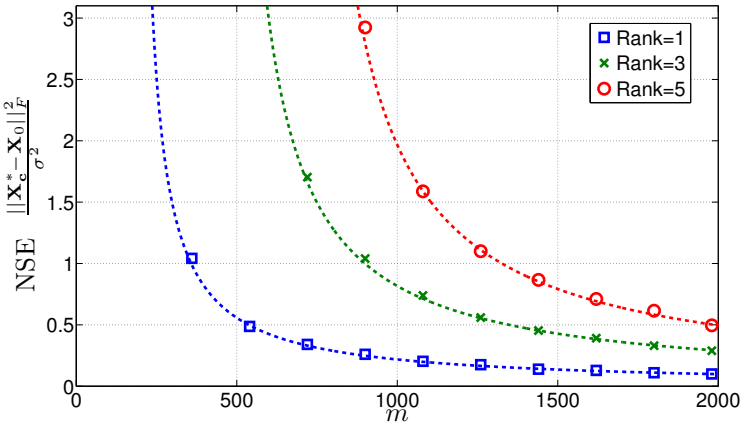


Figure 1: We have considered the Constrained-LASSO with nuclear norm minimization and fixed the signal to noise ratio $\frac{\|\mathbf{X}_0\|_F^2}{\sigma^2}$ to 10^5. Size of the underlying matrices are 40×40 and their ranks are $1, 3$ and 5. Based on [55,57], we estimate $\mathbf{D}_f(\mathbf{X}_0, \mathbb{R}^+) \approx 179, 450$ 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. Main Results

This section provides the formal statements of our main results. A more elaborate discussion follows in Section 4.

3.1. 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 (1.5), (1.6) and (1.7). 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 $\mathbf{D}_f(\mathbf{x}_0, \lambda), \mathbf{C}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ in (2.6), (2.7) and (2.8), respectively.

3.2. C-LASSO

Theorem 3.1 (NSE of C-LASSO). *Assume there exists a constant $\epsilon_L > 0$ such that, $(1-\epsilon_L)m \geq \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) \geq \epsilon_L m$ and m is sufficiently large. For any $\epsilon > 0$, there exists a constant $C = C(\epsilon, \epsilon_L) > 0$ such that, with probability $1 - \exp(-Cm)$,*

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|^2}{\sigma^2} \leq (1+\epsilon) \frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}, \tag{3.1}$$

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,

$$\left| \frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|^2}{\sigma^2} \times \frac{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)} - 1 \right| < \epsilon. \tag{3.2}$$

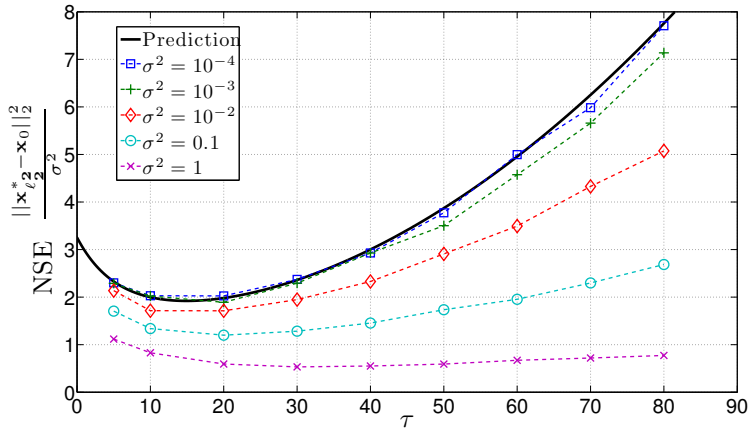


Figure 2: We considered ℓ_2^2-LASSO problem, for a k sparse signal of size $n = 1000$. We let $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$ and normalize the signal power by setting $\|\mathbf{x}_0\| = 1$. τ is varied from 0 to 80 and the signal-to-noise ratio (SNR) $\frac{\|\mathbf{x}_0\|^2}{\sigma^2}$ 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.

3.3. ℓ_2-LASSO

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

$$\mathcal{R}_{ON} = \left\{ \lambda > 0 \mid m - \mathbf{D}_f(\mathbf{x}_0, \lambda) > \max\{0, \mathbf{C}_f(\mathbf{x}_0, \lambda)\} \right\}.$$

Remark: *Section 8 fully characterizes $\mathcal{R}_{ON}$ and shows that it is an open interval.*

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

$$\left| \frac{\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|^2}{\sigma^2} \times \frac{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}{\mathbf{D}_f(\mathbf{x}_0, \lambda)} - 1 \right| < \epsilon. \tag{3.3}$$

3.4. ℓ_2^2-LASSO

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

$$map(\lambda) = \lambda \, \frac{m - \mathbf{D}_f(\mathbf{x}_0, \lambda) - \mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}. \tag{3.4}$$

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

Formula 1 (Conjecture on the NSE of ℓ_2^2-LASSO). *Assume $(1 - \epsilon_L)m \geq \min_{\lambda \geq 0} \mathbf{D}_f(\mathbf{x}_0, \lambda) \geq \epsilon_L m$ for a constant $\epsilon_L > 0$ 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 $\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|^2}{\sigma^2}$ for sufficiently small σ. Furthermore, we believe that the same expression upper bounds the NSE for arbitrary values of σ.

3.5. 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 $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$.*

Remark: Any norm in $\mathbb{R}^n$ is Lipschitz continuous [79].

Theorem 3.4 (Failure of Robust Recovery)**.** *Let $f(\cdot)$ be a Lipschitz continuous convex function Assume $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$. 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 $1 - 8\exp(-\frac{(\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)-m)^2}{4n})$, we have,*

$$\frac{\|\mathbf{x}^*_{\ell_2}(\mathbf{A}, \mathbf{z}) - \mathbf{x}_0\|^2}{\sigma^2} \geq C_{max}, \quad \text{and} \quad \frac{\|\mathbf{x}^*_{\ell_2^2}(\mathbf{A}, \mathbf{z}) - \mathbf{x}_0\|^2}{\sigma^2} \geq C_{max}. \tag{3.5}$$

3.6. Remarks

A detailed discussion of the results follows in Section 4. Before this, the following remarks are in place.

- Known results in the noiseless CS problem (1.2) 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 $\lambda \in \mathcal{R}_{\text{ON}}$. In Section 4 we elaborate on the behavior of the NSE for other values of the penalty parameter. Yet, the set of values $\mathcal{R}_{\text{ON}}$ 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 $\mathcal{R}_{\text{ON}}$.

(b) The function map$(\cdot)$ defined in Definition 3.2 proposes a bijective mapping from $\mathcal{R}_{\text{ON}}$ 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 $\mathcal{R}_{\text{ON}}$. Following this mapping, the exact characterization of the NSE of the ℓ_2-LASSO for $\lambda \in \mathcal{R}_{\text{ON}}$, 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 11. Section 11 also shows that, when $m > \min_{\lambda \geq 0} \mathbf{D}_f(\mathbf{x}_0, \lambda)$, 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 4. 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 [31]. A result of similar nature holds for the C-LASSO, as well. For the exact statement of this result and the proofs see Section 12.

3.7. Paper Organization

Section 4 contains a detailed discussion on our results and on their interpretation. Sections 5 and 6 contain the technical details of the framework as it was summarized in Section 2.6. In Sections 7 and 10, we prove the two parts of Theorem 3.1 on the NSE of the C-LASSO. Section 8 analyzes the ℓ_2-LASSO and Section 9 proves Theorem 3.2 regarding the NSE over $\mathcal{R}_{\text{ON}}$. Section 11 discusses the mapping between ℓ_2 and ℓ_2^2-LASSO, proves Theorem 3.3 and motivates Formula 1. In Section 12 we focus on the regime where robust estimation fails and prove Theorem 3.4.

Simulation results presented in Section 13 support our analytical predictions. Finally, directions for future work are discussed in Section 14. Some of the technical details are deferred to the Appendix.

4. Discussion of the Results

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

4.1. C-LASSO

We are able to characterize the estimation performance of the Constrained-LASSO in (1.5) solely based on $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$. Whenever $m > \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$, for sufficiently small σ, we prove that,

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|^2}{\sigma^2} \approx \frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}. \tag{4.1}$$

Furthermore, (4.1) holds for arbitrary values of σ when $\approx$ is replaced with $\lesssim$. Observe in (4.1) that as m approaches $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$, the NSE increases and when $m = \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$, NSE $= \infty$. This behavior is not surprising as when $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$, one cannot even recover $\mathbf{x}_0$ from noiseless observations via (1.2) hence it is futile to expect noise robustness. For purposes of illustration, notice that (4.1) can be further simplified for certain regimes as follows:

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

4.1.1 Relation to Proximal Denoising

We want to compare the NSE of the C-LASSO in (1.5) to the MSE risk of the constrained proximal denoiser in (1.4). For a fair comparison, the average signal power $\mathbb{E}[\|\mathbf{A}\mathbf{x}_0\|^2]$ in (1.5) should be equal to $\|\mathbf{x}_0\|^2$. 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 $\|\mathbf{x}_c^* - \mathbf{x}_0\|^2$ by σ^2. Thus, in this case, the formula (4.1) for the NSE is multiplied by m to result in $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) \cdot \frac{m}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}$ (see Section 4.5 for further explanation). Now, let us compare this with the results known for proximal denoising. There [41, 43], it is known that the normalized MSE is maximized when $\sigma \to 0$ and is equal to $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$. 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 $\frac{m}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)} > 1$. This factor can be interpreted as the penalty paid in the estimation error for using linear measurements.

4.2. 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.

4.2.1 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 (1.2), which is the case when $m > \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ [25, 31]. To translate this requirement in terms of $\mathbf{D}_f(\mathbf{x}_0, \lambda)$, recall (2.9) and Lemma 2.5, and define λ_{best} to be the *unique* minimizer of $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ over $\lambda \in \mathbb{R}^+$. We, then, write the regime of interest as $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$.

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.

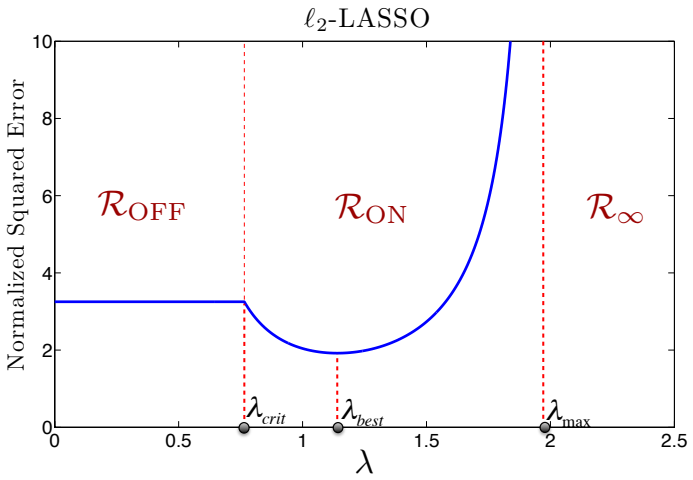


Figure 3: We consider the ℓ_1-penalized ℓ_2-LASSO problem for a k sparse signal in $\mathbb{R}^n$. x-axis is the penalty parameter λ. For $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$, we have $\lambda_{\text{crit}} \approx 0.76$, $\lambda_{\text{best}} \approx 1.14$, $\lambda_{\max} \approx 1.97$.

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

c) λ_{crit} : Over $0 \leq \lambda \leq \lambda_{\text{best}}$, if $m \leq n$, the equation $m - \mathbf{D}_f(\mathbf{x}_0, \lambda) = \mathbf{C}_f(\mathbf{x}_0, \lambda)$ 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) $\mathcal{R}_{\text{ON}} = \{\lambda \in \mathbb{R}^+ | \lambda_{\text{crit}} < \lambda < \lambda_{\max}\}$.

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

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

See Figure 4 for an illustration of the definitions above and Section 8 for the detailed proofs of the statements.

4.2.2 Characterizing the NSE in each Region

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

- $\mathcal{R}_{\text{OFF}}$: For $\lambda \in \mathcal{R}_{\text{OFF}}$, we empirically observe that the LASSO estimate $\mathbf{x}^*_{\ell_2}$ satisfies $\mathbf{y} = \mathbf{A}\mathbf{x}^*_{\ell_2}$ and the optimization (1.6) reduces to:

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

  which is the standard approach to solving the noiseless linear inverse problems (recall (1.2)). We prove that this reduction is indeed true for values of λ sufficiently small (see Lemma 9.2), while our empirical observations suggest that the claim is valid for all $\lambda \in \mathcal{R}_{\text{OFF}}$. 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 $\lambda \in \mathcal{R}_{\text{OFF}}$. Interestingly, this would also give the NSE formula for the particularly interesting problem (4.2). Simulation results in Section 13 validate the claim.

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

$$\frac{\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|}{\sigma^2} \approx \frac{\mathbf{D}_f(\mathbf{x}_0, \lambda)}{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}. \tag{4.2}$$

Also, empirical observations suggest that 4.2 holds for arbitrary σ when $\approx$ replaced with $\lesssim$. Finally, we should note that the NSE formula $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}$ is a convex function of λ over $\mathcal{R}_{\text{ON}}$.

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

4.2.3 Optimal Tuning of the Penalty Parameter

It is not hard to see that the formula in (4.2) is strictly increasing in $\mathbf{D}_f(\mathbf{x}_0, \lambda)$. Thus, when $\sigma \to 0$, the NSE achieves its minimum value when the penalty parameter is set to λ_{best}. Now, recall that $\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ and compare the formulae in (4.1) and (4.2), to conclude that the C-LASSO and ℓ_2-LASSO can be related by choosing $\lambda = \lambda_{\text{best}}$. In particular, we have,

$$\frac{\|\mathbf{x}^*_{\ell_2}(\lambda_{\text{best}}) - \mathbf{x}_0\|^2}{\sigma^2} \approx \frac{\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})}{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})} \approx \frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)} \approx \frac{\|\mathbf{x}^*_c - \mathbf{x}_0\|^2}{\sigma^2}. \tag{4.3}$$

4.3. ℓ_2^2-LASSO

4.3.1 Connection to ℓ_2-LASSO

We propose a mapping between the penalty parameters λ of the ℓ_2-LASSO program (1.6) and τ of the ℓ_2^2-LASSO program (1.7), 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 $\mathcal{R}_{\text{ON}}$, since $m > \mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $m - \mathbf{D}_f(\mathbf{x}_0, \lambda) > \mathbf{C}_f(\mathbf{x}_0, \lambda)$ for all $\lambda \in \mathcal{R}_{\text{ON}}$. Theorem 3.3 proves that $\text{map}(\cdot)$ defines a bijective mapping from $\mathcal{R}_{\text{ON}}$ 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 11 proves these properties and more, and contains a short technical discussion that motivates the proposed mapping function.

4.3.2 Proposed Formula

We use the mapping function in (3.4) to translate our results on the NSE of the ℓ_2-LASSO over $\mathcal{R}_{\text{ON}}$ (see formula (4.2)) 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 $\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|^2}{\sigma^2}$ for sufficiently small σ, and upper bounds $\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|^2}{\sigma^2}$ for arbitrary σ.

4.3.3 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 $\text{map}(\lambda_{\text{best}})$ we make use of Lemma 2.5 and the fact that $\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda} = -\frac{2}{\lambda}\mathbf{C}_f(\mathbf{x}_0, \lambda)$ for all $\lambda \geq 0$. Combine this with the fact that λ_{best} is the unique minimizer of $\mathbf{D}_f(\mathbf{x}_0, \lambda)$, to show that $\mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{best}}) = 0$, and to conclude with,

$$\tau_{\text{best}} = \lambda_{\text{best}} \sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}})}. \tag{4.4}$$

As a last comment, (4.4) simplifies even further if one uses the fact $\mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{best}}) \approx \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$, which is valid under reasonable assumptions, [31, 32, 41]. In this case, $\tau_{best} \approx \lambda_{\text{best}} \sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}$.

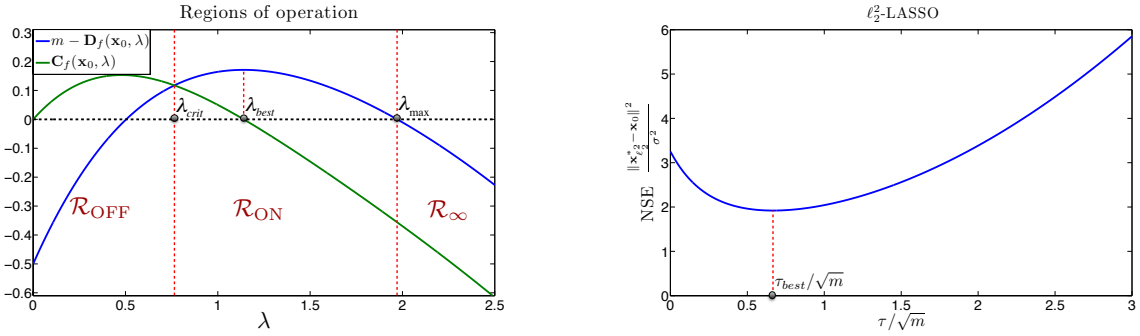


Figure 4: We consider the exact same setup of Figure 3. a) We plot $m - \mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ as a function of λ to illustrate the important penalty parameters $\lambda_{\text{crit}}, \lambda_{\text{best}}, \lambda_{\text{max}}$ and the regions of operation $\mathcal{R}_{\text{OFF}}, \mathcal{R}_{\text{ON}}, \mathcal{R}_{\infty}$. 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 $\sqrt{m}$.

4.4. Closed Form Calculations of the Formulae

	Normalized Squared Error
C-LASSO	$\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}$
ℓ_2-LASSO	$\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}$ for $\lambda \in \mathcal{R}_{\text{ON}}$
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 2: Summary of formulae for the NSE.

Table 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 $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ and $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ ($\mathbf{C}_f(\mathbf{x}_0, \lambda)$ 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 $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) \lesssim 2k(\log \frac{n}{k} + 1)$. Substituting this into the formula for the NSE of the C-LASSO results in the "closed-form" upper bound given in (1.9), i.e. one expressed only in terms of m,n and k. Analogous results have been derived [25,32,48,55] for other well-known signal models as well, including low rankness (see (1.10)) and block-sparsity (see (1.11)). The first row of Table 3 summarizes some of the results for $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ found in the literature (see [25,32]). The second row provides our closed form results on $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ 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 $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and closed form bounds, the reader is referred to [32]. The derivation of these results can be found in Section H of the Appendix. In the same section, we also provide exact formulas for $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ for the same signal models. Based on those formulas and Table 3, one simply needs to substitute $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ or $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ 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 [25]).

It follows from this discussion, that establishing new and tighter analytic bounds for $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ for more regularizers f is certainly an interesting direction for future research. In the case where such analytic bounds do

	k-**sparse,** $\mathbf{x}_0 \in \mathbb{R}^n$	**Rank** r**,** $\mathbf{X}_0 \in \mathbb{R}^{d\times d}$	k-**block sparse,** $\mathbf{x}_0 \in \mathbb{R}^{tb}$
$\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$	$2k(\log \frac{n}{k} + 1)$	$6dr$	$4k(\log \frac{t}{k} + b)$
$\mathbf{D}_f(\mathbf{x}_0, \lambda)$	$(\lambda^2+3)k$ 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: Closed form upper bounds for $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ ([25,32]) and $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ corresponding to (1.9), (1.10) and (1.11).

not already exist in literature or are hard to derive, one can numerically estimate $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ 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 $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ (see Lemma B.3) . Hence to compute $\mathbf{D}_f(\mathbf{x}_0, \lambda)$:

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

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

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

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

4.5. 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 $\frac{1}{n}$, [14,15,17]. For example, a variance of $\frac{1}{m}$ ensures that in expectation $\|\mathbf{A}\mathbf{x}\|^2$ is same as $\|\mathbf{x}\|^2$. 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,

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

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

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

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 $\sqrt{m}\lambda' \in \mathcal{R}_{\text{ON}}$ can be given as,

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

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

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

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

5. Applying Gordon's Lemma

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

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

5.1. 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, $\|\mathbf{y} - \mathbf{A}\mathbf{x}\| = \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|$. 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*:

$$f_p(\mathbf{w}) = f(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0), \tag{5.1}$$

$$\hat{f}_p(\mathbf{w}) = \hat{f}(\mathbf{x}_0 + \mathbf{w}) - f(\mathbf{x}_0) = \sup_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \mathbf{s}^T\mathbf{w}. \tag{5.2}$$

Then, the ℓ_2-LASSO will write as

$$\mathbf{w}^*_{\ell_2} = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + \lambda f_p(\mathbf{w}) \right\}. \tag{5.3}$$

and the C-LASSO as

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

or, equivalently,

$$\mathbf{w}^*_c = \arg\min_{\mathbf{w}} \left\{ \ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + \max_{\lambda \geq 0} \lambda f_p(\mathbf{w}) \ \right\}. \tag{5.4}$$

5.2. The Approximate LASSO Problem

In Section 2, and in particular in (2.3) 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:

$$\hat{\mathbf{w}}_{\ell_2} = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + \lambda \hat{f}_p(\mathbf{w}) \right\}. \tag{5.5}$$

Similarly, the approximated C-LASSO writes

$$\hat{\mathbf{w}}_c \ = \ \arg\min_{\mathbf{w}} \left\{ \ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + \max_{\lambda \geq 0} \lambda \hat{f}_p(\mathbf{w}) \ \right\}. \tag{5.6}$$

Denote $\hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v})$ and $\hat{\mathcal{F}}_{\ell 2}(\mathbf{A}, \mathbf{v})$ the optimal costs of problems (5.6) and (5.5), 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.

5.3. Technical Tool: Gordon's Lemma

As already noted the most important technical ingredient underlying our analysis is a Lemma proved by Gordon in [72]; recall Lemma 2.1 in Section 2. In fact, Gordon's key Lemma 2.1 is a Corollary of a more general theorem which establishes a probabilistic comparison between two centered Gaussian processes. The theorem was proved by Gordon in [73] and is stated below for completeness.

Theorem 5.1 (Gordon's Theorem, [72]). *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. $\mathbb{E}\left[X_{ij}^2\right] = \mathbb{E}\left[Y_{ij}^2\right]$,
2. $\mathbb{E}\left[X_{ij}X_{ik}\right] \geq \mathbb{E}\left[Y_{ij}Y_{ik}\right]$,
3. $\mathbb{E}\left[X_{ij}X_{\ell k}\right] \leq \mathbb{E}\left[Y_{ij}Y_{\ell k}\right], \quad$ *if* $i \neq \ell$.

Then,

$$\mathbb{P}\left(\cap_i \cup_j \left[Y_{ij} \geq \lambda_{ij}\right]\right) \geq \mathbb{P}\left(\cap_i \cup_j \left[X_{ij} \geq \lambda_{ij}\right]\right),$$

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

Application of Gordon's Theorem 5.1 to specific Gaussian processes results in Gordon's Lemma 2.1 [72]. In this work, we require a slightly modified version of this lemma, namely Lemma 5.1. The key idea is of course the same as in the original lemma, but the statement is modified to fit the setup of the current paper.

Lemma 5.1 (Modified Gordon's Lemma)**.** *Let* $\mathbf{G}$, $\mathbf{g}$, $\mathbf{h}$ *be defined as in Lemma 2.1 and let* $\psi(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *. Also, let* $\Phi_1 \subset \mathbb{R}^n$ *and* $\Phi_2 \subset \mathbb{R}^m$ *such that either both* Φ_1 *and* Φ_2 *are compact or* Φ_1 *is arbitrary and* Φ_2 *is a scaled unit sphere. Then, for any* $c \in \mathbb{R}$*:*

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

The proof of Lemma 5.1 closely parallels the proof of Lemma 5.1 in [72]. We defer the proof to Section C in the Appendix.

5.4. Simplifying the LASSO objective through Gordon's Lemma

Section 2.6 introduced the technical framework. Key feature in this framework is the application of Gordon's Lemma. In particular, we apply Gordon's Lemma 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 Lemma and we identify the corresponding key optimizations. Later, in Section 6, 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:

$$\mathcal{F}(\mathbf{A},\mathbf{v}) = \min_{\mathbf{w}} \{\ \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + p(\mathbf{w})\ \}, \tag{5.7}$$

where $p : \mathbf{R}^n \to \mathbf{R} \cup \infty$ is a proper convex function [86]. Choose the penalty function $p(\cdot)$ in the generic formulation (5.7) accordingly to end up with (5.3), (5.4), (5.5) or (5.6). To retrieve (5.4) and (5.6), choose $p(\mathbf{w})$ as the indicator function of the sets $\{\mathbf{w}|f_p(\mathbf{w}) \leq 0\}$ and $\left\{\mathbf{w}|\hat{f}_p(\mathbf{w}) \leq 0\right\}$ [83].

5.4.1 Lower Bound

The following corollary is a direct application of Lemma 5.1 to $\mathcal{F}(\mathbf{A},\mathbf{v})$ in (5.7).

Corollary 5.1. *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*

$$\mathcal{L}(\mathbf{g},\mathbf{h}) = \min_{\mathbf{w}}\left\{\sqrt{\|\mathbf{w}\|^2+\sigma^2}\|\mathbf{g}\| - \mathbf{h}^T\mathbf{w} + p(\mathbf{w})\right\}. \tag{5.8}$$

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

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

Proof. Notice that $\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| = \|\mathbf{A}_\mathbf{v}\mathbf{w}_\sigma\|$, 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 the modified Gordon's Lemma 5.1, with $\mathbf{x} = \mathbf{w}_\sigma$, $\Phi_1 = \{\mathbf{w}_\sigma|\mathbf{w} \in \mathbb{R}^n\}$, $\Phi_2 = \mathcal{S}^{m-1}$, $\mathbf{G} = \mathbf{A}_\mathbf{v}$, $\psi(\mathbf{w}_\sigma) = p(\mathbf{w})$. Further perform the trivial optimizations over $\mathbf{a}$ on both sides of the inequality. Namely, $\max_{\|\mathbf{a}\|=1}\mathbf{a}^T\mathbf{A}_\mathbf{z}[\mathbf{w}^T\ \sigma]^T = \|\mathbf{A}_\mathbf{z}\mathbf{w}_\sigma\|$ and, $\max_{\|\mathbf{a}\|=1}\mathbf{g}^T\mathbf{a} = \|\mathbf{g}\|$. □

5.4.2 Upper Bound

Similar to the lower bound derived in the previous section, we derive an upper bound for $\mathcal{F}(\mathbf{A},\mathbf{v})$. For this, we need to apply Gordon's Lemma to $-\mathcal{F}(\mathbf{A},\mathbf{v})$ and use the dual formulation of it. Lemma D in the Appendix shows that the dual of the minimization in (5.7) can be written as

$$-\mathcal{F}(\mathbf{A},\mathbf{v}) = \min_{\|\boldsymbol{\mu}\|\leq 1} \max_{\mathbf{w}} \left\{ \boldsymbol{\mu}^T (\mathbf{A}\mathbf{w} - \sigma\mathbf{v}) - p(\mathbf{w}) \right\}. \tag{5.9}$$

Lemma 5.1 requires the set over which maximization is performed to be compact. We thus apply Lemma 5.1 to the restricted problem,

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

Notice, that this still gives a valid lower bound to $-\mathcal{F}(\mathbf{A},\mathbf{v})$ since the optimal cost of this latter problem is no larger than $-\mathcal{F}(\mathbf{A},\mathbf{v})$. In Section 6, we will choose C_{up} so that the resulting lower bound is as tight as possible.

Corollary 5.2. *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,*

$$\mathcal{U}(\mathbf{g},\mathbf{h}) = -\min_{\|\boldsymbol{\mu}\|\leq 1} \max_{\|\mathbf{w}\|\leq C_{up}} \left\{ \sqrt{\|\mathbf{w}\|^2+\sigma^2}\, \mathbf{g}^T\boldsymbol{\mu} + \|\boldsymbol{\mu}\|\mathbf{h}^T\mathbf{w} - p(\mathbf{w}) \right\}. \tag{5.10}$$

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

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

Proof. Similar to the proof of Corollary 5.1 write $\|\sigma\mathbf{v}-\mathbf{A}\mathbf{w}\| = \|\mathbf{A}_{\mathbf{v}}\mathbf{w}_\sigma\|$. Then, apply the modified Gordon's Lemma 5.1, with $\mathbf{x}=\boldsymbol{\mu}$, $\alpha=\mathbf{w}_\sigma$, $\Phi_1=\mathcal{B}^{m-1}$, $\Phi_2=\left\{\mathbf{w}_\sigma \mid \frac{1}{C_{up}}\mathbf{w}\in\mathcal{B}^{n-1}\right\}$, $\mathbf{G}=\mathbf{A}_{\mathbf{v}}$, $\psi(\mathbf{w}_\sigma)=p(\mathbf{w})$, to find that for any $c\in\mathbb{R}$:

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

□

5.4.3 Deviation Analysis

Of interest in the deviation analysis of the LASSO problem (cf. Step 4 in Section 2.6) is the analysis of a restricted version of the *LASSO* problem, namely

$$\min_{\|\mathbf{w}\|\in S_{dev}} \{\|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\| + p(\mathbf{w})\} \tag{5.11}$$

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 (5.11). As usual, we apply Lemma 5.1 to our setup, to conclude the following.

Corollary 5.3. *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*

$$\mathcal{L}_{dev}(\mathbf{g},\mathbf{h}) = \min_{\|\mathbf{w}\|\in S_{dev}} \left\{ \sqrt{\|\mathbf{w}\|^2+\sigma^2}\|\mathbf{g}\| - \mathbf{h}^T\mathbf{w} + p(\mathbf{w}) \right\}. \tag{5.12}$$

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

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

Proof. Follows from Lemma 5.1 following exactly the same steps as in the proof of Corollary 5.1. □

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

5.4.4 Summary

We summarize the results of Corollaries 5.1, 5.2 and 5.3 in Lemma 5.2. Adding to a simple summary, we perform a further simplification of the corresponding statements. In particular, we discard the "distracting" term σh in Corollaries 5.1 and 5.3, as well as the term $\min_{0\leq\alpha\leq 1}\alpha\sigma h$ in Corollary 5.2. Recall the definitions of the key optimizations $\mathcal{L}$, $\mathcal{U}$ and $\mathcal{L}_{dev}$ in (5.8), (5.10) and (5.12).

Lemma 5.2. *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 $\epsilon>0$, the following are true:*

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

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

$$\mathbb{P}\left(|h|\leq\epsilon\sqrt{m}\right)\geq 1-2\exp(-\frac{\epsilon^2 m}{2}). \tag{5.13}$$

Thus,

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

Combine this with Corollary 5.1 to conclude with the first statement of Lemma 5.2. The proof of the third statement of the Lemma follows the exact same steps applied this time to Corollary 5.3. For the second statement write,

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

and use (5.13) as above. To conclude, combine with the statement of Corollary 5.2. □

6. After Gordon's Lemma: Analyzing the Key Optimizations

6.1. 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 (5.7) 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 (5.8), (5.10) and (5.12), we will be analyzing the following *key optimizations*,

$$\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})=\min_{\mathbf{w}}\left\{\sqrt{\|\mathbf{w}\|^2+\sigma^2}\|\mathbf{g}\|-\mathbf{h}^T\mathbf{w}+\max_{\mathbf{s}\in\mathcal{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{6.1a}$$

$$\hat{\mathcal{U}}(\mathbf{g},\mathbf{h})=-\min_{\|\boldsymbol{\mu}\|\leq 1}\max_{\|\mathbf{w}\|=C_{up}}\left\{\sqrt{\|\mathbf{w}\|^2+\sigma^2}\,\mathbf{g}^T\boldsymbol{\mu}+\|\boldsymbol{\mu}\|\mathbf{h}^T\mathbf{w}-\max_{\mathbf{s}\in\mathcal{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{6.1b}$$

$$\hat{\mathcal{L}}_{dev}(\mathbf{g},\mathbf{h})=\min_{\|\mathbf{w}\|\in S_{dev}}\left\{\sqrt{\|\mathbf{w}\|^2+\sigma^2}\|\mathbf{g}\|-\mathbf{h}^T\mathbf{w}+\max_{\mathbf{s}\in\mathcal{C}}\mathbf{s}^T\mathbf{w}\right\}, \tag{6.1c}$$

where $\mathcal{C}$ 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 (6.1b) we have constrained the feasible set of the inner maximization to the scaled sphere rather than ball. Following our discussion, in Section 5.4.2 this does not affect the validity of Lemma 5.2, while it facilitates our derivations here.

To be consistent with the definitions in (6.1), which treat the key optimizations of the C-LASSO and ℓ_2-LASSO under a common framework with introducing a generic set $\mathcal{C}$, we also define

$$\hat{\mathcal{F}}(\mathbf{A},\mathbf{v}) = \min_{\mathbf{w}} \left\{ \|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\| + \max_{\mathbf{s}\in\mathcal{C}} \mathbf{s}^T\mathbf{w} \right\}, \tag{6.2}$$

to correspond to (5.6) and (5.5), when setting $\mathcal{C} = \text{cone}(\partial f(\mathbf{x}_0))$ and $\mathcal{C} = \lambda\partial f(\mathbf{x}_0)$, respectively.

6.2. Some Notation

Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed and nonempty convex set. For any vector $\mathbf{x} \in \mathbb{R}^n$, we denote its (unique) projection onto $\mathcal{C}$ as $\text{Proj}(\mathbf{x},\mathcal{C})$, i.e.

$$\text{Proj}(\mathbf{x},\mathcal{C}) := \text{argmin}_{\mathbf{s}\in\mathcal{C}} \|\mathbf{x}-\mathbf{s}\|.$$

It will also be convenient to denote,

$$\Pi(\mathbf{x},\mathcal{C}) := \mathbf{x} - \text{Proj}(\mathbf{x},\mathcal{C}).$$

The distance of $\mathbf{x}$ to the set $\mathcal{C}$ can then be written as,

$$\text{dist}(\mathbf{x},\mathcal{C}) := \|\Pi(\mathbf{x},\mathcal{C})\|.$$

Finally, we denote,

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

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

$$\mathbf{D}(\mathcal{C}) := \mathbb{E}\left[\text{dist}^2(\mathbf{h},\mathcal{C}) \right], \tag{6.3a}$$

$$\mathbf{P}(\mathcal{C}) := \mathbb{E}\left[\|\text{Proj}(\mathbf{h},\mathcal{C})\|^2 \right], \tag{6.3b}$$

$$\mathbf{C}(\mathcal{C}) := \mathbb{E}\left[\text{corr}(\mathbf{h},\mathcal{C}) \right], \tag{6.3c}$$

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

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

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

6.3. Analysis

We perform a detailed analysis of the three key optimization problems $\hat{\mathcal{L}}$, $\hat{\mathcal{U}}$ and $\hat{\mathcal{L}}_{dev}$. For each one of them we summarize the results of the analysis in Lemmas 6.1, 6.2 and 6.3 below. Each Lemma includes three statements. In the 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 5.2. 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 E in the Appendix.

6.3.1 Lower Key Optimization

Lemma 6.1 (Properties of $\hat{\mathcal{L}}$). *Let* $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$ *and* $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *and*

$$\hat{\mathcal{L}}(\mathbf{g}, \mathbf{h}) = \min_{\mathbf{w}} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2} \|\mathbf{g}\| - \mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\}, \tag{6.4}$$

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

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

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

and,

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

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

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

6.3.2 Upper Key Optimization

Lemma 6.2 (Properties of $\hat{\mathcal{U}}$). *Let* $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$, $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *and*

$$\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h}) = -\min_{\|\boldsymbol{\mu}\| \leq 1} \max_{\|\mathbf{w}\| = C_{up}} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + \|\boldsymbol{\mu}\| \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\}. \tag{6.5}$$

The following hold true:

1. Scalarization: $\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h}) = -\min_{0 \leq \alpha \leq 1} \left\{ -\alpha \cdot \sqrt{C_{up}^2 + \sigma^2}\, \|\mathbf{g}\| + C_{up} dist(\alpha \mathbf{h}, \mathcal{C}) \right\}$.
2. Deterministic result: *If* $\mathbf{h} \notin \mathcal{C}$ *and*

$$C_{up} dist(\mathbf{h}, \mathcal{C}) + C_{up} \frac{corr(\mathbf{h}, \mathcal{C})}{dist(\mathbf{h}, \mathcal{C})} < \sqrt{C_{up}^2 + \sigma^2} \|\mathbf{g}\|, \tag{6.6}$$

then,

$$\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h}) = \sqrt{C_{up}^2 + \sigma^2} \|\mathbf{g}\| - C_{up} dist(\mathbf{h}, \mathcal{C}). \tag{6.7}$$

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

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

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

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

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

6.3.3 Deviation Key Optimization

Lemma 6.3 (Properties of $\hat{\mathcal{L}}_{dev}$). *Let* $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$ *and* $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ *and*

$$\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\|\mathbf{w}\| \in S_{dev}} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2} \|\mathbf{g}\| - \mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\}, \tag{6.8}$$

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: $\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\alpha \in S_{dev}} \left\{ \sqrt{\alpha^2 + \sigma^2} \|\mathbf{g}\| - \alpha \cdot dist(\mathbf{h}, \mathcal{C}) \right\}$.
2. Deterministic result: *If*

$$\frac{\sigma \cdot dist(\mathbf{h}, \mathcal{C})}{\sqrt{\|\mathbf{g}\|^2 - dist^2(\mathbf{h}, \mathcal{C})}} \notin S_{dev}, \tag{6.9}$$

   then,

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

3. Probabilistic result: *Assume* $(1 - \epsilon_L)m > \mathbf{D}(\mathcal{C}) > \epsilon_L m$, *for some* $\epsilon_0 > 0$ *and set*

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

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

$$\mathbb{P}\left(\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) \geq (1 + t)\sigma \sqrt{m - \mathbf{D}(\mathcal{C})} \right) \geq 1 - c_1 \exp\left(-c_2 m\right). \tag{6.10}$$

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

Application of Gordon's Lemma to $\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v})$ introduced the three key optimizations in Lemma 5.2. Next, in Lemmas 6.1, 6.2 and 6.3 we carried out the analysis of those problems. Here, we combine the results of the four Lemmas mentioned above in order to evaluate $\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v})$ and to compute an exact value for the norm of its optimizer $\hat{\mathbf{w}}(\mathbf{A}, \mathbf{v})$. Lemma 6.4 below formally states the results of the analysis and the proof of it follows.

Lemma 6.4. *Assume* $m \geq \max\{\mathbf{D}(\mathcal{C}), \mathbf{D}(\mathcal{C}) + \mathbf{C}(\mathcal{C})\} + \epsilon_L m$ *and* $\mathbf{D}(\mathcal{C}) \geq \epsilon_L m$ *for some* $\epsilon_L > 0$. *Also, assume* m *is sufficiently large and let* $\gamma(m, n) = m$ *if* $\mathcal{C}$ *is a cone and* $\min\{m, \frac{m^2}{n}\}$ *else. Then, the following statements are true.*

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

$$\left| \hat{\mathcal{F}}(\mathbf{A}, \mathbf{v}) - \sigma \sqrt{m - \mathbf{D}(\mathcal{C})} \right| \leq \epsilon \sigma \sqrt{m - \mathbf{D}(\mathcal{C})}. \tag{6.11}$$

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

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

$$\left| \|\mathbf{w}\| - \sigma \sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}} \right| \geq \delta_{dev} \sigma \sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}}, \tag{6.12}$$

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

$$\|\mathbf{A}\mathbf{w} - \sigma \mathbf{v}\| + \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \geq \hat{\mathcal{F}}(\mathbf{A}, \mathbf{v}) + t\sigma \sqrt{m}, \tag{6.13}$$

   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*

$$\left| \|\hat{\mathbf{w}}(\mathbf{A}, \mathbf{v})\| - \sigma \sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}} \right| \leq \delta \sigma \sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}}. \tag{6.14}$$

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

Proof. We prove each one of the three statements of Theorem 6.4 sequentially. Assume the regime where $m \geq \max\{\mathbf{D}(\mathcal{C}), \mathbf{D}(\mathcal{C}) + \mathbf{C}(\mathcal{C})\} + \epsilon_L m$ and $\mathbf{D}(\mathcal{C}) \geq \epsilon_L m$ for some $\epsilon_L > 0$ and also m is sufficiently large.
1. *Proof of* (6.11)*:* Consider any $\epsilon' > 0$. First, we establish a high probability lower bound for $\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v})$. From Lemma 6.1,

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

with probability $1 - \exp(-\mathcal{O}(m))$. Combine this with the first statement of Lemma 5.2 to conclude that

$$\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v}) \geq (1 - \epsilon')\sigma \sqrt{m - \mathbf{D}(\mathcal{C})} - \epsilon' \sigma \sqrt{m}, \tag{6.15}$$

with the same probability.

Similarly, for a high probability upper bound for $\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v})$ we have from Lemma 6.2, that

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

with probability $1 - \exp(-\mathcal{O}(\gamma(m, n)))$. Combine this with the second statement of Lemma 5.2 to conclude that

$$\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v}) \leq (1 + \epsilon')\sigma \sqrt{m - \mathbf{D}(\mathcal{C})} + \epsilon' \sigma \sqrt{m}, \tag{6.16}$$

with the same probability. To conclude the proof of (6.11) fix any positive constant $\epsilon > 0$, and observe that by choosing $\epsilon' = \epsilon \frac{\sqrt{\epsilon_L}}{1 + \sqrt{\epsilon_L}}$ in (6.15) and (6.16) we ensure that $\epsilon' \left(1 + \frac{\sqrt{m}}{\sqrt{m - \mathbf{D}(\mathcal{C})}}\right) \leq \epsilon$. It then follows from (6.15) and (6.16) that there exist $c_1, c_2 > 0$ such that

$$\left| \frac{\hat{\mathcal{F}}(\mathbf{A}, \mathbf{v})}{\sigma \sqrt{m - \mathbf{D}(\mathcal{C})}} - 1 \right| \leq \epsilon, \tag{6.17}$$

with probability $1 - c_1 \exp(-c_2 \gamma(m, n))$.
2. *Proof of* (6.13)*:* Fix any $\delta_{dev} > 0$. In accordance to its definition in previous sections define the set

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

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

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

Combining this with the third statement of Lemma 5.2, it suffices for the proof of (6.13) to show that there exists constant $t(\delta_{dev}) > 0$ such that

$$\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) \geq \hat{\mathcal{F}}(\mathbf{A}, \mathbf{v}) + 2t\sigma \sqrt{m}, \tag{6.18}$$

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

To show (6.18), start from Lemma 6.3 which gives that here exists $t'(\delta_{dev}) > 0$, such that

$$\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) \geq (1 + t')\sigma \sqrt{m - \mathbf{D}(\mathcal{C})}, \tag{6.19}$$

with probability $1-\exp(-\mathcal{O}(m))$. Furthermore, from the first statement of Lemma 6.4,

$$\hat{\mathcal{F}}(\mathbf{A},\mathbf{v}) \leq (1+\frac{t'}{2})\sigma\sqrt{m-\mathbf{D}(\mathcal{C})}, \tag{6.20}$$

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

$$2t\sigma\sqrt{m} \leq \frac{t'}{2}\sigma\sqrt{m-\mathbf{D}(\mathcal{C})}. \tag{6.21}$$

Combine (6.19), (6.20) and (6.21) to conclude that (6.18) indeed holds with the desired probability.
3. *Proof of* (6.14)*:* The third statement of Lemma 6.4 is a simple consequence of its second statement. Fix any $\epsilon>0$. The proof is by contradiction. Assume that $\hat{\mathbf{w}}(\mathbf{A},\mathbf{v})$ does not satisfy (6.14). It then satisfies (6.12) for $\delta_{dev}=\epsilon$. Thus, it follows from the second statement of Lemma 6.4, that there exists $t(\epsilon)>0$ such that

$$\hat{\mathcal{F}}(\mathbf{A},\mathbf{v}) \geq \hat{\mathcal{F}}(\mathbf{A},\mathbf{v}) + t\sigma\sqrt{m}, \tag{6.22}$$

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

□

7. The NSE of the C-LASSO

In this section, we prove the second statement of Theorem 3.1, namely (3.2). We restate the theorem here for ease of reference.

Theorem 3.1 (NSE of C-LASSO)**.** *Assume there exists a constant $\epsilon_L>0$ such that, $(1-\epsilon_L)m \geq \mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+) \geq \epsilon_L m$ and m is sufficiently large. For any $\epsilon>0$, there exists a constant $C=C(\epsilon,\epsilon_L)>0$ such that, with probability $1-\exp(-Cm)$,*

$$\frac{\|\mathbf{x}_c^*-\mathbf{x}_0\|^2}{\sigma^2} \leq (1+\epsilon)\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}, \tag{3.1}$$

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,

$$\left|\frac{\|\mathbf{x}_c^*-\mathbf{x}_0\|^2}{\sigma^2}\times\frac{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}-1\right|<\epsilon. \tag{3.2}$$

First, in Section 7.1 we focus on the approximated C-LASSO and prove that its NSE concentrates around $\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}$ for arbitrary values of σ. Later in Section 7.2, we use that result and fundamental properties of the approximated problem to prove (3.2), i.e. that the NSE of the original problem concentrates around the same quantity for small enough σ.

7.1. Approximated C-LASSO Problem

Recall the definition of the approximated C-LASSO problem in (5.6). As it has been argued previously, this is equivalent to the generic problem (6.2) with $\mathcal{C}=\text{cone}\{\partial f(\mathbf{x}_0)\}$. Hence, to calculate its NSE we will simply apply the results we obtained throughout Section 6. We first start by mapping the generic formulation in Section 6 to the C-LASSO.

Lemma 7.1. *Let $\mathcal{C}=cone\{\partial f(\mathbf{x}_0)\}$. Then,*

- $\mathbf{D}(\mathcal{C})=\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)$,
- $corr(\mathbf{h},\mathcal{C})=0$, *for all* $\mathbf{h}\in\mathbb{R}^n$,
- $\mathbf{C}(\mathcal{C})=0$.

Proof. The first statement follows by definition of the quantities involved. The second statement is a direct consequence of Moreau's decomposition theorem (Fact A.1) applied on the closed and convex cone cone$\{\partial f(\mathbf{x}_0)\}$. The last statement follows easily after taking expectation in both sides of the equality in the second statement. □

With this mapping, we can directly apply Lemma 6.4, where $\mathcal{C}$ is a cone, to conclude with the desired result. The following corollary summarizes the result.

Corollary 7.1. *Assume $(1-\epsilon_L)m \geq \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) \geq \epsilon_L m$, for some $\epsilon_L > 0$. Also, assume m is sufficiently large. Then, for any constants $\epsilon_1, \epsilon_2 > 0$, there exist constants $c_1, c_2 > 0$ such that with probability $1 - c_1 \exp(-c_2 m)$,*

$$\left| \frac{\hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v})}{\sigma\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}} - 1 \right| \leq \epsilon_1,$$

and

$$\left| \frac{\|\hat{\mathbf{w}}_c(\mathbf{A}, \mathbf{v})\|^2}{\sigma^2} - \frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)} \right| \leq \epsilon_2.$$

7.2. Original C-LASSO Problem

In this section we prove (3.2). For the proof we rely on Corollary 7.1. First, we require the introduction of some useful concepts from convex analysis.

7.2.1 Tangent Cone and Cone of the Subdifferential

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

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

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

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

where $Cl(\cdot)$ denotes the closure of a set. By definition, tangent cone $\mathcal{T}_{\mathcal{C}}(\mathbf{x}^*)$ and feasible set $F_{\mathcal{C}}(\mathbf{x}^*)$ should be *close* to each other around a small neighborhood of 0. The following proposition is a corollary of Proposition F.1 of [41] 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 7.1 (Approximating the tangent cone, [41])**.** *Let $\mathcal{C}$ be a closed convex set and $\mathbf{x}^* \in \mathcal{C}$. For any $\delta > 0$, there exists $\epsilon > 0$ such that*

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

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

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

Lemma 7.2 ([86])**.** *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,*

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

7.2.2 Proof of Theorem 3.1: Small σ regime

We prove here the second part of Theorem 3.1, namely (3.2). For a proof of (3.1) see Section 10. For the purposes of the proof, we will use $\mathcal{C} = \{\mathbf{x} \big| f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$. 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

$$\eta_c = \frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}.$$

Recalling the definition of the approximated C-LASSO problem in (5.6), we may write

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

where for the last equality we have used Lemma 7.2. Hence,

$$\hat{\mathbf{w}}_c \in \mathcal{T}_{\mathcal{C}}(\mathbf{x}_0). \tag{7.1}$$

At the same time, clearly,

$$\mathbf{w}_c^* \in F_{\mathcal{C}}(\mathbf{x}_0). \tag{7.2}$$

After Corollary 7.1, $\|\hat{\mathbf{w}}_c\|^2$ 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 7.1, hold with high probability for some arbitrary $\epsilon_1, \epsilon_2 > 0$. It suffices to prove that for any $\epsilon_3 > 0$ there exists $\sigma_0 > 0$ such that

$$\left| \frac{\|\mathbf{w}_c^*\|^2}{\sigma^2} - \eta_c \right| \leq \epsilon_3, \tag{7.3}$$

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 7.1, there exists σ_0 such that

$$\text{dist}(\mathbf{w}, F_{\mathcal{C}}(\mathbf{x}_0)) \leq \delta \|\mathbf{w}\| \tag{7.4}$$

for all $\mathbf{w} \in \mathcal{T}_{\mathcal{C}}(\mathbf{x}_0)$ satisfying $\|\mathbf{w}\| \leq C = C(\sigma_0, \epsilon_2) := \sigma_0\sqrt{(1+\epsilon_2)\eta_c}$.

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

- Using Corollary 7.1, with high probability $\hat{\mathbf{w}}_c$ satisfies,

$$\|\hat{\mathbf{w}}_c\| \leq \sigma\sqrt{(1+\epsilon_2)\eta_c} \leq C. \tag{7.5}$$

- $\mathbf{A}$ has independent standard normal entries. Hence, its spectral norm satisfies $\|\mathbf{A}\|_2 \leq 2(\sqrt{n} + \sqrt{m})$ with probability $1 - \exp(-\mathcal{O}(\max\{m, n\}))$, [71].

- Using (6.13) of Lemma 6.4 with $\mathcal{C} = \text{cone}(\partial f(\mathbf{x}_0))$, there exists a constant $t = t(\epsilon_3)$ so that for all $\mathbf{w}$ satisfying $|\frac{\|\mathbf{w}\|^2}{\sigma^2} - \eta_c| \geq \epsilon_3$, we have,

$$\|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\| + \max_{\mathbf{s} \in \text{cone}(\partial f(\mathbf{x}_0))} \mathbf{s}^T\mathbf{w} \geq \hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + t(\epsilon_3)\sigma\sqrt{m}. \tag{7.6}$$

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

$$\mathbf{p}(\hat{\mathbf{w}}_c) := \text{Proj}(\hat{\mathbf{w}}_c, F_{\mathcal{C}}(\mathbf{x}_0)) = \hat{\mathbf{w}}_c - \mathbf{\Pi}(\hat{\mathbf{w}}_c, F_{\mathcal{C}}(\mathbf{x}_0)). \tag{7.7}$$

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

$$\begin{aligned}\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\| &\leq \|\mathbf{A}\hat{\mathbf{w}}_c - \sigma\mathbf{v}\| + \|\mathbf{A}\hat{\mathbf{w}}_c - \mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c)\| \\ &\leq \hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + \|\mathbf{A}\|_2 \text{dist}(\hat{\mathbf{w}}_c, F_{\mathcal{C}}(\mathbf{x}_0)) \\ &\leq \hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + \|\mathbf{A}\|_2 \sigma\delta\sqrt{(1+\epsilon_2)\eta_c} \\ &\leq \hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + 2(\sqrt{m}+\sqrt{n})\sigma\delta\sqrt{(1+\epsilon_2)\eta_c}. \end{aligned} \tag{7.8}$$

The first inequality is an application of the triangle inequality and the second one follows from (7.7). For the third inequality, we have used (7.1) and combined (7.4) with (7.5).

Next, we show that if (7.3) was not true then a suitable choice of δ would make $\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\|$ much larger than the optimal $\hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v})$ than (7.8) allows. Therefore, concluding a desired contradiction. More precisely, assuming (7.3) does not hold, we have

$$\begin{aligned}\|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\| &\geq \|\mathbf{A}\mathbf{w}_c^* - \sigma\mathbf{v}\| \\ &\geq \hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + t(\epsilon_3)\sigma\sqrt{m}. \end{aligned} \tag{7.9}$$

The first inequality above follows since $\mathbf{p}(\hat{\mathbf{w}}_c) \in F_{\mathcal{C}}(\mathbf{x}_0)$ and from the optimality of $\mathbf{w}_c^* \in F_{\mathcal{C}}(\mathbf{x}_0)$. To get the second inequality, recall that (7.3) is not true. Also, from (7.2), $\max_{\mathbf{s}\in\text{cone}(\partial f(\mathbf{x}_0))} \mathbf{s}^T\mathbf{w}_c^* = \max_{\mathbf{s}\in(\mathcal{T}(\mathbf{x}_0))^\circ} \mathbf{s}^T\mathbf{w}_c^* = 0$. Combine these and invoke (7.6).

To conclude, choose σ_0 sufficiently small to ensure $\delta < \frac{t(\epsilon_3)\sqrt{m}}{2(\sqrt{m}+\sqrt{n})\sqrt{(1+\epsilon_2)\eta_c}}$ and combine (7.8) and (7.9) to obtain the following contradiction.

$$\begin{aligned}\hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + 2(\sqrt{m}+\sqrt{n})\delta\sigma\sqrt{(1+\epsilon_2)\eta_c} &\geq \|\mathbf{A}\mathbf{p}(\hat{\mathbf{w}}_c) - \sigma\mathbf{v}\| \\ &\geq \hat{\mathcal{F}}_c(\mathbf{A}, \mathbf{v}) + t(\epsilon_3)\sigma\sqrt{m}. \end{aligned}$$

σ_0 is a deterministic number that is a function of $m, n, f, \mathbf{x}_0, \epsilon_3$.

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 $\mathcal{R}_{OFF}$, $\mathcal{R}_{ON}$ and $\mathcal{R}_\infty$. 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 9.

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 $\mathbf{D}_f(\mathbf{x}_0, \lambda)$, projection $\mathbf{P}_f(\mathbf{x}_0, \lambda)$ and correlation $\mathbf{C}_f(\mathbf{x}_0, \lambda)$. 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 8.1. *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. $\mathbf{D}_f(\mathbf{x}_0,\lambda)+2\mathbf{C}_f(\mathbf{x}_0,\lambda)+\mathbf{P}_f(\mathbf{x}_0,\lambda)=n$.
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. $\lim_{\lambda\to\infty}\mathbf{D}_f(\mathbf{x}_0,\lambda)=\infty$, $\lim_{\lambda\to\infty}\mathbf{P}_f(\mathbf{x}_0,\lambda)=\infty$, *and* $\lim_{\lambda\to\infty}\mathbf{C}_f(\mathbf{x}_0,\lambda)=-\infty$.
4. $\mathbf{P}_f(\mathbf{x}_0,\lambda)$, $\mathbf{C}_f(\mathbf{x}_0,\lambda)$ *and* $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ *are all continuous functions of* $\lambda\geq 0$.
5. $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ *is strictly convex and attains its minimum at a unique point. Denote* λ_{best} *the unique minimizer of* $\mathbf{D}_f(\mathbf{x}_0,\lambda)$.
6. $\mathbf{P}_f(\mathbf{x}_0,\lambda)$ *is an increasing function for* $\lambda\geq 0$.
7. $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ *is differentiable for* $\lambda>0$. *For* $\lambda>0$,

$$\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda}=-\frac{2}{\lambda}\mathbf{C}_f(\mathbf{x}_0,\lambda).$$

   For $\lambda=0$, *interpret* $\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda}$ *as a right derivative.*

8. $$\mathbf{C}_f(\mathbf{x}_0,\lambda)\begin{cases}\geq 0 & ,\lambda\in[0,\lambda_{best}]\\ =0 & ,\lambda=\lambda_{best}\\ \leq 0 & ,\lambda\in[\lambda_{best},\infty)\end{cases}$$

9. $\mathbf{D}_f(\mathbf{x}_0,\lambda)+\mathbf{C}_f(\mathbf{x}_0,\lambda)$ *is strictly decreasing for* $\lambda\in[0,\lambda_{best}]$.

Some of the statements in Lemma 8.1 are easy to prove, while others require more work. Statements 5 and 7 have been recently proved in [31]. We defer the proofs of all statements to Appendix G.

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.

8.2.1 λ_{best}

The first key parameter is λ_{best} which was defined in Lemma 8.1 to be the unique minimum of $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ 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 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, $\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{best}})\leq\mathbf{D}_f(\mathbf{x}_0,0)=n$.

8.2.2 λ_{max}

The second key parameter λ_{max} is defined as the unique $\lambda\geq\lambda_{\text{best}}$ that satisfies $\mathbf{D}_f(\mathbf{x}_0,\lambda)=m$. We formally repeat this definition in the following Lemma.

Lemma 8.2. *Suppose* $\mathbf{D}_f(\mathbf{x}_0,\lambda_{best})<m$ *and consider the following equation over* $\lambda\geq\lambda_{best}$*:*

$$\mathbf{D}_f(\mathbf{x}_0,\lambda)=m,\quad \lambda\geq\lambda_{best}. \tag{8.1}$$

Equation (8.1) *has a unique solution, which we denote* λ_{max}.

Proof. We make use of Lemma 8.1. First, we show that equation (8.1) has at most one solution: $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ is a strictly convex function of $\lambda\geq 0$ and thus strictly increasing for $\lambda\geq\lambda_{\text{best}}$. Next, we show that (8.1) 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, $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ is continuous in λ. Combining those facts and using the intermediate value theorem we conclude with the desired result. □

8.2.3 λ_{crit}

The third key parameter λ_{crit} is defined to be the unique $\lambda \leq \lambda_{\text{best}}$ that satisfies $m - \mathbf{D}_f(\mathbf{x}_0, \lambda) = \mathbf{C}_f(\mathbf{x}_0, \lambda)$ when $m \leq n$ or to be 0 when $m > n$. We formally repeat this definition in the following Lemma.

Lemma 8.3. *Suppose $\mathbf{D}(\mathbf{x}_0, \lambda_{best}) < m$ and consider the following equation over $0 \leq \lambda \leq \lambda_{best}$:*

$$m - \mathbf{D}_f(\mathbf{x}_0, \lambda) = \mathbf{C}_f(\mathbf{x}_0, \lambda), \quad 0 \leq \lambda \leq \lambda_{best}. \tag{8.2}$$

- *If $m \leq n$, then (8.2) has a unique solution, which we denote as λ_{crit}.*
- *If $m > n$, then (8.2) has no solution. Then $\lambda_{crit} = 0$.*

Proof. We repeatedly make use of Lemma 8.1. For convenience define the function

$$g(\lambda) = \mathbf{D}_f(\mathbf{x}_0, \lambda) + \mathbf{C}_f(\mathbf{x}_0, \lambda),$$

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 (8.2) 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 (8.2) has no solution. □

8.3. Regions of Operation: $\mathcal{R}_{OFF}$, $\mathcal{R}_{ON}$, $\mathcal{R}_{\infty}$

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 8.1. *Define the following regions of operation for the ℓ_2-LASSO problem:*

- $\mathcal{R}_{OFF} = \{\lambda \mid 0 \leq \lambda \leq \lambda_{crit}\}$,
- $\mathcal{R}_{ON} = \{\lambda \mid \lambda_{crit} < \lambda < \lambda_{\text{max}}\}$,
- $\mathcal{R}_{\infty} = \{\lambda \mid \lambda \geq \lambda_{\text{max}}\}$.

Remark: The definition of $\mathcal{R}_{\text{ON}}$ in Definition 8.1 is consistent to the Definition in 3.1. In other words, $\lambda_{\text{crit}} \leq \lambda \leq \lambda_{\text{max}}$ if and only if $m \geq \max\{\mathbf{D}_f(\mathbf{x}_0, \lambda), \mathbf{D}_f(\mathbf{x}_0, \lambda) + \mathbf{C}_f(\mathbf{x}_0, \lambda)\}$. This follows after combining Lemmas 8.2 and 8.3 with the Lemma 8.4 below.

Lemma 8.4. *The following hold:*

1. *$m - \mathbf{D}_f(\mathbf{x}_0, \lambda) \leq \mathbf{C}_f(\mathbf{x}_0, \lambda)$ for all $\lambda \in \mathcal{R}_{OFF}$ if $\lambda_{crit} \neq 0$.*
2. *$m - \mathbf{D}_f(\mathbf{x}_0, \lambda) > \max\{0, \mathbf{C}_f(\mathbf{x}_0, \lambda)\}$ for all $\lambda \in \mathcal{R}_{ON}$,*
3. *$m \leq \mathbf{D}_f(\mathbf{x}_0, \lambda)$ for all $\lambda \in \mathcal{R}_{\infty}$.*

Proof. We prove the statements in the order they appear. We use Lemma 8.1 throughout.
1. The function $\mathbf{D}_f(\mathbf{x}_0, \lambda) + \mathbf{C}_f(\mathbf{x}_0, \lambda)$ is strictly decreasing in $[0, \lambda_{\text{best}}]$. Thus, assuming $\lambda_{\text{crit}} \neq 0$, $\mathbf{D}_f(\mathbf{x}_0, \lambda) + \mathbf{C}_f(\mathbf{x}_0, \lambda) \geq \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{crit}}) + \mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{crit}}) = m$ for all $\lambda \in [0, \lambda_{\text{crit}}]$.
2. Since $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ is strictly convex, $m - \mathbf{D}_f(\mathbf{x}_0, \lambda)$ is strictly concave and has a unique maximum at λ_{best}. Therefore, for all $\lambda \in [\lambda_{\text{crit}}, \lambda_{\text{max}}]$,

$$m - \mathbf{D}_f(\mathbf{x}_0, \lambda) \geq \max\{ \underbrace{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{crit}})}_{=\mathbf{C}_f(\mathbf{x}_0, \lambda_{\text{crit}}) \geq 0}, \underbrace{m - \mathbf{D}_f(\mathbf{x}_0, \lambda_{\text{max}})}_{=0} \} \geq 0.$$

Furthermore, $\mathbf{D}_f(\mathbf{x}_0,\lambda)+\mathbf{C}_f(\mathbf{x}_0,\lambda)$ is strictly decreasing in $[0,\lambda_{\text{best}}]$. Thus, $\mathbf{D}_f(\mathbf{x}_0,\lambda)+\mathbf{C}_f(\mathbf{x}_0,\lambda)<\mathbf{D}_f(\mathbf{x}_0,\lambda_{\text{crit}})+\mathbf{C}_f(\mathbf{x}_0,\lambda_{\text{crit}})\leq m$ for all $\lambda\in(\lambda_{\text{crit}},\lambda_{\text{best}}]$. For $\lambda\in[\lambda_{\text{best}},\lambda_{\max})$, we have $m-\mathbf{D}_f(\mathbf{x}_0,\lambda)>0\geq\mathbf{C}_f(\mathbf{x}_0,\lambda)$.
3. $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ is strictly convex. Hence, $m-\mathbf{D}_f(\mathbf{x}_0,\lambda)$ is strictly decreasing in $[\lambda_{\text{best}},\infty)$. This proves that $m-\mathbf{D}_f(\mathbf{x}_0,\lambda)\leq m-\mathbf{D}_f(\mathbf{x}_0,\lambda_{\max})=0$ for all $\lambda\geq\lambda_{\max}$.

□

9. The NSE of the ℓ_2-LASSO

We split our analysis in three sections, one for each of the three regions $\mathcal{R}_{\text{OFF}}$, $\mathcal{R}_{\text{ON}}$ and $\mathcal{R}_{\infty}$. We start from $\mathcal{R}_{\text{ON}}$, for which the analysis is similar in nature to C-LASSO.

9.1. $\mathcal{R}_{ON}$

In this section we prove Theorem 3.2 which characterizes the NSE of the ℓ_2-LASSO in the region $\mathcal{R}_{\text{ON}}$. We repeat the statement of the theorem here, for ease of reference.

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

$$\left|\frac{\|\mathbf{x}^*_{\ell_2}-\mathbf{x}_0\|^2}{\sigma^2}\times\frac{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}{\mathbf{D}_f(\mathbf{x}_0,\lambda)}-1\right|<\epsilon. \tag{3.3}$$

As usual, we first focus on the approximated ℓ_2-LASSO problem in Section 9.1.1. Next, in Section 9.1.2, we translate this result to the original ℓ_2-LASSO problem.

9.1.1 Approximated ℓ_2-LASSO

The approximated ℓ_2-LASSO problem is equivalent to the generic problem (6.2) after taking $\mathcal{C}=\lambda\partial f(\mathbf{x}_0)$. Hence, we simply need to apply the result of Lemma 6.4. with $\mathbf{D}(\mathcal{C})$ and $\mathbf{C}(\mathcal{C})$ corresponding to $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ and $\mathbf{C}_f(\mathbf{x}_0,\lambda)$. We conclude with the following result.

Corollary 9.1. *Let $m\geq\min_{\lambda\geq0}\mathbf{D}_f(\mathbf{x}_0,\lambda)$ and assume there exists constant $\epsilon_L>0$ such that $(1-\epsilon_L)m\geq\max\{\mathbf{D}_f(\mathbf{x}_0,\lambda),\mathbf{D}_f(\mathbf{x}_0,\lambda)+\mathbf{C}_f(\mathbf{x}_0,\lambda)\}$ and $\mathbf{D}_f(\mathbf{x}_0,\lambda)\geq\epsilon_L m$. Further assume that m is sufficiently large. Then, for any constants $\epsilon_1,\epsilon_2>0$, there exist constants $c_1,c_2>0$ such that with probability $1-c_1\exp(-c_2\min\{m,\frac{m^2}{n}\})$,*

$$\left|\frac{\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})}{\sigma\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}}-1\right|\leq\epsilon_1, \tag{9.1}$$

and

$$\left|\frac{\|\hat{\mathbf{w}}_{\ell_2}(\mathbf{A},\mathbf{v})\|^2}{\sigma^2}-\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}\right|\leq\epsilon_2. \tag{9.2}$$

9.1.2 Original ℓ_2-LASSO: Proof of Theorem 3.2

Next, we use Corollary 9.1 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 9.1 (Max formula, [77,78])**.** *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,*

$$\lim_{\epsilon\to0^+}\frac{f(\mathbf{x}+\epsilon\mathbf{v})-f(\mathbf{x})}{\epsilon}=\sup_{\mathbf{s}\in\partial f(\mathbf{x})}\langle\mathbf{s},\mathbf{v}\rangle.$$

In particular, the subdifferential $\partial f(\mathbf{x})$ is nonempty.

Proposition 9.1 considers a fixed direction $\mathbf{v}$, and compares $f(\mathbf{x}_0+\epsilon\mathbf{v})$ and $\hat{f}(\mathbf{x}_0+\epsilon\mathbf{v})$. 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 [85].

Proposition 9.2 (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 (2.2). Then, for any $\delta>0$, there exists $\epsilon>0$ such that,*

$$f(\mathbf{x}_0+\mathbf{w})-\hat{f}(\mathbf{x}_0+\mathbf{w})\leq\delta\|\mathbf{w}\|, \tag{9.3}$$

for all $\mathbf{w}\in\mathbb{R}^n$ with $\|\mathbf{w}\|\leq\epsilon$.

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,

$$\eta_{\ell_2}=\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}.$$

After Corollary 9.1, $\|\hat{\mathbf{w}}_{\ell_2}\|^2$ 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 9.1 hold with high probability for some arbitrary $\epsilon_1,\epsilon_2>0$. It suffices to prove that for any $\epsilon_3>0$ there exists $\sigma_0>0$ such that

$$\left|\frac{\|\mathbf{w}^*_{\ell_2}\|^2}{\sigma^2}-\eta_{\ell_2}\right|\leq\epsilon_3, \tag{9.4}$$

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 9.2, there exists σ_0 such that

$$f(\mathbf{x}_0+\mathbf{w})-\hat{f}(\mathbf{x}_0+\mathbf{w})\leq\delta\|\mathbf{w}\| \tag{9.5}$$

for all $\mathbf{w}$ satisfying $\|\mathbf{w}\|\leq C=C(\sigma_0,\epsilon_2):=\sigma_0\sqrt{(1+\epsilon_2)\eta_{\ell_2}}$. Now, fix any $\sigma<\sigma_0$. We will make use of the fact that the following three events hold with high probability.

- Using Corollary 9.1, with high probability $\hat{\mathbf{w}}_{\ell_2}$ satisfies,

$$\|\hat{\mathbf{w}}_{\ell_2}\|\leq\sigma\sqrt{(1+\epsilon_2)\eta_{\ell_2}}\leq C. \tag{9.6}$$

- Using (6.13) of Lemma 6.4 with $\mathcal{C}=\lambda\partial f(\mathbf{x}_0)$, there exists a constant $t=t(\epsilon_3)$ so that for any $\mathbf{w}$ satisfying $|\frac{\|\mathbf{w}\|^2}{\sigma^2}-\eta_{\ell_2}|\geq\epsilon_3$, we have,

$$\|\mathbf{A}\mathbf{w}-\sigma\mathbf{v}\|+\max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)}\mathbf{s}^T\mathbf{w}\geq\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})+t(\epsilon_3)\sigma\sqrt{m}. \tag{9.7}$$

Combine (9.6) with (9.5) to find that

$$\begin{aligned}\|\mathbf{A}\hat{\mathbf{w}}_{\ell_2}-\sigma\mathbf{v}\|+\lambda(f(\mathbf{x}_0+\hat{\mathbf{w}}_{\ell_2})-f(\mathbf{x}_0))&\leq\underbrace{\|\mathbf{A}\hat{\mathbf{w}}_{\ell_2}-\sigma\mathbf{v}\|+\lambda(\hat{f}(\mathbf{x}_0+\hat{\mathbf{w}}_{\ell_2})-f(\mathbf{x}_0))}_{=\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})}+\delta\|\hat{\mathbf{w}}_{\ell_2}\|\\&\leq\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})+\delta\sigma\sqrt{(1+\epsilon_2)\eta_{\ell_2}}.\end{aligned} \tag{9.8}$$

Now, assume that $\|\mathbf{w}^*_{\ell_2}\|$ does not satisfy (9.4). Then,

$$\begin{aligned}\|\mathbf{A}\hat{\mathbf{w}}_{\ell_2}-\sigma\mathbf{v}\|+\lambda(f(\mathbf{x}_0+\hat{\mathbf{w}}_{\ell_2})-f(\mathbf{x}_0))&\geq\mathcal{F}^*_{\ell_2}(\mathbf{A},\mathbf{v}) &(9.9)\\&\geq\|\mathbf{A}\mathbf{w}^*_{\ell_2}-\sigma\mathbf{v}\|+\lambda\max_{\mathbf{s}\in\lambda\partial f(\mathbf{x}_0)}\mathbf{s}^T\mathbf{w}^*_{\ell_2} &(9.10)\\&\geq\hat{\mathcal{F}}_{\ell_2}(\mathbf{A},\mathbf{v})+t(\epsilon_3)\sigma\sqrt{m}. &(9.11)\end{aligned}$$

(9.9) follows from optimality of $\mathbf{w}^*_{\ell_2}$. For (9.10) 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, (9.11) follows from (9.7).

To complete the proof, choose $\delta<\frac{t\sqrt{m}}{\sqrt{(1+\epsilon_2)\eta_{\ell_2}}}$. This will result in contradiction between (9.8) and (9.11). Observe that, our choice of δ and σ_0 is deterministic and depends on $m,\mathbf{x}_0,f(\cdot),\epsilon_3$.

9.1.3 A Property of the NSE Formula

Theorem 3.2 shows that the asymptotic NSE formula in $\mathcal{R}_{\text{ON}}$ is $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}$. The next lemma provides a useful property of this formula as a function of λ on $\mathcal{R}_{\text{ON}}$.

Lemma 9.1. *$\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}$ is a convex function of λ over $\mathcal{R}_{ON}$.*

Proof. From 8.1, $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ 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 $\mathcal{R}_{\text{ON}}$. Consequently, the asymptotic NSE formula is a composition of an increasing convex function with a convex function, and is thus itself convex [83]. □

9.2. $\mathcal{R}_{OFF}$

Our analysis, unfortunately, does not extend to $\mathcal{R}_{OFF}$, and we have no proof that characterizes the NSE in this regime. On the other hand, our extensive numerical experiments (see Section 13) show that, in this regime, the optimal estimate $\mathbf{x}^*_{\ell_2}$ of (1.6) 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,

$$\min f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{9.12}$$

Here, we provide some intuition to why it is reasonable to expect this to be the case. Recall that $\lambda \in \mathcal{R}_{\text{OFF}}$ iff $0 \leq \lambda \leq \lambda_{\text{crit}}$, and so the "small" values of the penalty parameter λ are in $\mathcal{R}_{\text{OFF}}$. As λ gets smaller, $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|$ becomes the dominant term, and ℓ_2-LASSO penalizes this term more. So, at least for sufficiently small λ, the reduction to problem (9.12) would not be surprising. Lemma 9.2 formalizes this idea for the small λ regime.

Lemma 9.2. *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 $1 - \exp(-\mathcal{O}(n))$. Here, $o(1)$ term is arbitrarily small positive constant.*

Proof. When $m \leq \alpha n$ for some constant $0 < \alpha < 1$, $\sqrt{n} - \sqrt{m} = \mathcal{O}\left(\sqrt{n}\right)$. Then, from standard concentration results (see [71]), with probability $1 - \exp(-\mathcal{O}(n))$, 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 $\|\mathbf{p}\| = 0$. Denote $\mathbf{w}_2 := \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{p}$. Using (9.13), with the same probability,

$$\|\mathbf{w}_2\|^2 = \mathbf{p}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{p} \leq \frac{\|\mathbf{p}\|^2}{(\sigma_{min}(\mathbf{A}^T))^2} \leq \frac{\|\mathbf{p}\|^2}{((\sqrt{n} - \sqrt{m})(1 - o(1)))^2}, \tag{9.13}$$

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,

$$\begin{aligned} 0 &\geq \|\mathbf{p}\| + \lambda f(\mathbf{x}^*_{\ell_2}) - \lambda f(\mathbf{x}_2) \\ &\geq \|\mathbf{p}\| - \lambda L\|\mathbf{x}^*_{\ell_2} - \mathbf{x}^*_2\| = \|\mathbf{p}\| - \lambda L\|\mathbf{w}_2\| \qquad (9.14) \\ &\geq \|\mathbf{p}\|(1 - \lambda\frac{L}{(\sqrt{n} - \sqrt{m})(1 - o(1))}). \qquad (9.15) \end{aligned}$$

The inequality in (9.14) follows from Lipschitzness of $f(\cdot)$, while we use (9.13) to find (9.15). For the sake of contradiction, assume that $\|\mathbf{p}\| \neq 0$, then (9.15) reduces to $0 > 0$, clearly, a contradiction. □

For an illustration of Lemma 9.2, consider the case where $f(\cdot) = \|\cdot\|_1$. ℓ_1-norm is Lipschitz with $L = \sqrt{n}$ (see [58] for related discussion). Lemma 9.2 would, then, require $\lambda < 1 - \sqrt{\frac{m}{n}}$ to be applicable. As an example, considering the setup in Figure 3, Lemma 9.2 would yield $\lambda < 1 - \sqrt{\frac{1}{2}} \approx 0.292$ whereas $\lambda_{\text{crit}} \approx 0.76$. While Lemma 9.2 supports our claims on $\mathcal{R}_{\text{OFF}}$, it does not say much about the exact location of the transition point, at which the ℓ_2-LASSO reduces to (9.12). We claim this point is $\lambda = \lambda_{\text{crit}}$.

9.3. $\mathcal{R}_\infty$

In this region $m \leq \mathbf{D}_f(\mathbf{x}_0, \lambda)$. In this region, we expect *no* noise robustness, namely, $\frac{\|\mathbf{x}^*_{\ell_2} - \mathbf{x}_0\|^2}{\sigma^2} \to \infty$ as $\sigma \to 0$. In this work, we show this under a stricter assumption, namely, $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$. See Theorem 3.4 and Section 12 for more details. Our proof method relies on results of [31] rather than Gordon's Lemma. On the other hand, we believe, application of Gordon's Lemma can give the desired result for the wider regime $m < \mathbf{D}_f(\mathbf{x}_0, \lambda)$. We leave this as a future work.

10. Constrained-LASSO Analysis for Arbitrary σ

In Section 7 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 (3.1), 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 Gordon's Lemma.

10.1. 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 (5.1) and (5.2). Finally, we will be making use of the following functions:

$$\mathcal{F}(\mathbf{w}; \mathbf{A}, \mathbf{v}) := \|\mathbf{A}\mathbf{w} - \sigma\mathbf{v}\|,$$

$$\mathcal{L}(\mathbf{w}; \mathbf{g}, \mathbf{h}) := \sqrt{\|\mathbf{w}\|^2 + \sigma^2}\|\mathbf{g}\| - \mathbf{h}^T\mathbf{w}, \tag{10.1}$$

$$L(\alpha; a, b) := \sqrt{\alpha^2 + \sigma^2}a - \alpha b. \tag{10.2}$$

Using this notation, and denoting the optimal cost of the (original) C-LASSO (see (1.5)) as $\mathcal{F}^*_c(\mathbf{A}, \mathbf{v})$, we write

$$\mathcal{F}^*_c(\mathbf{A}, \mathbf{v}) = \min_{f_p(\mathbf{w}) \leq 0} \mathcal{F}(\mathbf{w}; \mathbf{A}, \mathbf{v}) = \mathcal{F}(\mathbf{w}^*_c; \mathbf{A}, \mathbf{v}). \tag{10.3}$$

10.2. Lower Key Optimization

As a first step in our proof, we apply Gordon's Lemma to the original C-LASSO problem in (10.3). Recall, that application of Corollary 5.1 to the approximated problem resulted in the following key optimization:

$$\hat{\mathcal{L}}(\mathbf{g}, \mathbf{h}) = \min_{\hat{f}_p(\mathbf{w}) \leq 0} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2}\|\mathbf{g}\| - \mathbf{h}^T\mathbf{w} \right\} = \min_{\hat{f}_p(\mathbf{w}) \leq 0} \mathcal{L}(\mathbf{w}; \mathbf{g}, \mathbf{h}). \tag{10.4}$$

Denote the minimizer of (10.4), as $\hat{\mathbf{w}}_{low}$. Using Corollary 5.1, the lower key optimization corresponding to the original C-LASSO has the following form:

$$\mathcal{L}^*(\mathbf{g}, \mathbf{h}) = \min_{f_p(\mathbf{w}) \leq 0} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2}\|\mathbf{g}\| - \mathbf{h}^T\mathbf{w} \right\} = \min_{f_p(\mathbf{w}) \leq 0} \mathcal{L}(\mathbf{w}; \mathbf{g}, \mathbf{h}). \tag{10.5}$$

Recall that in both (10.4) and (10.5), $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$. In Lemma 6.1 in Section 6 we solved explicitly for the optimizer $\hat{\mathbf{w}}_{low}$ of problem (10.4). In a similar nature, Lemma 10.1 below identifies a critical property of the optimizer $\mathbf{w}^*_{low}$ of the key optimization (10.5): $\|\mathbf{w}^*\|$ is no larger than $\|\hat{\mathbf{w}}_{low}\|$.

Lemma 10.1. *Let $\mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$ be given and $\|\mathbf{g}\| > dist_{\mathbb{R}^+}(\mathbf{h})$. Denote the minimizer of the problem (10.5) as $\mathbf{w}^*_{low} = \mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h})$. Then,*

$$\frac{\|\mathbf{w}^*_{low}\|^2}{\sigma^2} \leq \frac{dist_{\mathbb{R}^+}(\mathbf{h})^2}{\|\mathbf{g}\|^2 - dist_{\mathbb{R}^+}(\mathbf{h})^2} = \frac{\|\hat{\mathbf{w}}_{low}\|^2}{\sigma^2}. \tag{10.6}$$

For the proof of Lemma 10.1, we require the following result on the tangent cone of the feasible set of (10.5).

Lemma 10.2. *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 $\mathcal{C} = \{\mathbf{w} | f(\mathbf{x}_0 + \mathbf{w}) \leq f(\mathbf{x}_0)\}$. Then, for all $\mathbf{w}^* \in \mathcal{C}$,*

$$\mathcal{T}_{\mathcal{C}}(\mathbf{w}^*)^\circ = \begin{cases} cone(\partial f(\mathbf{x}_0 + \mathbf{w}^*)) & \text{if } f(\mathbf{x}_0 + \mathbf{w}^*) = f(\mathbf{x}_0), \\ \{0\} & \text{if } f(\mathbf{x}_0 + \mathbf{w}^*) < f(\mathbf{x}_0). \end{cases} \tag{10.7}$$

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

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 $\epsilon > 0$ such that $f(\mathbf{x}_0 + \mathbf{w}^* + \epsilon\mathbf{u}) \in \mathcal{C}$. Hence, $\mathcal{T}_{\mathcal{C}}(\mathbf{w}^*) = \text{cone}(\text{Cl}(F_{\mathcal{C}}(\mathbf{w}^*))) = \mathbb{R}^n \implies (\mathcal{T}_{\mathcal{C}}(\mathbf{w}^*))^\circ = \{0\}$ in this case.

Now, assume $f(\mathbf{x}_0 + \mathbf{w}^*) = f(\mathbf{x}_0)$. Then, $F_{\mathcal{C}}(\mathbf{w}^*) = \{\mathbf{u} | f(\mathbf{x}_0 + \mathbf{w}^* + \mathbf{u}) \leq f(\mathbf{x}_0) = f(\mathbf{x}_0 + \mathbf{w}^*)\} = F_{\mathcal{C}'}(\mathbf{x}_0 + \mathbf{w}^*)$, where $F_{\mathcal{C}'}(\mathbf{x}_0 + \mathbf{w}^*)$ denotes the set of feasible directions in $\mathcal{C}' := \{\mathbf{x} | f(\mathbf{x}) \leq f(\mathbf{x}_0 + \mathbf{w}^*)\}$ at $\mathbf{x}_0 + \mathbf{w}^*$. Thus, $\mathcal{T}_{\mathcal{C}}(\mathbf{w}^*) = \mathcal{T}_{\mathcal{C}'}(\mathbf{x}_0 + \mathbf{w}^*) = \text{cone}(\partial f(\mathbf{x}_0 + \mathbf{w}^*))^\circ$, where the last equality follows from Lemma 7.2, 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 of Lemma 10.1. 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 (10.5) is also feasible for (10.4). This implies that $\mathcal{L}^*(\mathbf{g}, \mathbf{h}) \geq \hat{\mathcal{L}}(\mathbf{g}, \mathbf{h})$. Also, from Lemma 6.1, $\hat{\mathcal{L}}(\mathbf{g}, \mathbf{h}) = \sigma\sqrt{\|\mathbf{g}\|^2 - \text{dist}_{\mathbb{R}^+}(\mathbf{h})^2}$. Combining,

$$\mathcal{L}^*(\mathbf{g}, \mathbf{h}) \geq \sigma\sqrt{\|\mathbf{g}\|^2 - \text{dist}_{\mathbb{R}^+}(\mathbf{h})^2} > 0. \tag{10.8}$$

Using the scalarization result of Lemma 6.1 with $\mathcal{C} = \text{cone}(\partial f(\mathbf{x}_0))$, for any $\alpha \geq 0$,

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

Hence, using Lemma F.1 in the appendix shows that, when $\|\mathbf{g}\| > \text{dist}_{\mathbb{R}^+}(\mathbf{h})$,

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

Combining this with (10.8) shows that $\mathcal{L}^*(\mathbf{g}, \mathbf{h})$ is strictly positive, and that $\|\mathbf{w}^*_{low}\|$ and $\mathbf{w}^*_{low}$ is finite.

The minimizer $\mathbf{w}^*_{low}$ satisfies the KKT optimality conditions of (10.5) [82]:

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

or, equivalently,

$$\mathbf{w}^*_{low} = \sigma \frac{\mathbf{h} - \mathbf{s}^*}{\sqrt{\|\mathbf{g}\|^2 - \|\mathbf{h} - \mathbf{s}^*\|^2}}, \tag{10.9}$$

where, from Lemma 10.2,

$$\mathbf{s}^* \in \begin{cases} \text{cone}\left(\partial f(\mathbf{x}_0 + \mathbf{w}^*_{low})\right) & \text{if } f_p(\mathbf{w}_{low}) = 0, \\ \{0\} & \text{if } f_p(\mathbf{w}_{low}) < 0. \end{cases} \tag{10.10}$$

First, consider the scenario in (10.10) where $f_p(\mathbf{w}^*_{low}) < 0$ and $\mathbf{s}^* = 0$. Then, from (10.9) $\mathbf{h} = c_h \mathbf{w}^*_{low}$ for some constant $c_h > 0$. But, from feasibility constraints, $\mathbf{w}^*_{low} \in \mathcal{T}_f(\mathbf{x}_0)$, hence, $\mathbf{h} \in \mathcal{T}_f(\mathbf{x}_0) \implies \mathbf{h} = \text{dist}_{\mathbb{R}^+}(\mathbf{h})$ which implies equality in (10.6).

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 $\|\mathbf{h} - \mathbf{s}^*\| \leq \|\text{dist}_{\mathbb{R}^+}(\mathbf{h})\|$. 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,

$$f(\mathbf{x}_0 + \mathbf{w}^*_{low}) = f(\mathbf{x}_0) \geq f(\mathbf{x}_0 + \mathbf{w}^*_{low}) - \langle \theta \mathbf{s}^*, \mathbf{w}^*_{low} \rangle \implies \langle \mathbf{s}^*, \mathbf{w}^*_{low} \rangle \geq 0. \tag{10.11}$$

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

$$\langle \mathbf{w}^*_{low}, \mathbf{s}_0 \rangle \leq 0. \tag{10.12}$$

Combine (10.11) and (10.12), and further use (10.9) 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,

$$(\text{dist}_{\mathbb{R}^+}(\mathbf{h}))^2 = \|(\mathbf{h} - \mathbf{s}^*) + (\mathbf{s}^* - \mathbf{s}_0)\|^2 \geq \|\mathbf{h} - \mathbf{s}^*\|^2, \tag{10.13}$$

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. □

10.3. Upper Key Optimization

In this section we find a high probability upper bound for $\mathcal{F}^*_c(\mathbf{A}, \mathbf{v})$. Using Corollary 5.2 of Section 5.4.2, application of Gordon's Lemma to the dual of the C-LASSO results in the following key optimization:

$$\mathcal{U}^*(\mathbf{g}, \mathbf{h}) = \max_{\|\mu\| \leq 1} \left\{ \min_{\substack{f_p(\mathbf{w}) \leq 0 \\ \|\mathbf{w}\| \leq C_{up}}} \sqrt{\|\mathbf{w}\|^2 + \sigma^2} \mu^T \mathbf{g} - \|\mu\| \mathbf{h}^T \mathbf{w} \right\}, \tag{10.14}$$

where

$$C_{up} = 2\sqrt{\frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}}.$$

Normalizing the inner terms in (10.14) by $\|\mu\|$ for $\mu \neq 0$, this can be equivalently be written as,

$$\begin{aligned}
\mathcal{U}^*(\mathbf{g}, \mathbf{h}) &= \max_{\|\mu\| \leq 1} \left\{ \|\mu\| \min_{\substack{f_p(\mathbf{w}) \leq 0 \\ \|\mathbf{w}\| \leq C_{up}}} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2} \|\mathbf{g}\| - \mathbf{h}^T \mathbf{w} \right\} \right\} \\
&= \max \left\{ 0, \min_{\substack{f_p(\mathbf{w}) \leq 0 \\ \|\mathbf{w}\| \leq C_{up}}} \mathcal{L}(\mathbf{w}; \mathbf{g}, \mathbf{h}) \right\} \\
&= \max \left\{ 0, \ \mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h}) \right\},
\end{aligned} \tag{10.15}$$

where we additionally defined

$$\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h}) := \min_{\substack{f_p(\mathbf{w}) \leq 0 \\ \|\mathbf{w}\| \leq C_{up}}} \mathcal{L}(\mathbf{w}; \mathbf{g}, \mathbf{h}). \tag{10.16}$$

Observe the similarity of the upper key optimization (10.15) to the lower key optimization (10.5). The next lemma proves that $\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ and $\mathcal{U}^*(\mathbf{g}, \mathbf{h})$ are Lipschitz functions.

Lemma 10.3 (Lipschitzness of $\mathcal{U}^*(\mathbf{g}, \mathbf{h})$)**.** *$\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ and, consequently, $\mathcal{U}^*(\mathbf{g}, \mathbf{h})$ are Lipschitz with Lipschitz constants at most $2\sigma\sqrt{C^2_{up} + 1}$.*

Proof. First, we prove that $\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ 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 (10.16). W.l.o.g., assume that $\mathcal{L}^*_{up}(\mathbf{g}_1, \mathbf{h}_1) \geq \mathcal{L}^*_{up}(\mathbf{g}_2, \mathbf{h}_2)$. Then,

$$\begin{aligned}
\mathcal{L}^*_{up}(\mathbf{g}_1, \mathbf{h}_1) - \mathcal{L}^*_{up}(\mathbf{g}_2, \mathbf{h}_2) &= \mathcal{L}(\mathbf{w}_1; \mathbf{g}_1, \mathbf{h}_2) - \mathcal{L}(\mathbf{w}_2; \mathbf{g}_2, \mathbf{h}_2) \\
&\leq \mathcal{L}(\mathbf{w}_2; \mathbf{g}_1, \mathbf{h}_1) - \mathcal{L}(\mathbf{w}_2; \mathbf{g}_2, \mathbf{h}_2) \\
&= \sqrt{\|\mathbf{w}_2\|^2 + \sigma^2}(\|\mathbf{g}_1\| - \|\mathbf{g}_2\|) - (\mathbf{h}_1 - \mathbf{h}_2)^T \mathbf{w}_2 \\
&\leq \sqrt{\sigma^2 C^2_{up} + \sigma^2} \|\mathbf{g}_1 - \mathbf{g}_2\| + \|\mathbf{h}_1 - \mathbf{h}_2\| \sigma C_{up},
\end{aligned} \tag{10.17}$$

where, we have used the fact that $\|\mathbf{w}_2\| \leq \sigma C_{up}$. From (10.17), it follows that $\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ is indeed Lipschitz and

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

To prove that $\mathcal{U}^*(\mathbf{g}, \mathbf{h})$ is Lipschitz with the same constant, assume w.l.o.g that $\mathcal{U}^*(\mathbf{g}_1, \mathbf{h}_1) \geq \mathcal{U}^*(\mathbf{g}_2, \mathbf{h}_2)$. Then, from (10.15),

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

□

10.4. Matching Lower and Upper key Optimizations

Comparing (10.5) to (10.15), 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 $\mathbb{E}[\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})]$.

Lemma 10.4. *Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m), \mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ and independently generated. Assume $(1 - \epsilon_0)m \geq \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) \geq \epsilon_0 m$ for some constant $\epsilon_0 > 0$ and m sufficiently large. For any $\epsilon > 0$, there exists $c > 0$ such that, with probability $1 - \exp(-cm)$, we have,*

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

In Lemma 10.3 we proved that $\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ is Lipschitz. Gaussian concentration of Lipschitz functions (see Lemma A.4) implies, then, that $\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ concentrates with high probability around its mean $\mathbb{E}[\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})]$. According to Lemma 10.4, under certain conditions implied by its assumptions, $\mathcal{U}^*(\mathbf{g}, \mathbf{h})$ and $\mathcal{L}^*(\mathbf{g}, \mathbf{h})$ also concentrate around the same quantity $\mathbb{E}[\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})]$. The way to prove this fact is by showing that when these conditions hold, $\mathcal{U}^*(\mathbf{g}, \mathbf{h})$ and $\mathcal{L}^*(\mathbf{g}, \mathbf{h})$ are equal to $\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})$ with high probability. Once we have shown that, we require the following result to complete the proof.

Lemma 10.5. *Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$. Assume f_1 is L-Lipschitz and, $\mathbb{P}(f_1(\mathbf{g}) = f_2(\mathbf{g})) > 1 - \epsilon$. Then, for all $t > 0$,*

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

Proof. From standard concentration result on Lipschitz functions (see Lemma A.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 $1 - \epsilon$. Combine those facts to complete the proof as follows,

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

□

Now, we complete the proof of Lemma 10.4 using the result of Lemma 10.5.

Proof of Lemma 10.4. We prove the two statements of the lemma in the order they appear.
1. First, we prove that under the assumptions of the lemma, $\mathcal{U}^* = \mathcal{L}^*_{up}$ w.h.p.. By (10.15), it suffices to show that $\mathcal{L}^*_{up} \geq 0$ w.h.p.. Constraining the feasible set of a minimization problem cannot result in a decrease in its optimal cost, hence,

$$\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h}) \geq \mathcal{L}^*(\mathbf{g}, \mathbf{h}) \geq \hat{\mathcal{L}}(\mathbf{g}, \mathbf{h}). \tag{10.18}$$

where recall $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$ is the lower key optimization of the approximated C-LASSO (see (10.4)). From Lemma 6.1, since $m \geq \mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+) + \epsilon_0 m$, we have that

$$\hat{\mathcal{L}}(\mathbf{g},\mathbf{h}) \geq (1-\epsilon)\sigma\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)} \geq 0,$$

with $1-\exp(-\mathcal{O}(m))$. Combine this with (10.18) to find that $\mathcal{L}^*_{up}(\mathbf{g},\mathbf{h}) \geq 0$ or $\mathcal{U}^* = \mathcal{L}^*_{up}$ with probability $1-\exp(-\mathcal{O}(m))$. Furthermore, from Lemma 10.3, $\mathcal{L}^*_{up}(\mathbf{g},\mathbf{h})$ is Lipschitz with constant $L = 2\sigma\sqrt{C_{up}^2+1}$. We now apply Lemma 10.5 setting $f_1 = \mathcal{L}^*_{up}(\mathbf{g},\mathbf{h})$, $f_2 = \mathcal{U}^*$ and $t = \epsilon\sigma\sqrt{m}$, to find that

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

with probability $1-\exp(-\mathcal{O}(m))$. In writing the exponent in the probability as $\mathcal{O}(m)$, we made use of the fact that $C_{up} = 2\sqrt{\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}}$ is bounded below by a constant, since $(1-\epsilon_0)m \geq \mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+) \geq \epsilon_0 m$.

2. As in the first statement, we apply Lemma 10.5, this time setting $f_1 = \mathcal{L}^*_{up}$, $f_2 = \mathcal{L}^*$ and $t = \epsilon\sigma\sqrt{m}$. The result is immediate after application of the lemma, but first we need to show that $\mathcal{L}^*(\mathbf{g},\mathbf{h}) = \mathcal{L}^*_{up}(\mathbf{g},\mathbf{h})$ w.h.p.. We will show equivalently that the minimizer $\mathbf{w}^*_{low}$ of (10.5) satisfies $\mathbf{w}^*_{low} \in S_{up}$. From Lemma 10.1, $\|\mathbf{w}^*_{low}\| \leq \frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\|\mathbf{g}\|-\text{dist}_{\mathbb{R}^+}(\mathbf{h})}$. On the other hand, using standard concentration arguments (Lemma B.2), with probability $1-\exp(-\mathcal{O}(m))$, $\frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\|\mathbf{g}\|-\text{dist}_{\mathbb{R}^+}(\mathbf{h})} \leq \frac{2\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)} = C_{up}$. Combining these completes the proof. □

10.5. Deviation Bound

Resembling the approach developed in Section 6, we show that if we restrict the norm of the error vector $\|\mathbf{w}\|$ in (10.3) as follows

$$\|\mathbf{w}\| \in S_{dev} := \left\{\ell \middle| \ell \geq (1+\epsilon_{dev})\sigma\sqrt{\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}}\right\}, \tag{10.19}$$

then, this results in a significant increase in the cost of C-LASSO. To lower bound the deviated cost, we apply Corollary 5.3 of Section 5.4.3 to the restricted original C-LASSO, which yields the following key optimization

$$\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) = \min_{\substack{f_p(\mathbf{w})\leq 0 \\ \|\mathbf{w}\|\in S_{dev}}} \mathcal{L}(\mathbf{w};\mathbf{g},\mathbf{h}). \tag{10.20}$$

Lemma 10.6. *Let $\mathbf{g} \sim \mathcal{N}(0,\mathbf{I}_m)$, $\mathbf{h} \sim \mathcal{N}(0,\mathbf{I}_n)$. Assume $(1-\epsilon_L)m > \mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+) > \epsilon_L m$ and m is sufficiently large. Then, there exists a constant $\delta_{dev} = \delta_{dev}(\epsilon_{dev}) > 0$ such that, with probability $1-\exp(-\mathcal{O}(m))$, we have,*

$$\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathbb{E}[\mathcal{L}^*_{up}(\mathbf{g},\mathbf{h})] \geq \sigma\delta_{dev}\sqrt{m}. \tag{10.21}$$

As common, our analysis begins with a deterministic result, which builds towards the proof of the probabilistic statement in Lemma 10.6.

10.5.1 Deterministic Result

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

$$\eta_d := \sigma\sqrt{\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}},$$

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

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

Also, recall the definition of the scalar function $L(\alpha;a,b)$ in (10.2).

Lemma 10.7. *Let $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{h} \in \mathbb{R}^n$ be such that $\|\mathbf{g}\| > dist_{\mathbb{R}^+}(\mathbf{h})$ and $\eta_s(\mathbf{g},\mathbf{h}) \leq (1+\epsilon_{dev})\eta_d$. Then,*

$$\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathcal{L}^*(\mathbf{g},\mathbf{h}) \geq L\left((1+\epsilon_{dev})\eta_d; \|\mathbf{g}\|, dist_{\mathbb{R}^+}(\mathbf{h})\right) - L\left(\eta_s(\mathbf{g},\mathbf{h}); \|\mathbf{g}\|, dist_{\mathbb{R}^+}(\mathbf{h})\right) \tag{10.22}$$

Proof. First assume that $\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) = \infty$. Since $\mathcal{L}^*(\mathbf{g},\mathbf{h}) \leq \hat{\mathcal{L}}(\mathbf{0};\mathbf{g},\mathbf{h}) = \sigma\|\mathbf{g}\|$ and the right hand side of (10.22) is finite, we can easily conclude with the desired result.

Hence, in the following assume that $\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) < \infty$ and denote $\mathbf{w}^*_{dev}$ the minimizer of the restricted problem (10.20). From feasibility constraints, we have $f_p(\mathbf{w}_{dev}) \leq 0$ and $\|\mathbf{w}^*_{dev}\| \in S_{dev}$. Define $\bar{\mathbf{w}}_{dev} = c\mathbf{w}^*_{dev}$ where $c := \frac{\eta_s}{\|\mathbf{w}^*_{dev}\|}$. Notice, $\|\mathbf{w}^*_{dev}\| \geq (1+\epsilon_{dev})\eta_d \geq \eta_s(\mathbf{g},\mathbf{h})$, thus, $c \leq 1$. Then, from convexity of $f(\cdot)$,

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

This shows that $\bar{\mathbf{w}}_{dev}$ is feasible for the minimization (10.5). Hence,

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

Starting with this, we write,

$$\begin{aligned}
\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) - \mathcal{L}^*(\mathbf{g},\mathbf{h}) &\geq \mathcal{L}(\mathbf{w}^*_{dev};\mathbf{g},\mathbf{h}) - \mathcal{L}(\bar{\mathbf{w}}_{dev};\mathbf{g},\mathbf{h}) \\
&= (\sqrt{\|\mathbf{w}^*_{dev}\|^2+\sigma^2} - \sqrt{\|\bar{\mathbf{w}}_{dev}\|^2+\sigma^2})\|\mathbf{g}\| - \mathbf{h}^T(\mathbf{w}^*_{dev} - \bar{\mathbf{w}}_{dev}) \\
&= (\sqrt{\|\mathbf{w}^*_{dev}\|^2+\sigma^2} - \sqrt{\|\bar{\mathbf{w}}_{dev}\|^2+\sigma^2})\|\mathbf{g}\| - (1-c)\mathbf{h}^T\mathbf{w}^*_{dev}.
\end{aligned} \tag{10.23}$$

Since, $f_p(\mathbf{w}^*_{dev}) \leq 0$, $\mathbf{w}^*_{dev} \in \mathcal{T}_f(\mathbf{x}_0)$. Hence, and using Moreau's decomposition Theorem (see Fact A.1), we have

$$\begin{aligned}
\mathbf{h}^T\mathbf{w}^*_{dev} &= \left\langle \text{Proj}(\mathbf{h}, \mathcal{T}_f(\mathbf{x}_0)), \mathbf{w}^*_{dev} \right\rangle + \underbrace{\left\langle \text{Proj}(\mathbf{h}, \left(\mathcal{T}_f(\mathbf{x}_0)\right)^\circ), \mathbf{w}^*_{dev} \right\rangle}_{\leq 0} \\
&\leq \text{dist}_{\mathbb{R}^+}(\mathbf{h})\|\mathbf{w}^*_{dev}\|.
\end{aligned} \tag{10.24}$$

Use (10.24) in (10.23), to write

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

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

10.5.2 Probabilistic result

We now prove the main result of the section, Lemma 10.6.

Proof of Lemma 10.6. The proof is based on the results of Lemma 10.7. First, we show that under the assumptions of Lemma 10.6, the assumptions of Lemma 10.7 hold w.h.p.. In this direction, using standard concentration arguments provided in Lemmas B.5 and B.3, we find that,

1. $\|\mathbf{g}\| \geq \text{dist}_{\mathbb{R}^+}(\mathbf{h})$,
2. $\frac{\text{dist}_{\mathbb{R}^+}(\mathbf{h})}{\sqrt{\|\mathbf{g}\|^2 - \text{dist}_{\mathbb{R}^+}(\mathbf{h})^2}} \leq (1+\epsilon_{dev})\frac{m}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}$.

3. For any constant $\epsilon > 0$,

$$|\|\mathbf{g}\|^2 - m| \leq \epsilon m \quad \text{and} \quad |\left(\text{dist}_{\mathbb{R}^+}(\mathbf{h})\right)^2 - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)| < \epsilon m, \tag{10.25}$$

all with probability $1 - \exp\left(-\mathcal{O}\left(m\right)\right)$. It follows from the first two statements that Lemma 10.7 is applicable and we can use (10.22). Thus, it suffinces to find a lower bound for the right hand side of (10.22).

Lemma F.1 in the Appendix analyzes in detail many properties of the scalar function $L(\alpha; a, b)$, which appears in (10.22). Here, we use the sixth statement of that Lemma (in a similar manner to the proof of Lemma 6.3). In particular, apply Lemma F.1 with the following mapping:

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

Application of the lemma is valid since (10.25) is true, and gives that with probability $1 - \exp\left(-\mathcal{O}\left(m\right)\right)$,

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

for some constant δ_{dev}. Combining this with Lemma 10.7, we may conclude

$$\mathcal{L}^*_{dev}(\mathbf{g}, \mathbf{h}) - \mathcal{L}^*(\mathbf{g}, \mathbf{h}) \geq 2\sigma\delta_{dev}\sqrt{m}. \tag{10.26}$$

On the other hand, from Lemma 10.4,

$$|\mathcal{L}^*(\mathbf{g}, \mathbf{h}) - \mathbb{E}[\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})]| \leq \sigma\delta_{dev}\sqrt{m} \tag{10.27}$$

with the desired probability. Union bounding over (10.26) and (10.27), we conclude with the desired result. □

10.6. 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 (1.5) of the original C-LASSO problem and also the definition of the set S_{dev} in (10.19).

Lemma 10.8. *Assume there exists a constant ϵ_L such that, $(1-\epsilon_L)m \geq \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) \geq \epsilon_L m$. Further assume, m is sufficiently large. The following hold:*

1. *For any $\epsilon_{up} > 0$, there exists $c_{up} > 0$ such that, with probability $1 - \exp(-c_{up}m)$, we have,*

$$\mathcal{F}^*_c(\mathbf{A}, \mathbf{v}) \leq \mathbb{E}[\mathcal{L}^*_{up}(f, \mathbf{g}, \mathbf{h})] + \epsilon_{up}\sigma\sqrt{m} \tag{10.28}$$

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,*

$$\min_{\|\mathbf{w}\| \in S_{dev},\ f_p(\mathbf{w}) \leq 0} \mathcal{F}(\mathbf{w}; \mathbf{A}, \mathbf{v}) \geq \mathbb{E}[\mathcal{L}^*_{up}(f, \mathbf{g}, \mathbf{h})] + \delta_{dev}\sigma\sqrt{m} \tag{10.29}$$

3. *For any $\epsilon_{dev} > 0$, there exists $c > 0$ such that, with probability $1 - \exp(-cm)$,*

$$\|\mathbf{x}^*_c - \mathbf{x}_0\|^2 \leq \sigma^2(1+\epsilon_{dev})\frac{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}{m - \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)}.$$

Proof. We prove the statements of the lemma in the order that they appear.
1. For notational simplicity denote $\xi = \mathbb{E}[\mathcal{L}^*_{up}(\mathbf{g}, \mathbf{h})]$. We combine second statement of Lemma 5.2 with Lemma 10.4. For any constant ϵ_{up}, we have,

$$\begin{aligned}
\mathbb{P}(\mathcal{F}^*_c(\mathbf{A}, \mathbf{v}) \leq \xi + 2\sigma\epsilon_{up}\sqrt{m}) &\geq 2\mathbb{P}(\mathcal{U}^*(\mathbf{g}, \mathbf{h}) + \sigma\epsilon\sqrt{m} \leq \xi + 2\sigma\epsilon_{up}\sqrt{m}) - 1 - \exp(-\mathcal{O}\left(m\right)) \\
&= 2\mathbb{P}(\mathcal{U}^*(\mathbf{g}, \mathbf{h}) \leq \xi + \sigma\epsilon_{up}\sqrt{m}) - 1 - \exp(-\mathcal{O}\left(m\right)) \\
&\geq 1 - \exp(-\mathcal{O}\left(m\right)),
\end{aligned}$$

where we used the first statement of Lemma (10.4) to lower bound the $\mathbb{P}(\mathcal{U}^*(\mathbf{g},\mathbf{h}) \leq \xi + \sigma\epsilon_{up}\sqrt{m})$.
2. Pick a small constant $\epsilon > 0$ satisfying $\epsilon < \frac{\delta_{dev}}{2}$ in the third statement of Lemma 5.2. Now, using Lemma 10.6 and this choice of ϵ, with probability $1 - \exp(-\mathcal{O}(m))$, we have,

$$\mathbb{P}(\min_{\mathbf{w}\in S_{dev},\, f_p(\mathbf{w})\leq 0} \mathcal{F}(\mathbf{w};\mathbf{A},\mathbf{v}) \geq \xi + \sigma\frac{\delta_{dev}}{2}\sqrt{m}) \geq 2\mathbb{P}(\mathcal{L}^*_{dev}(\mathbf{g},\mathbf{h}) \geq \xi + \sigma\delta_{dev}\sqrt{m} - \epsilon\sigma\sqrt{m}) - 1 - \exp(-\mathcal{O}(m))$$
$$\geq 1 - \exp(-\mathcal{O}(m)),$$

where we used (10.21) of Lemma 10.6.
3. Apply Statements 1. and 2. of the lemma, choosing $\epsilon_{up} = \frac{\delta_{dev}}{8}$. Union bounding we find that

$$\mathbb{P}(\min_{\mathbf{w}\in S_{dev},\, f_p(\mathbf{w})\leq 0} \mathcal{F}(\mathbf{w};\mathbf{A},\mathbf{v}) \geq \mathcal{F}^*_c(\mathbf{A},\mathbf{v}) + \sigma\frac{\delta_{dev}}{4}) \geq 1 - \exp(-\mathcal{O}(m)),$$

which implies with the same probability $\|\mathbf{w}^*_c\| \notin S_{dev}$, i.e., $\|\mathbf{w}^*_c\| \leq (1+\epsilon_{dev})\sigma\sqrt{\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}}$.

□

11. ℓ_2^2-LASSO

As we have discussed throughout our main results, one of the critical contributions of this paper 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 Gordon's Lemma.

11.1. 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 (3.4). The following lemma gives a simple condition for ℓ_2-LASSO and ℓ_2^2-LASSO to have the same solution.

Lemma 11.1. *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 $\tau = \lambda \cdot \frac{\|\mathbf{A}\mathbf{x}^*_{\ell_2}-\mathbf{y}\|}{\sigma}$.*

Proof. The optimality condition for the ℓ_2^2-LASSO problem (1.6), implies the existence of $\mathbf{s}_{\ell_2} \in \partial f(\mathbf{x}^*_{\ell_2})$ such that,

$$\lambda\mathbf{s}_{\ell_2} + \frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}^*_{\ell_2} - \mathbf{y})}{\|\mathbf{A}\mathbf{x}^*_{\ell_2} - \mathbf{y}\|} = 0 \tag{11.1}$$

On the other hand, from the optimality conditions of (1.7), $\mathbf{x}$ is a minimizer of the ℓ_2^2-LASSO if there exists $\mathbf{s} \in \partial f(\mathbf{x})$ such that,

$$\sigma\tau\mathbf{s} + \mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{y}) = 0. \tag{11.2}$$

Observe that, for $\tau = \lambda \cdot \frac{\|\mathbf{A}\mathbf{x}^*_{\ell_2}-\mathbf{y}\|}{\sigma}$, using (11.1), $\mathbf{x}^*_{\ell_2}$ satisfies (11.2) and is thus a minimizer of the ℓ_2^2-LASSO. □

In order to evaluate the mapping function as proposed in Lemma 11.1, we need to estimate $\|\mathbf{y} - \mathbf{A}\mathbf{x}^*_{\ell_2}\|$. We do this relying again on the approximated ℓ_2-LASSO problem in (5.5). 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 (5.5) and Lemma 6.4,

$$\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}^*_{\ell_2}\| = \hat{\mathcal{F}}^*_{\ell_2}(\mathbf{A},\mathbf{v}) - \lambda\hat{f}_p(\hat{\mathbf{w}}^*_{\ell_2})$$
$$\approx \sigma\sqrt{m - \mathbf{D}_f(\mathbf{x}_0,\lambda)} - \lambda\hat{f}_p(\hat{\mathbf{w}}^*_{\ell_2}). \tag{11.3}$$

Arguing that,

$$\lambda \hat{f}_p(\mathbf{w}^*_{\ell_2}) \approx \sigma \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}, \tag{11.4}$$

and substituting this in (11.3) will result in the desired mapping formula given in (3.4).

In the remaining lines we provide justification supporting our belief that (11.4) is true. Not surprisingly at this point, the core of our argument relies on application of Gordon's Lemma. Following the lines of our discussion in Section 6, we use the minimizer $\mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h})$ of the simple optimization (2.5) 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 11.2 below shows that

$$\begin{aligned}\lambda \hat{f}_p(\mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h})) &= \sigma \frac{\langle \Pi(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)), \text{Proj}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \rangle}{\sqrt{\|\mathbf{g}\|^2 - \text{dist}(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))^2}} \\ &\approx \sigma \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}},\end{aligned}$$

where the second (approximate) equality follows via standard concentration inequalities.

Lemma 11.2. *Assume $(1 - \epsilon_L)m \geq \mathbf{D}_f(\mathbf{x}_0, \lambda)$ and m is sufficiently large. Then, for any constant $\epsilon > 0$, with probability $1 - \exp(-\mathcal{O}\left(\min\{m, \frac{m^2}{n}\}\right))$,*

$$\left| \lambda \hat{f}_p(\mathbf{w}^*_{low}) - \sigma \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}} \right| < \epsilon \sqrt{m}. \tag{11.5}$$

Proof. Recall that $\mathbf{w}^*_{low}(\mathbf{g}, \mathbf{h}) = \sigma \frac{\Pi(\mathbf{h}, \mathcal{C})}{\sqrt{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0))}}$ for $\mathcal{C} = \lambda \partial f(\mathbf{x}_0)$. Combining this with Fact A.2, we obtain,

$$\hat{f}_p(\mathbf{w}_{low}) = \max_{\mathbf{s} \in \mathcal{C}} \langle \mathbf{w}_{low}, \mathbf{s} \rangle = \frac{\langle \Pi(\mathbf{h}, \mathcal{C}), \text{Proj}(\mathbf{h}, \mathcal{C}) \rangle}{\sqrt{\|\mathbf{g}\|^2 - \text{dist}(\mathbf{h}, \mathcal{C})^2}}.$$

What remains is to show the right hand side concentrates around $\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}$ with the desired probability. Fix a constant $\epsilon > 0$. Consider the denominator. Using Lemma B.5, with probability $1 - \exp(-\mathcal{O}(m))$,

$$\left| \frac{\sqrt{\|\mathbf{g}\|^2 - \text{dist}(\mathbf{h}, \mathcal{C})^2}}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}} - 1 \right| < \epsilon. \tag{11.6}$$

We now apply Lemma B.3 for $\mathbf{C}(\mathcal{C})$ where we choose $t = \frac{m}{\sqrt{\max\{m,n\}}}$ and use the fact that $m > \mathbf{D}(\mathcal{C})$. Then, with probability $1 - \exp(-\mathcal{O}\left(\min\{m, \frac{m^2}{n}\}\right))$, we have,

$$|\text{corr}(\mathbf{h}, \mathcal{C}) - \mathbf{C}(\mathcal{C})| \leq \epsilon m.$$

Combining this with (11.6) choosing $\epsilon > 0$, sufficiently small (according to ϵ_L), we find (11.5) with the desired probability. □

The lemma above shows that, $\lambda \hat{f}_p(\mathbf{w}^*_{low})$ is around $\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}$ 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 $\hat{\mathcal{L}}^*_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\mathbf{w} \in S_{dev}} \hat{\mathcal{L}}(\mathbf{w}; \mathbf{g}, \mathbf{h})$ where we modify the set S_{dev} to,

$$S_{dev} = \{\mathbf{w} \big| \left| \frac{\lambda \hat{f}_p(\mathbf{w})}{\sigma} - \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}} \right| > \epsilon_{dev} \sqrt{m} \}.$$

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 $\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h})$. While this approach may be promising, we believe it is more challenging than our ℓ_2 norm analysis of $\|\mathbf{w}^*_{\ell_2}\|$ and it will not be topic of this paper.

The next section shows that there exists a one-to-one (monotone) mapping of the region $\mathcal{R}_{ON}$ to the entire possible regime of penalty parameters of the ℓ_2^2-LASSO.

11.2. Properties of map(λ)

The following result shows that $\mathbf{P}(\lambda\mathcal{C}), \mathbf{D}(\lambda\mathcal{C}), \mathbf{C}(\lambda\mathcal{C})$ (see (6.3)) are Lipschitz continuous and will be useful for the consequent discussion. The proof can be found in Appendix B.

Lemma 11.3. *Let $\mathcal{C}$ 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 $\max_{\mathbf{s}\in\mathcal{C}} \|\mathbf{s}\| = R$. Then, viewing $\mathbf{P}(\lambda\mathcal{C}), \mathbf{D}(\lambda\mathcal{C}), \mathbf{C}(\lambda\mathcal{C})$ 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 $\mathcal{R}_{\text{ON}}$ from Definition 8.1.

Proposition 11.1. *Assume $m > \mathbf{D}_f(\mathbf{x}_0, \lambda_{best})$. Recall that $\mathcal{R}_{ON} = (\lambda_{crit}, \lambda_{\max})$. calib$(\lambda) = \frac{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)-\mathbf{C}_f(\mathbf{x}_0,\lambda)}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}}$ and map$(\lambda) = \lambda \cdot$ calib(λ) have the following properties over $\{\lambda_{crit}\} \cup \mathcal{R}_{ON} \to \{0\} \cup \mathbb{R}^+$.*

- *calib(λ) is a nonnegative, increasing and continuous function over $\{\lambda_{crit}\} \cup \mathcal{R}_{ON}$.*
- *map(λ) is nonnegative, strictly increasing and continuous at all $\lambda \in \{\lambda_{crit}\} \cup \mathcal{R}_{ON}$.*
- *map$(\lambda_{crit}) = 0$. $\lim_{\lambda\to\lambda_{\max}}$ map$(\lambda) = \infty$. Hence, map$(\lambda) : \{\lambda_{crit}\} \cup \mathcal{R}_{ON} \to \{0\} \cup \mathbb{R}^+$ is bijective.*

Proof. Proof of the first statement: Assume $\lambda \in \mathcal{R}_{\text{ON}}$, from Lemma 8.4, $m > \max\{\mathbf{D}_f(\mathbf{x}_0, \lambda), \mathbf{D}_f(\mathbf{x}_0, \lambda) + \mathbf{C}_f(\mathbf{x}_0, \lambda)\}$ and $\lambda > 0$. Hence, calib(λ) is strictly positive over $\lambda \in \mathcal{R}_{\text{ON}}$. Recall that,

$$\text{calib}(\lambda) = \frac{m - \mathbf{D}_f(\mathbf{x}_0, \lambda) - \mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}} = \sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)} - \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}.$$

Let $h > 0$. We will investigate the change in calib(λ) by considering calib$(\lambda + h) -$ calib(λ) as $h \to 0^+$. Since $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ is differentiable, $\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}$ is differentiable as well and gives,

$$\frac{\partial\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}{\partial\lambda} = \frac{-\mathbf{D}_f(\mathbf{x}_0, \lambda)'}{2\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}. \tag{11.7}$$

For the second term, consider the following,

$$\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda + h)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda + h)}} - \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}} = h[E_1(\lambda, h) + E_2(\lambda, h)],$$

where,

$$E_1(\lambda, h) = \frac{1}{h}\left[\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda + h)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda + h)}} - \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda + h)}}\right],$$
$$E_2(\lambda, h) = \frac{1}{h}\left[\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda + h)}} - \frac{\mathbf{C}_f(\mathbf{x}_0, \lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}}\right].$$

As $h \to 0^+$, we have,

$$\lim_{h\to 0^+} E_2(\lambda,h) = \mathbf{C}_f(\mathbf{x}_0,\lambda)\frac{\partial \frac{1}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}}}{\partial\lambda} = \frac{\mathbf{C}_f(\mathbf{x}_0,\lambda)\mathbf{D}_f(\mathbf{x}_0,\lambda)'}{2(m-\mathbf{D}_f(\mathbf{x}_0,\lambda))^{3/2}} \leq 0, \tag{11.8}$$

since $\operatorname{sgn}(\mathbf{C}_f(\mathbf{x}_0,\lambda)) = -\operatorname{sgn}(\mathbf{D}_f(\mathbf{x}_0,\lambda)')$.

Fix arbitrary $\epsilon_D > 0$ and let $R = \sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \|\mathbf{s}\|$. Using continuity of $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ and Lemma 11.3, choose h sufficiently small to ensure,

$$\left|\frac{1}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}} - \frac{1}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda+h)}}\right| < \epsilon_D, \quad |\mathbf{C}_f(\mathbf{x}_0,\lambda+h) - \mathbf{C}_f(\mathbf{x}_0,\lambda)| < 3R(\sqrt{n}+\lambda R)h.$$

We then have,

$$E_1(\lambda,h) \leq \frac{\mathbf{C}_f(\mathbf{x}_0,\lambda+h)-\mathbf{C}_f(\mathbf{x}_0,\lambda)}{h}\frac{1}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}} + 3\epsilon_D R(\sqrt{n}+\lambda R). \tag{11.9}$$

Denote $\frac{\mathbf{C}_f(\mathbf{x}_0,\lambda+h)-\mathbf{C}_f(\mathbf{x}_0,\lambda)}{h}, \frac{\mathbf{D}_f(\mathbf{x}_0,\lambda+h)-\mathbf{D}_f(\mathbf{x}_0,\lambda)}{h}$ by $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{D}}$. Combining (11.8), (11.9) and (11.7), for sufficiently small h, we find,

$$\limsup_{h\to 0^+} \frac{\operatorname{calib}(\lambda+h)-\operatorname{calib}(\lambda)}{h} = \limsup_{h\to 0}\Big[\frac{-\tilde{\mathbf{D}}}{2\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}} - \frac{\tilde{\mathbf{C}}}{\sqrt{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}} - \frac{\mathbf{C}_f(\mathbf{x}_0,\lambda)\mathbf{D}_f(\mathbf{x}_0,\lambda)'}{2(m-\mathbf{D}_f(\mathbf{x}_0,\lambda))^{3/2}} + 3\epsilon_D R(\sqrt{n}+\lambda R)\Big].$$

We can let ϵ_D go to 0 as $h \to 0^+$ and $-\tilde{\mathbf{D}} - 2\tilde{\mathbf{C}}$ is always nonnegative as $\mathbf{P}_f(\mathbf{x}_0,\lambda)$ is nondecreasing due to Lemma 8.1. Hence, the right hand side is nonnegative. Observe that the increase is strict for $\lambda \neq \lambda_{\text{best}}$, as we have $\mathbf{C}_f(\mathbf{x}_0,\lambda)\mathbf{D}_f(\mathbf{x}_0,\lambda)' > 0$ whenever $\lambda \neq \lambda_{\text{best}}$ due to the fact that $\mathbf{D}_f(\mathbf{x}_0,\lambda)'$ (and $\mathbf{C}_f(\mathbf{x}_0,\lambda)$) 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 $\operatorname{calib}(\lambda)$ is continuous for all $\lambda \in \{\lambda_{\text{crit}}\} \cup \mathcal{R}_{\text{ON}}$ (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 $\operatorname{calib}(\lambda)$, observe that, for any $\lambda \in \mathcal{R}_{\text{ON}} \cup \{\lambda_{\text{crit}}\}$, $m - \mathbf{D}_f(\mathbf{x}_0,\lambda) > 0$ and from Lemma 11.3, $\mathbf{D}_f(\mathbf{x}_0,\lambda), \mathbf{C}_f(\mathbf{x}_0,\lambda)$ are continuous functions which ensures continuity of $m - \mathbf{D}_f(\mathbf{x}_0,\lambda) - \mathbf{C}_f(\mathbf{x}_0,\lambda)$ and $m - \mathbf{D}_f(\mathbf{x}_0,\lambda)$. Hence, $\operatorname{calib}(\lambda)$ is continuous as well.

Proof of the second statement: Since $\operatorname{calib}(\lambda)$ is strictly increasing on $\mathcal{R}_{\text{ON}}$, $\lambda \cdot \operatorname{calib}(\lambda)$ is strictly increasing over $\mathcal{R}_{\text{ON}}$ as well. Increase at $\lambda = \lambda_{\text{crit}}$ follows from the fact that $\operatorname{map}(\lambda_{\text{crit}}) = 0$ (see next statement). Since $\operatorname{calib}(\lambda)$ is continuous, $\lambda \cdot \operatorname{calib}(\lambda)$ is continuous as well.

Proof of the third statement: From Lemma 8.3, if $\operatorname{calib}(\lambda_{\text{crit}}) > 0$, $\lambda_{\text{crit}} = 0$ hence $\operatorname{map}(\lambda_{\text{crit}}) = 0$. If $\operatorname{calib}(\lambda_{\text{crit}}) = 0$, then $\operatorname{map}(\lambda_{\text{crit}}) = \lambda_{\text{crit}} \cdot \operatorname{calib}(\lambda_{\text{crit}}) = 0$. In any case, $\operatorname{map}(\lambda_{\text{crit}}) = 0$. 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, $\operatorname{calib}(\lambda) \to \infty$. This ensures $\operatorname{map}(\lambda) \to \infty$ as well. Since $\operatorname{map}(\lambda)$ is continuous and strictly increasing and achieves the values 0 and ∞, it maps $\{\lambda_{\text{crit}}\} \cup \mathcal{R}_{\text{ON}}$ to $\{0\} \cup \mathbb{R}^+$ bijectively. □

11.3. On the stability of ℓ_2^2-LASSO

As it has been discussed in Section 11.2 in detail, $\operatorname{map}(\cdot)$ takes the interval $[\lambda_{\text{crit}}, \lambda_{\max})$ to $[0,\infty)$ and Theorem 3.2 gives tight stability guarantees for $\lambda \in \mathcal{R}_{\text{ON}}$. 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 11.4. *Consider the ℓ_2^2-LASSO problem (1.7). 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 $(1-\epsilon_L)m \geq \mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)$ for a constant $\epsilon_L > 0$ and m is sufficiently large. Then, there exists a number $C > 0$ independent of σ, such that, with probability $1 - \exp(-\mathcal{O}(m))$,*

$$\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|^2}{\sigma^2} \leq C. \tag{11.10}$$

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,

$$\mathcal{T}_f(\mathbf{x}_0, \epsilon_0) = \mathrm{Cl}(\{\alpha \cdot \mathbf{w} \big| f(\mathbf{x}_0 + \mathbf{w}) \leq f(\mathbf{x}_0) + \epsilon_0 \|\mathbf{w}\|,\ \alpha \geq 0\}).$$

Appendix I investigates basic properties of this set. In particular, we will make use of Lemma I.2. We can choose sufficiently small numbers $\epsilon_0, \epsilon_1 > 0$ (independent of σ) such that,

$$\min_{\mathbf{w} \in \mathcal{T}_f(\mathbf{x}_0, \epsilon_0), \|\mathbf{w}\|=1} \|\mathbf{A}\mathbf{w}\| \geq \epsilon_1, \tag{11.11}$$

with probability $1 - \exp(-\mathcal{O}(m))$ as $\sqrt{m-1} - \sqrt{\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)} \gtrsim (1 - \sqrt{1-\epsilon_L})\sqrt{m}$. Furthermore, we will make use of the following fact that $\|\mathbf{z}\| \leq 2\sigma\sqrt{m}$ with probability $1 - \exp(-\mathcal{O}(m))$, where we let $\mathbf{z} = \sigma\mathbf{v}$ (see Lemma B.2).

Assuming these hold, we will show the existence of $C > 0$ satisfying (11.10). 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,

$$\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}^*_{\ell_2^2}\|^2 + \sigma\tau f(\mathbf{x}^*_{\ell_2^2}) = \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}) \leq \frac{1}{2}\|\mathbf{z}\|^2.$$

On the other hand, expanding the terms,

$$\frac{1}{2}\|\mathbf{z}\|^2 \geq \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}) \geq \frac{1}{2}\|\mathbf{z}\|^2 - \|\mathbf{z}\|\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\| + \frac{1}{2}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}).$$

Using $\|\mathbf{z}\| \leq 2\sigma\sqrt{m}$, this implies,

$$2\sigma\sqrt{m}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\| \geq \|\mathbf{z}\|\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\| \geq \frac{1}{2}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 + \sigma\tau f_p(\mathbf{w}^*_{\ell_2^2}). \tag{11.12}$$

Normalizing by σ,

$$2\sqrt{m}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\| \geq \frac{1}{2\sigma}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 + \tau f_p(\mathbf{w}^*_{\ell_2^2}).$$

The rest of the proof will be split into two cases.

Case 1: Let L be the Lipschitz constant of $f(\cdot)$. If $\mathbf{w}^*_{\ell_2^2} \in \mathcal{T}_f(\mathbf{x}_0, \epsilon_0)$, using (11.11),

$$2\sqrt{m}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\| \geq \frac{1}{2\sigma}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 - \tau L\|\mathbf{w}^*_{\ell_2^2}\| \geq \frac{1}{2\sigma}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|^2 - \frac{\tau L}{\epsilon_1}\|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\|.$$

Further simplifying, we find, $2\sigma(2\sqrt{m} + \frac{\tau L}{\epsilon_1}) \geq \|\mathbf{A}\mathbf{w}^*_{\ell_2^2}\| \geq \epsilon_1\|\mathbf{w}^*_{\ell_2^2}\|$. Hence, indeed, $\frac{\|\mathbf{w}^*_{\ell_2^2}\|}{\sigma}$ is upper bound by $\frac{4\sqrt{m}}{\epsilon_1} + \frac{2\tau L}{\epsilon_1^2}$.

Case 2: Assume $\mathbf{w}^*_{\ell_2^2} \notin \mathcal{T}_f(\mathbf{x}_0, \epsilon_0)$. Then $f_p(\mathbf{w}^*_{\ell_2^2}) \geq \epsilon_0\|\mathbf{w}^*_{\ell_2^2}\|$. Using this and letting $\hat{\mathbf{w}} = \frac{\mathbf{w}^*_{\ell_2^2}}{\sigma}$, we can rewrite (11.12) without σ as,

$$\frac{1}{2}\|\mathbf{A}\hat{\mathbf{w}}\|^2 - 2\sqrt{m}\|\mathbf{A}\hat{\mathbf{w}}\| + 2m + (\tau\epsilon_0\|\hat{\mathbf{w}}\| - 2m) \leq 0.$$

Finally, observing $\frac{1}{2}\|\mathbf{A}\hat{\mathbf{w}}\|^2 - 2\sqrt{m}\|\mathbf{A}\hat{\mathbf{w}}\| + 2m = \frac{1}{2}(\|\mathbf{A}\hat{\mathbf{w}}\| - 2\sqrt{m})^2$, we find,

$$\tau\epsilon_0\|\hat{\mathbf{w}}\| - 2m \leq 0 \implies \frac{\|\mathbf{w}^*_{\ell_2^2}\|}{\sigma} \leq \frac{2m}{\tau\epsilon_0}.$$

□

12. Converse Results

Until now, we have stated the results assuming m is sufficiently large. In particular, we have assumed that $m \geq \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ or $m \geq \mathbf{D}_f(\mathbf{x}_0, \lambda)$. It is important to understand the behavior of the problem when m is small. Showing a converse result for $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ or $m < \mathbf{D}_f(\mathbf{x}_0, \lambda)$ will illustrate the tightness of our analysis. In this section, we focus our attention on the case where $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ and show that the NSE approaches infinity as $\sigma \to 0$. As it has been discussed previously, $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ is the compressed sensing threshold which is the number of measurements required for the success of the noiseless problem (1.2):

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}_0. \tag{12.1}$$

For our analysis, we use Proposition 12.1 below, which is a slight modification of Theorem 1 in [31].

Proposition 12.1. *[[31]] 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 $m \leq \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) - t\sqrt{n}$. Then, $\mathbf{x}_0$ is not a minimizer of (12.1) with probability at least $1 - 4\exp(-\frac{t^2}{4})$.*

Proposition 12.1 leads to the following useful Corollary.

Corollary 12.1. *Consider the same setting as in Proposition 12.1 and denote $\mathbf{x}^*$ the minimizer of (12.1). For a given $t > 0$, there exists an $\epsilon > 0$ such that, with probability $1 - 8\exp(-\frac{t^2}{4})$, we have,*

$$f(\mathbf{x}^*) \leq f(\mathbf{x}_0) - \epsilon$$

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 12.1, $\mathbb{P}(E) \geq 1 - 4\exp(-\frac{t^2}{4})$. 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 $\epsilon = n_0^{-1}$, to conclude the proof. □

The results discussed in this section, hold under the following assumption.

Assumption 12.1. *Assume $m_{lack} := \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+) - m > 0$. $\mathbf{x}_0$ is not a minimizer of the convex function $f(\cdot)$. $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz function, i.e., there exists constant $L > 0$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$.*

12.1. Converse Result for C-LASSO

Recall the C-LASSO problem (1.5):

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}_0 + \sigma\mathbf{v} - \mathbf{A}\mathbf{x}\| \quad \text{subject to} \quad f(\mathbf{x}) \leq f(\mathbf{x}_0). \tag{12.2}$$

(12.2) 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 12.1. *Suppose Assumption 12.1 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 (12.2), $\mathbf{x}_c^*$, such that,*

$$\frac{\|\mathbf{x}_c^* - \mathbf{x}_0\|^2}{\sigma^2} \geq C_{max} \tag{12.3}$$

Proof. From Corollary 12.1, with probability $1-8\exp(-\frac{m_{lack}^2}{4n})$, there exists $\epsilon>0$ and $\mathbf{x}'$ satisfying $f(\mathbf{x}')\leq f(\mathbf{x}_0)-\epsilon$ and $\mathbf{A}\mathbf{x}'=\mathbf{A}\mathbf{x}_0$.Denote $\mathbf{w}'=\mathbf{x}'-\mathbf{x}_0$ and pick a minimizer of (12.2) namely, $\mathbf{x}_0+\mathbf{w}^*$. Now, let $\mathbf{w}_2^*=\mathbf{w}^*+\mathbf{w}'$. Observe that $\|\sigma\mathbf{v}-\mathbf{A}\mathbf{w}^*\|=\|\sigma\mathbf{v}-\mathbf{A}\mathbf{w}_2^*\|$. 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,

$$f(\mathbf{x}_0+\mathbf{w}_2^*)=f(\mathbf{x}'+\mathbf{w}^*)\leq f(\mathbf{x}')+L\|\mathbf{w}^*\|,$$

Hence, if $\|\mathbf{w}^*\|\leq\frac{f(\mathbf{x}_0)-f(\mathbf{x}')}{L}$, $\mathbf{w}_2^*+\mathbf{x}_0$ is a minimizer. Let $C_w=\min\{\frac{f(\mathbf{x}_0)-f(\mathbf{x}')}{L},\frac{1}{2}\|\mathbf{w}'\|\}$ and consider,

$$\mathbf{w}_3^*=\begin{cases}\mathbf{w}^* & \text{if} \quad \|\mathbf{w}^*\|\geq C_w,\\ \mathbf{w}_2^* & \text{otherwise.}\end{cases}$$

From the discussion above, $\mathbf{x}_0+\mathbf{w}_3^*$ is guaranteed to be feasible and minimizer. Now, since $f(\mathbf{x}')\leq f(\mathbf{x}_0)-\epsilon$ and $f(\cdot)$ is Lipschitz, we have that $\|\mathbf{w}'\|\geq\frac{\epsilon}{L}$. Consequently, if $\|\mathbf{w}^*\|\geq C_w$, then, we have, $\frac{\|\mathbf{w}_3^*\|}{\sigma}\geq\frac{\epsilon}{2L\sigma}$. Otherwise, $\|\mathbf{w}^*\|\leq\frac{\|\mathbf{w}'\|}{2}$, and so,

$$\frac{\|\mathbf{w}_3^*\|}{\sigma}=\frac{\|\mathbf{w}_2^*\|}{\sigma}\geq\frac{|\|\mathbf{w}'\|-\|\mathbf{w}^*\||}{\sigma}\geq\frac{\|\mathbf{w}'\|}{2\sigma}\geq\frac{\epsilon}{2L\sigma}.$$

In any case, we find that, $\frac{\|\mathbf{w}_3^*\|}{\sigma}$ is lower bounded by $\frac{\epsilon}{2L\sigma}$ with the desired probability. To conclude with (12.3), we can choose σ_0 sufficiently small to ensure $\frac{\epsilon^2}{4L^2\sigma_0^2}\geq C_{max}$.

□

12.2. 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 $m\geq\mathbf{D}_f(\mathbf{x}_0,\lambda)$. However, hereby, the converse guarantee we give is slightly looser, namely, $m\leq\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)$ where $\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)\leq\mathbf{D}_f(\mathbf{x}_0,\lambda)$ by definition. This is mostly because of the nature of our proof which uses Proposition 12.1 and we believe it is possible to get a converse result for $m\leq\mathbf{D}_f(\mathbf{x}_0,\lambda)$ via Gordon's Lemma. We leave this to future work. Recall ℓ_2-LASSO in (1.6):

$$\min_{\mathbf{x}}\|\mathbf{A}\mathbf{x}_0+\sigma\mathbf{v}-\mathbf{A}\mathbf{x}\|+\lambda f(\mathbf{x}) \tag{12.4}$$

The following theorem is a restatement of Theorem 3.4 and summarizes our result on the ℓ_2-LASSO when m is small.

Theorem 12.2. *Suppose Assumption 12.1 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 (12.4), $\mathbf{x}_{\ell_2}^*$, satisfies,*

$$\frac{\|\mathbf{x}_{\ell_2}^*-\mathbf{x}_0\|^2}{\sigma^2}\geq C_{max}. \tag{12.5}$$

Proof. From Corollary 12.1, with probability $1-8\exp(-\frac{m_{lack}^2}{4n})$, there exists $\epsilon>0$ and $\mathbf{x}'$ satisfying $f(\mathbf{x}')\leq f(\mathbf{x}_0)-\epsilon$ 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 (12.4) and let $\mathbf{w}_2^*=\mathbf{w}^*+\mathbf{w}'$. Clearly, $\|\mathbf{A}\mathbf{w}_2^*-\sigma\mathbf{v}\|=\|\mathbf{A}\mathbf{w}^*-\sigma\mathbf{v}\|$. 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 $f(\cdot)$,

$$f(\mathbf{x}_0+\mathbf{w}_2^*)=f(\mathbf{x}'+\mathbf{w}^*)\leq f(\mathbf{x}')+L\|\mathbf{w}^*\|,$$

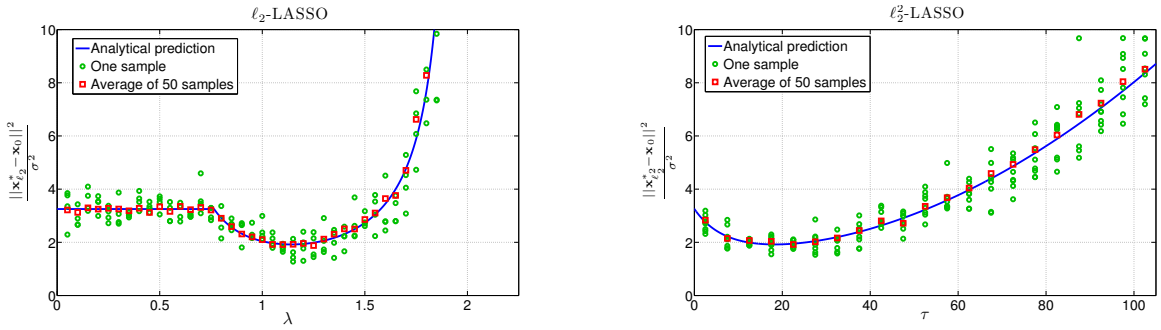
and

$$f(\mathbf{x}_0+\mathbf{w}^*)\geq f(\mathbf{x}_0)-L\|\mathbf{w}^*\|.$$

Combining those, we find,

$$f(\mathbf{x}')+L\|\mathbf{w}^*\|\geq f(\mathbf{x}_0+\mathbf{w}_2^*)\geq f(\mathbf{x}_0+\mathbf{w}^*)\geq f(\mathbf{x}_0)-L\|\mathbf{w}^*\|,$$

which implies, $\|\mathbf{w}^*\|\geq\frac{f(\mathbf{x}_0)-f(\mathbf{x}')}{2L}\geq\frac{\epsilon}{2L}$, and gives the desired result (12.5) when $\sigma_0\leq\frac{\epsilon}{4L\sqrt{C_{max}}}$. □



(a) (b)

Figure 5: 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).

For the ℓ_2^2-LASSO result, let us rewrite (1.7) as,

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}_0 + \sigma\mathbf{v} - \mathbf{A}\mathbf{x}\|^2 + \sigma\tau f(\mathbf{x}) \tag{12.6}$$

The next theorem shows that ℓ_2^2-LASSO does not recover $\mathbf{x}_0$ stably when $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$. Its proof is identical to the proof of Theorem 12.2.

Theorem 12.3. *Suppose Assumption 12.1 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* (12.6)*,* $\mathbf{x}^*_{\ell_2^2}$*, satisfies,*

$$\frac{\|\mathbf{x}^*_{\ell_2^2} - \mathbf{x}_0\|^2}{\sigma^2} \geq C_{max}.$$

13. 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.

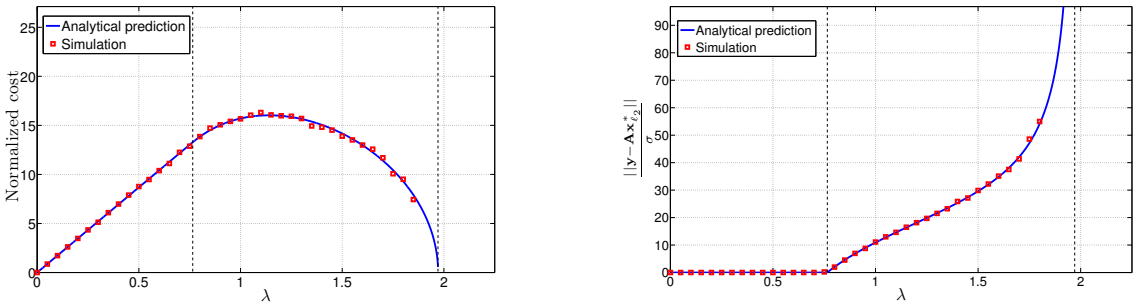
13.1. Sparse Signal Estimation

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 and 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 5.

ℓ_2-LASSO: λ is varied from 0 to 2. The analytical predictions are calculated via the formulas given in Appendix H for the regime $\frac{k}{n} = 0.1$ and $\frac{m}{n} = 0.5$. We have investigated three properties.

- **NSE:** In Figure 5(a), we plot the simulation results with the small σ NSE formulas. Based on Theorem 3.2 and Section 9, over $\mathcal{R}_{\text{ON}}$, we plotted $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{m-\mathbf{D}_f(\mathbf{x}_0,\lambda)}$ and over $\mathcal{R}_{\text{OFF}}$, 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 $\lambda_{\max}$.



(a) (b)

Figure 6: 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 $\mathcal{R}_{\text{ON}}$. In $\mathcal{R}_{\text{OFF}}$, $(\lambda \leq \lambda_{\text{crit}} \approx 0.76)$ observe that $\mathbf{y} = \mathbf{A}\mathbf{x}^*_{\ell_2}$ indeed holds.

- **Normalized cost:** We plotted the cost of ℓ_2-LASSO normalized by σ in Figure 6(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 $\mathcal{R}_{\text{ON}}$, this should be around $\sqrt{m - \mathbf{D}_f(\mathbf{x}_0, \lambda)}$ due to Theorem 6.4. In $\mathcal{R}_{\text{OFF}}$, 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 6(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 11. In $\mathcal{R}_{\text{ON}}$, we analytically expect this to be $\frac{m - \mathbf{D}_f(\mathbf{x}_0,\lambda) - \mathbf{C}_f(\mathbf{x}_0,\lambda)}{\sqrt{m - \mathbf{D}_f(\mathbf{x}_0,\lambda)}}$. In $\mathcal{R}_{\text{OFF}}$, as discussed in Section 9.2, the problem behaves as (1.2) 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 5(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.

13.2. Low-Rank Matrix Estimation

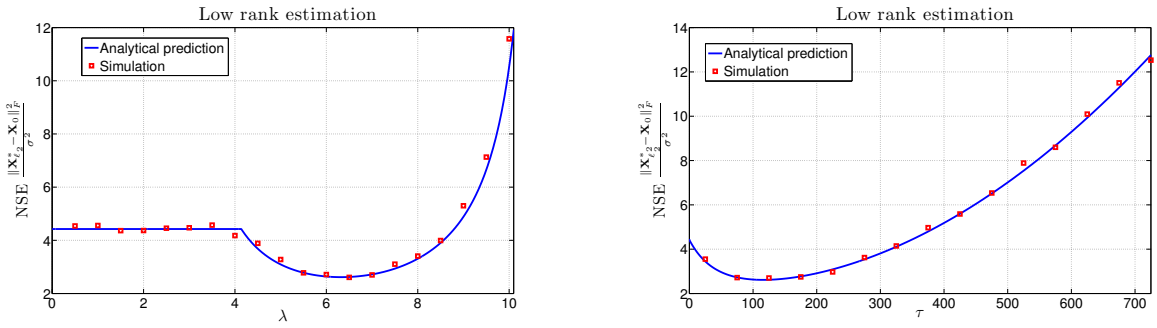
For low rank estimation, we choose the nuclear norm $\|\cdot\|_\star$ as a surrogate for rank [50]. Nuclear norm is the sum of singular values of a matrix and basically takes the role of ℓ_1 minimization.

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,

$$\min_{\mathbf{X}\in\mathbb{R}^{d\times d}} \|\mathbf{y} - \mathbf{A}\cdot\text{vec}(\mathbf{X})\| + \lambda\|\mathbf{X}\|_\star.$$

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 7(b) and 7(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 H, we estimated $\mathbf{D}_f(\mathbf{x}_0, \lambda), \mathbf{C}_f(\mathbf{x}_0, \lambda)$ 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.



(a) (b)

Figure 7: $d = 45$, $m = 0.6d^2$, $r = 6$. 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.

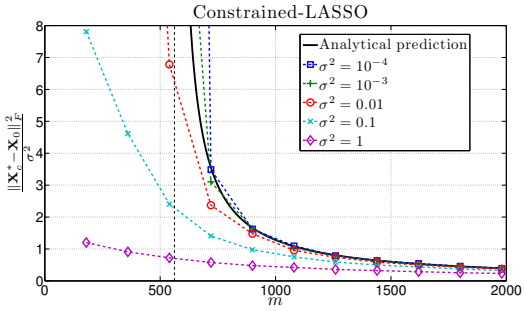


Figure 8: $\mathbf{X}_0$ is a 40×40 matrix with rank 4. As σ decreases, NSE increases. The vertical dashed lines marks the estimated $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ where we expect a transition in stability.

13.3. C-LASSO with varying σ

Consider the low rank estimation problem as in Section 13.2, but use the C-LASSO as an estimator:

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

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 13.2. 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 8. 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 $\frac{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}{m-\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)}$ indeed provides a good estimate of NSE for $\sigma^2 = 10^{-4}$ and upper bounds the remaining ones. In particular, we estimate $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ to be around 560. Based on Theorems 3.4 and 3.1, as m moves from $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$ to $m > \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$, 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 $\sigma^2 \le 10^{-2}$, the NSE becomes noticeably high for the regime $m < \mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$.

14. Future Directions

We believe that our work 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 11 provides justification behind Formula 1, a rigorous proof is arguably the most important point missing in this paper. Such a proof would close the gap in this paper and will extend results of [14, 15] to arbitrary convex functions.

- **Error formulas for arbitrary σ:** Another issue that hasn't been fully explored in this paper 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 $\sigma \to 0$. Empirical observations suggest that the same is true 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.

- **Extension to multiple structures:** Throughout this work, 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 [64–67]. 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 $\mathbf{s}_0$ [32, 60]. 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 [32, 33] 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:

$$\min_{\mathbf{x},\mathbf{s}} \|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{s}\| \quad \text{subject to} \quad g(\mathbf{s}) \leq g(\mathbf{s}_0) \ \text{ and } \ f(\mathbf{x}) \leq f(\mathbf{x}_0).$$

  where $g(\cdot)$ is typically the ℓ_1 norm.

- **Application specific results:** In this paper, 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 [14, 15]. 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, [24, 46, 58, 59, 64, 67]. In particular, one can find an NSE upper bound to a LASSO problem as long as he has an upper bound to $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ or $\mathbf{D}_f(\mathbf{x}_0, \mathbb{R}^+)$.

- **Different $\mathbf{A}, \mathbf{v}$:** Throughout the paper, $\mathbf{A}$ and $\mathbf{v}$ were assumed to be independent with i.i.d. standard normal entries. It might be interesting to consider different measurement ensembles such as matrices with subgaussian entries or even a different noise setup such as "adversarial noise", in which case the error vector $\mathbf{v}$ is generated to maximize the NSE. For example, in the literature of compressed sensing phase transitions, it is widely observed that measurement matrices with subgaussian entries behave same as gaussian ones, [35, 36].

- **Mean-Squared-Error (MSE) Analysis:** In this paper, 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.

Acknowledgments

Authors would like to thank Joel Tropp, Arian Maleki and Kishore Jaganathan for stimulating discussions and helpful comments. S.O. would also like to thank Adrian Lewis for pointing out Proposition 9.2.

References

[1] E. J. Candès, J. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements". Comm. Pure Appl. Math., 59:1207–1223, 2006.

[2] E. J. Candès and T. Tao, "Decoding by linear programming," IEEE Trans. Inform. Theory, 51 4203–4215.

[3] E. J. Candès, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". IEEE Trans. Inform. Theory, 52 489–509.

[4] D. L. Donoho, "Compressed Sensing," IEEE Trans. on Information Theory, 52(4), pp. 1289 – 1306, April 2006.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society, 58:267–288, 1996.

[6] P. J. Bickel , Y. Ritov and A. Tsybakov "Simultaneous analysis of LASSO and Dantzig Selector". The Annals of Statistics, 37(4):1705–1732, 2009.

[7] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. Electronic Journal of Statistics, 1:169–194, 2007.

[8] S.S. Chen and D. Donoho. "Examples of basis pursuit". Proceeding of wavelet applications in signal and image processing III, 1995.

[9] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. "Sparsity oracle inequalities for the lasso." Electronic Journal of Statistics, 1:169–194, 2007.

[10] M J Wainwright. "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-constrained quadratic programming" Information Theory, IEEE Transactions on 55.5 (2009): 2183-2202.

[11] P. Zhao and B. Yu. "On model selection consistency of Lasso". Journal of Machine Learning Research, 7:2541Ð2567, 2006.

[12] D. L. Donoho, M. Elad, and V. M. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Info Theory, 52(1):6Ð18, January 2006.

[13] N. Meinshausen and B. Yu. "Lasso-type recovery of sparse representations for high-dimensional data." Ann. Statist., 37(1):246–270, 2009.

[14] M. Bayati and A. Montanari. "The dynamics of message passing on dense graphs, with applications to compressed sensing." IEEE Transactions on Information Theory, Vol. 57, No. 2, 2011.

[15] M. Bayati and A. Montanari, "The LASSO risk for gaussian matrices", IEEE Transactions on Information Theory, Vol. 58, No. 4, 2012.

[16] A. Maleki, L. Anitori, A. Yang, and R. Baraniuk, "Asymptotic Analysis of Complex LASSO via Complex Approximate Message Passing (CAMP)", Information Theory, IEEE Transactions on, vol. 59, no. 7, pp. 4290–4308, 2011.

[17] E. J. Candès and M. A. Davenport "How well can we estimate a sparse vector?", arXiv:1104.5246.

[18] E. J. Candès and T. Tao. "The Dantzig selector: Statistical estimation when p is much larger than n". Ann. Stat., 35(6):2313–2351, 2007.

[19] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming". Linear algebra and its applications, 284(1), 193–228.

[20] L. Meier, S. van de Geer, and P. Buhlmann, "The group Lasso for logistic regression". J. Roy. Statist. Soc. Ser. B 70 53–71, 2008.

[21] N. Meinshausen and B. Yu. "Lasso-type recovery of sparse representations for high-dimensional data." The Annals of Statistics (2009): 246–270.

[22] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear norm penalization and optimal rates for noisy matrix completion", Annals of Statistics, 2011.

[23] E. J. Candès and Y. Plan. "Matrix completion with noise". Proceedings of the IEEE 98(6), 925–936.

[24] D. Needell and R. Ward, "Stable image reconstruction using total variation minimization", arXiv preprint arXiv:1202.6429, (2012).

[25] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The Convex Geometry of Linear Inverse Problems", Foundations of Computational Mathematics. Online First, October 2012.

[26] M. Stojnic, "Various thresholds for ℓ_1 - optimization in compressed sensing", arXiv:0907.3666v1.

[27] D. L. Donoho, J. Tanner, "Thresholds for the recovery of sparse solutions via l1 minimization", Conf. on Information Sciences and Systems, 2006.

[28] D. L. Donoho and J. Tanner, "Neighborliness of randomly-projected simplices in high dimensions," Proc. National Academy of Sciences, 102(27), pp. 9452-9457, 2005.

[29] D. L. Donoho, "High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension", Comput. Geometry, (online) Dec. 2005.

[30] M. Stojnic, "A rigorous geometry-probability equivalence in characterization of ℓ_1-optimization", arXiv:1303.7287.

[31] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: A geometric theory of phase transitions in convex optimization". arXiv:1303.6672.

[32] R. Foygel and L. Mackey, "Corrupted Sensing: Novel Guarantees for Separating Structured Signals", arXiv:1305.2524.

[33] M. B. McCoy and J. A. Tropp, " The achievable performance of convex demixing", arXiv:1309.7478.

[34] F. Bach "Structured sparsity-inducing norms through submodular functions", NIPS 2010.

[35] M. Bayati, M. Lelarge, and A. Montanari. “Universality in polytope phase transitions and message passing algorithms”. Available at arxiv.org/abs/1207.7321, 2012.

[36] D. L. Donoho and J. Tanner. “Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing.” Phil. Trans. R. Soc. A 13 November 2009 vol. 367 no. 1906 4273–4293.

[37] M. Stojnic, “A framework to characterize performance of LASSO algorithms”, arXiv:1303.7291.

[38] M. Stojnic, “A performance analysis framework for SOCP algorithms in noisy compressed sensing”, arXiv:1304.0002.

[39] M. Stojnic, “Regularly random duality”, arXiv:1304.0002.

[40] D. L. Donoho, “De-noising by soft-thresholding,” Information Theory, IEEE Transactions on, vol. 41, no. 3, pp. 613–627, 1995.

[41] S. Oymak and B. Hassibi, “Asymptotically Exact Denoising in Relation to Compressed Sensing”, arXiv:1305.2714.

[42] S. Oymak, and B. Hassibi. “On a relation between the minimax risk and the phase transitions of compressed recovery.” Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on. IEEE, 2012.

[43] V. Chandrasekaran, and M. I. Jordan, “Computational and statistical tradeoffs via convex relaxation”, Proceedings of the National Academy of Sciences 110.13 (2013): E1181-E1190.

[44] D. Donoho, I. Johnstone, and A. Montanari. "Accurate Prediction of Phase Transitions in Compressed Sensingvia a Connection to Minimax Denoising." (2013): 1-1.

[45] D. L. Donoho, M. Gavish, “Minimax Risk of Matrix Denoising by Singular Value Thresholding”, arXiv:1304.2085.

[46] Y. C. Eldar, P. Kuppinger, and H. BŽlcskei, “Block-Sparse Signals: Uncertainty Relations and Efficient Recovery”, IEEE Trans. on Signal Proc., Vol. 58, No. 6, June 2010.

[47] M. Stojnic, F. Parvaresh, and B. Hassibi, “On the reconstruction of block-sparse signals with an optimal number of measurements”. IEEE Trans. on Signal Processing, vol.57, no.8, pp.3075-3085, Aug. 2009.

[48] M. Stojnic, “Block-length dependent thresholds in block-sparse compressed sensing”, arXiv:0907.3679.

[49] N. Rao, B. Recht, and R. Nowak, “Tight Measurement Bounds for Exact Recovery of Structured Sparse Signals”. In Proceedings of AISTATS, 2012.

[50] B. Recht, M. Fazel, P. Parrilo, “Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization”. SIAM Review, Vol 52, no 3, pages 471–501, 2010.

[51] M. Fazel, “Matrix Rank Minimization with Applications”. Elec. Eng. Dept, Stanford University, March 2002.

[52] E. J. Candès and B. Recht. “Exact matrix completion via convex optimization”. Found. of Comput. Math., 9 717-772.

[53] E. J. Candès and T. Tao. “The power of convex relaxation: Near-optimal matrix completion”. IEEE Trans. Inform. Theory 56(5), 2053–2080.

[54] E. J. Candès and Y. Plan. “Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements.” IEEE Transactions on Information Theory 57(4), 2342–2359.

[55] S. Oymak and B. Hassibi, “New Null Space Results and Recovery Thresholds for Matrix Rank Minimization”, arXiv:1011.6326.

[56] S. Oymak and B. Hassibi. “Tight recovery thresholds and robustness analysis for nuclear norm minimization.” Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on. IEEE, 2011.

[57] D. L. Donoho, M. Gavish, and A. Montanari, “The Phase Transition of Matrix Recovery from Gaussian Measurements Matches the Minimax MSE of Matrix Denoising”, arXiv:1302.2331.

[58] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, “Simultaneously Structured Models with Application to Sparse and Low-rank Matrices”, arXiv:1212.3753.

[59] E. Richard, P. Savalle, and N. Vayatis, “Estimation of Simultaneously Sparse and Low Rank Matrices”, in Proceedings of the 29th International Conference on Machine Learning (ICML 2012).

[60] X. Li, “Compressed sensing and matrix completion with constant proportion of corruptions.” Constructive Approximation 37(1), 73–99.

[61] D. L. Donoho, A. Maleki, and A. Montanari, “Message Passing Algorithms for Compressed Sensing”, PNAS November 10, 2009 vol. 106 no. 45 18914–18919.

[62] D. L. Donoho, A. Maleki, A. Montanari, “The Noise-Sensitivity Phase Transition in Compressed Sensing”, IEEE Trans. Inform. Theory, 57 6920–6941 .

[63] A. Maleki, "Analysis of approximate message passing algorithm", CISS 2010.

[64] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive Principal Component Pursuit", arXiv:1202.4596.

[65] E. J. Candès, X. Li, Y. Ma, and J. Wright. "Robust Principal Component Analysis?" Journal of ACM 58(1), 1–37.

[66] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, A. S. Willsky, "Rank-Sparsity Incoherence for Matrix Decomposition", SIAM Journal on Optimization, Vol. 21, Issue 2, pp. 572–596, 2011.

[67] M. B. McCoy and J. A. Tropp, " Sharp recovery bounds for convex deconvolution, with applications", arXiv:1205.1580.

[68] CVX Research, Inc. "CVX: Matlab software for disciplined convex programming", version 2.0 beta. http://cvxr.com/cvx, September 2012.

[69] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for Optimization over Symmetric Cones", Optimization Methods and Software 11(12), pp. 625–653, 1999.

[70] S. J. Press, "Applied multivariate analysis: using Bayesian and frequentist methods of inference", Courier Dover Publications, 2012.

[71] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices", Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012.

[72] Y. Gordon, "On MilmanÕs inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$", in Geometric Aspects of Functional Analysis, volume 1317 of Lecture Notes in Mathematics, pages 84–106. Springer, 1988.

[73] Y. Gordon,"Some inequalities for Gaussian processes and applications." Israel Journal of Mathematics 50.4 (1985): 265-289.

[74] M. Ledoux, M. Talagrand, "Probability in Banach Spaces: Isoperimetry and Processes". Springer, 1991.

[75] M. Ledoux. "The concentration of measure phenomenon", volume 89 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI 2001.

[76] V. I. Bogachev. "Gaussian measures", volume 62 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 1998.

[77] J. Borwein and A. Lewis. "Convex analysis and nonlinear optimization: theory and examples". Vol. 3. Springer, 2006.

[78] J.M. Borwein. "A note on the existence of subgradients". Mathematical Programming, 24:225–228, 1982.

[79] L. R. Scott, "Numerical Analysis", Princeton University Press, 2011.

[80] J. J. Moreau. "Fonctions convexes duales et points proximaux dans un espace hilbertien." C.R. Acad. Sci. Paris Sèr. A Math., 255:1897 2899, 1962.

[81] Y. Nesterov, "Introductory Lectures on Convex Optimization. A Basic Course", 2004.

[82] D. Bertsekas with A. Nedic and A.E. Ozdaglar, "Convex Analysis and Optimization". Athena Scientific, 2003.

[83] S. Boyd and L. Vandenberghe, "Convex Optimization" Cambridge University Press, 2004.

[84] R. T. Rockafellar, "Second-order convex analysis". Journal of Nonlinear and Convex Analysis 1 (1999), 1–16.

[85] JB Hiriart-Urruty and C LemarŐchal. "Convex Analysis and Minimization Algorithms: Part 1: Fundamentals". Vol. 1. Springer, 1996.

[86] R. T. Rockafellar, "Convex Analysis", Vol. 28. Princeton university press, 1997.

APPENDIX

A. Useful Facts

Fact A.1 (Moreau's decomposition theorem). *Let $\mathcal{C}$ be a closed and convex cone in $\mathbb{R}^n$. For any $\mathbf{v} \in \mathbb{R}^n$, the following two are equivalent:*

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

Fact A.2 (Properties of the projection, [82, 83]). *Assume $\mathcal{C} \subseteq \mathbb{R}^n$ is a nonempty, closed and convex set and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are arbitrary points. Then,*

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

Fact A.3 (Variance of Lipschitz functions). *Assume $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_p)$ and let $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ be an L-Lipschitz function. Then,*

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

Fact A.4 (Gaussian concentration Inequality for Lipschitz functions). *Let $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ be an L-Lipschitz function and $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_p)$. Then,*

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

B. Auxiliary Results

Lemma B.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,*

$$\sqrt{\mathbb{E}\left[(f(\mathbf{g}))^2\right] - L^2} - t \leq f(\mathbf{g}) \leq \sqrt{\mathbb{E}\left[(f(\mathbf{g}))^2\right]} + t,$$

with probability $1 - 2\exp(-\frac{t^2}{2L^2})$.

Proof. From Fact A.4,

$$|f(\mathbf{g}) - \mathbb{E}[f(\mathbf{g})]| \leq t, \tag{B.1}$$

holds with probability $1 - 2\exp(-\frac{t^2}{2L^2})$. Furthermore,

$$\mathbb{E}[(f(\mathbf{g}))^2] - L^2 \leq (\mathbb{E}[f(\mathbf{g})])^2 \leq \mathbb{E}[(f(\mathbf{g}))^2]. \tag{B.2}$$

The left hand side inequality in B.2 follows from an application of Fact A.3 and the right hand side follows from Jensen's Inequality.

Combining (B.1) and (B.2) completes the proof. □

For the statements of the lemmas below, recall the definitions of $\mathbf{D}(\mathcal{C})$,$\mathbf{P}(\mathcal{C})$ and $\mathbf{C}(\mathcal{C})$ in Section 6.2.

Lemma B.2. *Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m), \mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ and let $\mathcal{C} \in \mathbb{R}^n$ be a closed and convex set. Given $t > 0$, each of the followings hold with probability $1 - 2\exp\left(\frac{-t^2}{2}\right)$.*

- $\sqrt{m-1} - t \leq \|\mathbf{g}\|_2 \leq \sqrt{m} + t$
- $\sqrt{\mathbf{D}(\mathcal{C}) - 1} - t \leq dist(\mathbf{h}, \mathcal{C}) \leq \sqrt{\mathbf{D}(\mathcal{C})} + t$
- $\sqrt{\mathbf{P}(\mathcal{C}) - 1} - t \leq \|Proj(\mathbf{h}, \mathcal{C})\|_2 \leq \sqrt{\mathbf{P}(\mathcal{C})} + t$

Proof. The result is an immediate application of Lemma B.1. The functions $\|\cdot\|$, $\|\text{Proj}(\cdot,\mathcal{C})\|$ and $\text{dist}(\cdot,\mathcal{C})$ are all 1-Lipschitz. Furthermore, $\mathbb{E}[\|\mathbf{g}\|_2^2] = m$ and $\mathbb{E}[\text{Proj}(\mathbf{h},\mathcal{C})^2] = \mathbf{P}(\mathcal{C})$, $\mathbb{E}[\text{dist}(\mathbf{h},\mathcal{C})^2] = \mathbf{D}(\mathcal{C})$ by definition. □

Lemma B.3. *Let $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ and let $\mathcal{C} \in \mathbb{R}^n$ be a convex and closed set. Then, given $t > 0$,*

- $|dist(\mathbf{h},\mathcal{C})^2 - \mathbf{D}(\mathcal{C})| \leq 2t\sqrt{\mathbf{D}(\mathcal{C})} + t^2 + 1.$
- $|\|Proj(\mathbf{h},\mathcal{C})\|^2 - \mathbf{P}(\mathcal{C})| \leq 3t\sqrt{n + \mathbf{D}(\mathcal{C})} + t^2 + 1.$
- $|corr(\mathbf{h},\mathcal{C}) - \mathbf{C}(\mathcal{C})| \leq 3t\sqrt{n + \mathbf{D}(\mathcal{C})} + t^2 + 1.$

with probability $1 - 4\exp(-\frac{t^2}{2})$.

Proof. The first two statements follow trivially from Lemma B.2. For the second statement, use again Lemma B.2 and also upper bound $\mathbf{P}(\mathcal{C})$ by $2(n + \mathbf{D}(\mathcal{C}))$ via Lemma B.4. To obtain the third statement, we write,

$$\text{corr}(\mathbf{h},\mathcal{C}) = \frac{n - (\|\text{Proj}(\mathbf{h},\mathcal{C})\|^2 + \text{dist}(\mathbf{h},\mathcal{C})^2)}{2}$$

and use the fact that first two statements hold with probability $1 - 4\exp(-\frac{t^2}{2})$. This will give,

$$|\text{corr}(\mathbf{h},\mathcal{C}) - \mathbf{C}(\mathcal{C})| \leq t(\sqrt{\mathbf{D}(\mathcal{C})} + \sqrt{\mathbf{P}(\mathcal{C})}) + t^2 + 1,$$

which when combined with Lemma B.4 concludes the proof. □

Lemma B.4. *Let $\mathcal{C} \in \mathbb{R}^n$ be a convex and closed set. Then, the following holds,*

$$\max\{\mathbf{C}(\mathcal{C}), \mathbf{P}(\mathcal{C})\} \leq 2(n + \mathbf{D}(\mathcal{C})).$$

Proof. From triangle inequality, for any $\mathbf{h} \in \mathbb{R}^n$,

$$\|\text{Proj}(\mathbf{h},\mathcal{C})\| \leq \|\mathbf{h}\| + \text{dist}(\mathbf{h},\mathcal{C}).$$

We also have,

$$\mathbb{E}[\|\mathbf{h}\| \cdot \text{dist}(\mathbf{h},\mathcal{C})] \leq \frac{1}{2}(\mathbb{E}[\|\mathbf{h}\|^2] + \mathbb{E}[\text{dist}(\mathbf{h},\mathcal{C})^2]) = \frac{n + \mathbf{D}(\mathcal{C})}{2}.$$

From these, we may write,

$$\begin{aligned}\mathbf{C}(\mathcal{C}) &= \mathbb{E}[\langle \Pi(\mathbf{h},\mathcal{C}), \text{Proj}(\mathbf{h},\mathcal{C})\rangle] \\ &\leq \mathbb{E}[\text{dist}(\mathbf{h},\mathcal{C})\|\text{Proj}(\mathbf{h},\mathcal{C})\|] \\ &\leq \frac{n + 3\mathbf{D}(\mathcal{C})}{2}.\end{aligned}$$

Similarly, we have,

$$\mathbf{P}(\mathcal{C}) = \mathbb{E}[\|\text{Proj}(\mathbf{h},\mathcal{C})\|^2] \leq \mathbb{E}[\|\mathbf{h}\| + \text{dist}(\mathbf{h},\mathcal{C})^2] \leq 2(n + \mathbf{D}(\mathcal{C})).$$

□

Lemma B.5. *Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_m)$ and $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$. Let $\mathcal{C}$ be a closed and convex set in $\mathbb{R}^n$. Assume $m(1 - \epsilon_L) > \mathbf{D}(\mathcal{C}) > \epsilon_L m$ for some constant $\epsilon_L > 0$ and m is sufficiently large. Then, for any constant $\epsilon > 0$, each of the following holds with probability $1 - \exp(-\mathcal{O}(m))$,*

- $\|\mathbf{g}\| > dist(\mathbf{h},\mathcal{C}).$
- $\left|\frac{\|\mathbf{g}\|^2 - dist^2(\mathbf{h},\mathcal{C})}{m - \mathbf{D}(\mathcal{C})} - 1\right| < \epsilon.$
- $\left|\frac{dist^2(\mathbf{h},\mathcal{C})}{\|\mathbf{g}\|^2 - dist^2(\mathbf{h},\mathcal{C})} \times \frac{m - \mathbf{D}(\mathcal{C})}{\mathbf{D}(\mathcal{C})} - 1\right| < \epsilon.$

Proof. Let δ be a constant to be determined. For sufficiently large m, using Lemma B.2, with probability $1 - \exp(-\mathcal{O}(m))$, we have,

$$|\|\mathbf{g}\|^2 - m| < \delta m, \quad |\text{dist}(\mathbf{h}, \mathcal{C})^2 - \mathbf{D}(\mathcal{C})| < \delta m$$

Now, choose $\delta < \frac{\epsilon_L}{2}$, which gives,

$$\|\mathbf{g}\| \geq \sqrt{m(1-\delta)} > \sqrt{\mathbf{D}(\mathcal{C}) + \epsilon_L m - \delta m} > \sqrt{\mathbf{D}(\mathcal{C}) + \delta m} \geq \text{dist}(\mathbf{h}, \mathcal{C})$$

This gives the first statement. For the second statement, observe that,

$$1 + \frac{2\delta}{\epsilon_L} \geq \frac{m - \mathbf{D}(\mathcal{C}) + 2\delta}{m - \mathbf{D}(\mathcal{C})} \geq \frac{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \mathcal{C})}{m - \mathbf{D}(\mathcal{C})} \geq \frac{m - \mathbf{D}(\mathcal{C}) - 2\delta}{m - \mathbf{D}(\mathcal{C})} \geq 1 - \frac{2\delta}{\epsilon_L}.$$

Choose $\frac{\delta}{\epsilon_L} < \frac{\epsilon}{2}$ to ensure the desired result. For the last statement, we similarly have,

$$\frac{1 + \frac{\delta}{\epsilon_L}}{1 - \frac{2\delta}{\epsilon_L}} \geq \frac{\text{dist}^2(\mathbf{h}, \mathcal{C})}{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \mathcal{C})} \times \frac{m - \mathbf{D}(\mathcal{C})}{\mathbf{D}(\mathcal{C})} \geq \frac{1 - \frac{\delta}{\epsilon_L}}{1 + \frac{2\delta}{\epsilon_L}} \tag{B.3}$$

To conclude, notice that we can choose $\frac{\delta}{\epsilon_L}$ sufficiently small (constant) to ensure that the left and right bounds in (B.3) above are between $1 \pm \epsilon$.

□

Proof of Lemma 11.3. We will show the results for $L_{\mathbf{P}}(\lambda)$ and $L_{\mathbf{D}}(\lambda)$. $L_{\mathbf{C}}(\lambda)$ follows from the fact that $\mathbf{P}(\lambda\mathcal{C}) + \mathbf{D}(\lambda\mathcal{C}) + 2\mathbf{C}(\lambda\mathcal{C}) = n$. Let $\mathbf{h} \in \mathbb{R}^n$. Then, for $\lambda + \epsilon, \lambda > 0$,

$$\|\text{Proj}(\mathbf{h}, (\lambda+\epsilon)\mathcal{C})\| = \frac{\lambda+\epsilon}{\lambda}\|\text{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\epsilon}, \lambda\mathcal{C})\| = \|\text{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\epsilon}, \lambda\mathcal{C})\| + \frac{\epsilon}{\lambda}\|\text{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\epsilon}, \lambda\mathcal{C})\|$$

This gives,

$$\left|\|\text{Proj}(\mathbf{h}, (\lambda+\epsilon)\mathcal{C})\| - \|\text{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\epsilon}, \lambda\mathcal{C})\|\right| \leq |\epsilon| R$$

Next, observe that,

$$\left|\|\text{Proj}(\frac{\lambda\mathbf{h}}{\lambda+\epsilon}, \lambda\mathcal{C})\| - \|\text{Proj}(\mathbf{h}, \lambda\mathcal{C})\|\right| \leq \frac{|\epsilon|\|\mathbf{h}\|}{\lambda+\epsilon}$$

Combining, letting $\mathbf{h} \sim \mathcal{N}(0, \mathbf{I}_n)$ and using $\|\text{Proj}(\mathbf{h}, \lambda\mathcal{C})\| \leq \lambda R$, we find,

$$\begin{aligned}\mathbf{P}((\lambda+\epsilon)\mathcal{C}) &\leq \mathbb{E}[(\|\text{Proj}(\mathbf{h}, \lambda\mathcal{C})\| + \frac{|\epsilon|\|\mathbf{h}\|}{\lambda+\epsilon} + |\epsilon|R)^2] \\ &\leq \mathbf{P}(\lambda\mathcal{C}) + 2\lambda R|\epsilon|(\frac{\mathbb{E}[\|\mathbf{h}\|]}{\lambda+\epsilon} + R) + |\epsilon|^2\mathbb{E}[(\frac{\|\mathbf{h}\|}{\lambda+\epsilon} + R)^2]\end{aligned}$$

Obtaining the similar lower bound on $\mathbf{P}((\lambda+\epsilon)\mathcal{C})$ and letting $\epsilon \to 0$,

$$L_{\mathbf{P}}(\lambda) = \limsup_{\epsilon\to 0}\left|\frac{\mathbf{P}((\lambda+\epsilon)\mathcal{C}) - \mathbf{P}(\lambda\mathcal{C})}{\epsilon}\right| \leq \lim_{\epsilon\to 0} 2\lambda R(\frac{\mathbb{E}[\|\mathbf{h}\|]}{\lambda+\epsilon} + R + \mathcal{O}(|\epsilon|)) \leq 2R(\sqrt{n} + \lambda R)$$

For $\lambda = 0$, observe that for any $\epsilon > 0, \mathbf{h} \in \mathbb{R}^n$, $\|\text{Proj}(\mathbf{h}, \epsilon\mathcal{C})\| \leq \epsilon R$ which implies $\mathbf{P}(\epsilon\mathcal{C}) \leq \epsilon^2 R^2$. Hence,

$$L_{\mathbf{P}}(0) = \lim_{\epsilon\to 0^+} \epsilon^{-1}(\mathbf{P}(\epsilon\mathcal{C}) - \mathbf{P}(0)) = 0 \tag{B.4}$$

Next, consider $\mathbf{D}(\lambda\mathcal{C})$. Using differentiability of $\mathbf{D}(\lambda\mathcal{C})$, for $\lambda > 0$,

$$L_{\mathbf{D}}(\lambda) = |\mathbf{D}(\lambda\mathcal{C})'| = \frac{2}{\lambda}|\mathbf{C}(\lambda\mathcal{C})| \leq \frac{2 \cdot \mathbb{E}[\|\text{Proj}(\mathbf{h}, \lambda\mathcal{C})\| \cdot \text{dist}(\mathbf{h}, \lambda\mathcal{C})]}{\lambda} \leq 2R \cdot \mathbb{E}[\text{dist}(\mathbf{h}, \lambda\mathcal{C})] \leq 2R(\sqrt{n} + \lambda R)$$

For $\lambda = 0$, see the "Continuity at zero" part of the proof of Lemma B.2 in [31], which gives the upper bound $2R\sqrt{n}$ on $L_{\mathbf{D}}(0)$. □

C. Proof of (modified) Gordon's Lemma

In this section we prove the modified Gordon's Lemma 5.1. The Lemma is a consequence of Theorem 5.1. We repeat the statement of the Lemma for ease of reference.

Lemma 5.1 (Modified Gordon's Lemma)**.** *Let* $\mathbf{G}$, $\mathbf{g}$, $\mathbf{h}$ *be defined as in Lemma 2.1 and let* $\psi(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *. Also, let* $\Phi_1 \subset \mathbb{R}^n$ *and* $\Phi_2 \subset \mathbb{R}^m$ *such that either both* Φ_1 *and* Φ_2 *are compact or* Φ_1 *is arbitrary and* Φ_2 *is a scaled unit sphere. Then, for any* $c \in \mathbb{R}$*:*

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

Our proof will closely parallel the proof of the original Gordon's Lemma 3.1 in [72].

Proof. For $\mathbf{x}\in\Phi_1$ and $\mathbf{a}\in\Phi_2$ define the two processes,

$$Y_{\mathbf{x},\mathbf{a}} = \mathbf{x}^T\mathbf{G}\mathbf{a} + \|\mathbf{a}\|\|\mathbf{x}\|g \quad \text{and} \quad X_{\mathbf{x},\mathbf{a}} = \|\mathbf{x}\|\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{h}^T\mathbf{x}$$

where $\mathbf{G},\mathbf{g},\mathbf{h}$ are as defined in the statement of the lemma 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 5.1:

$$\mathbb{E}[X_{\mathbf{x},\mathbf{a}}^2] = \|\mathbf{x}\|^2\|\mathbf{a}\|^2 + \|\mathbf{a}\|^2\|\mathbf{x}\|^2 = \mathbb{E}[Y_{\mathbf{x},\mathbf{a}}^2],$$

and

$$\begin{aligned}\mathbb{E}[X_{\mathbf{x},\mathbf{a}}X_{\mathbf{x}',\mathbf{a}'}] - \mathbb{E}[Y_{\mathbf{x},\mathbf{a}}Y_{\mathbf{x}',\mathbf{a}'}] &= \|\mathbf{x}\|\|\mathbf{x}'\|(\mathbf{a}^T\mathbf{a}') + \|\mathbf{a}\|^2(\mathbf{x}^T\mathbf{x}') - (\mathbf{x}^T\mathbf{x}')(\mathbf{a}^T\mathbf{a}') - \|\mathbf{a}\|\|\mathbf{a}'\|\|\mathbf{x}\|\|\mathbf{x}'\| \\ &= \left(\underbrace{\|\mathbf{x}\|\|\mathbf{x}'\| - (\mathbf{x}^T\mathbf{x}')}_{\geq 0}\right)\left(\underbrace{(\mathbf{a}^T\mathbf{a}') - \|\mathbf{a}\|\|\mathbf{a}'\|}_{\leq 0}\right),\end{aligned}$$

which is non positive and equal to zero when $x = x'$. Also, on the way of applying Theorem 5.1 for the two processes defined above, let

$$\lambda_{\mathbf{x},\mathbf{a}} = \psi(\mathbf{x},\mathbf{a}) + c.$$

The only caveat in directly applying Theorem 5.1 is now that it requires the processes to be discrete. This technicality is addressed by Gordon in [72] (see Lemma 3.1 therein), for the case where Φ_1 is arbitrary and Φ_2 is a scaled unit sphere. In Lemma C.1, we show that the minimax inequality can be translated from discrete to continuous processes, as well, in the case where both Φ_1 and Φ_2 are compact sets. To conclude, applying Theorem 5.1 we have,

$$\begin{aligned}\mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\mathbf{a}^T\mathbf{G}\mathbf{x}+\|\mathbf{a}\|\|\mathbf{x}\|g-\psi(\mathbf{x},\mathbf{a})\right\}\geq c\right)\geq& \\ \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\|\mathbf{x}\|\mathbf{g}^T\mathbf{a}-\|\mathbf{a}\|\mathbf{h}^T\mathbf{x}-\psi(\mathbf{x},\mathbf{a})\right\}\geq c\right) &:= q. \qquad \text{(C.1)}\end{aligned}$$

Since $g\sim\mathcal{N}(0,1)$, we can write the left hand side of (C.1) as, $p = \frac{p_+ + p_-}{2}$ where we define p_+, p_-, p_0 as,

$$\begin{aligned}p_- &= \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\mathbf{a}^T\mathbf{G}\mathbf{x}+\|\mathbf{a}\|\|\mathbf{x}\|g-\psi(\mathbf{x},\mathbf{a})\right\}\geq c \mid g\leq 0\right), \\ p_+ &= \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\mathbf{a}^T\mathbf{G}\mathbf{x}+\|\mathbf{a}\|\|\mathbf{x}\|g-\psi(\mathbf{x},\mathbf{a})\right\}\geq c \mid g> 0\right), \\ p_0 &= \mathbb{P}\left(\min_{\mathbf{x}\in\Phi_1}\max_{\mathbf{a}\in\Phi_2}\left\{\mathbf{a}^T\mathbf{G}\mathbf{x}-\psi(\mathbf{x},\mathbf{a})\right\}\geq c\right)\end{aligned}$$

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.

□

Lemma C.1. *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,*

$$\mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2} \{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c) \geq \mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2} \{\|\mathbf{x}\|\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c)$$

Then,

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

Proof. Let $R(\Phi_i) = \sup_{\mathbf{v}\in\Phi_i} \|\mathbf{v}\|$ 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 $\|\mathbf{v}' - \mathbf{v}\| \leq \epsilon$. 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,

$$|[\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})] - [\mathbf{a}'^T\mathbf{G}\mathbf{x}' - \psi(\mathbf{x}',\mathbf{a}')]| \leq \epsilon(R(\Phi_1) + R(\Phi_2) + \epsilon)\|\mathbf{G}\|_2 + \delta \tag{C.2}$$

$$|[\|\mathbf{x}\|\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})] - [\|\mathbf{x}'\|\mathbf{g}^T\mathbf{a}' - \|\mathbf{a}'\|\mathbf{h}^T\mathbf{x}' - \psi(\mathbf{x}',\mathbf{a}')]| \leq \epsilon(R(\Phi_1) + R(\Phi_2) + \epsilon)(\|\mathbf{g}\| + \|\mathbf{h}\|) + \delta \tag{C.3}$$

Next, using Lipschitzness of $\|\mathbf{g}\|, \|\mathbf{h}\|, \|\mathbf{G}\|_2$ and Lemma B.2, for $t > 1$, we have,

$$\mathbb{P}(\max\{\|\mathbf{g}\| + \|\mathbf{h}\|, \|\mathbf{G}\|_2\} \leq t(\sqrt{n} + \sqrt{m})) \geq 1 - 4\exp(-\frac{(t-1)^2(m+n)}{2}) := p(t) \tag{C.4}$$

Let $C(t,\epsilon) = t\epsilon(R(\Phi_1) + R(\Phi_2) + \epsilon)(\sqrt{m} + \sqrt{n}) + \delta$. Then, since (C.2) and (C.3) holds for all $\mathbf{a}, \mathbf{x}$, using (C.4),

$$\mathbb{P}(\min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{a}\in\Phi_2}\{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c - C(t,\epsilon)) \geq \mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2}\{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c) - p(t) \tag{C.5}$$

$$\mathbb{P}(\min_{\mathbf{x}\in S_1} \max_{\mathbf{a}\in S_2}\{\|\mathbf{x}\|\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c) \geq \mathbb{P}(\min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{a}\in\Phi_2}\{\|\mathbf{x}\|\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c + C(t,\epsilon)) - p(t) \tag{C.6}$$

Combining (C.5) and (C.6), for all $\epsilon > 0, t > 1$, the following holds,

$$\mathbb{P}(\min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{a}\in\Phi_2}\{\mathbf{a}^T\mathbf{G}\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c - C(t,\epsilon)) \geq \mathbb{P}(\min_{\mathbf{x}\in\Phi_1} \max_{\mathbf{a}\in\Phi_2}\{\|\mathbf{x}\|\mathbf{g}^T\mathbf{a} - \|\mathbf{a}\|\mathbf{h}^T\mathbf{x} - \psi(\mathbf{x},\mathbf{a})\} \geq c + C(t,\epsilon)) - 2p(t)$$

Setting $t = \epsilon^{-1/2}$ and letting $\epsilon \to 0$, we obtain the desired result as $C(t,\epsilon), p(t), \delta \to 0$. □

D. The Dual of the LASSO

To derive the dual we write the problem in (5.7) equivalently as

$$\begin{aligned}\mathcal{F}(\mathbf{A},\mathbf{v}) = &\min_{\mathbf{w},\mathbf{b}} \{\|\mathbf{b}\| + p(\mathbf{w})\}\\ &\text{s.t.} \quad \mathbf{b} = \mathbf{A}\mathbf{w} - \sigma\mathbf{v},\end{aligned}$$

and then reduce it to

$$\min_{\mathbf{w},\mathbf{b}} \max_{\boldsymbol{\mu}} \left\{\|\mathbf{b}\| + \boldsymbol{\mu}^T(\mathbf{b} - \mathbf{A}\mathbf{w} + \sigma\mathbf{v}) + p(\mathbf{w})\right\}.$$

The dual of the problem above is

$$\max_{\boldsymbol{\mu}} \min_{\mathbf{w},\mathbf{b}} \{\|\mathbf{b}\| + \boldsymbol{\mu}^T(\mathbf{b} - \mathbf{A}\mathbf{w} + \sigma\mathbf{v}) + p(\mathbf{w})\}. \tag{D.1}$$

The minimization over $\mathbf{b}$ above is easy to perform. A simple application of the Cauchy–Schwarz inequality gives

$$\begin{aligned}\|\mathbf{b}\| + \boldsymbol{\mu}^T\mathbf{b} &\geq \|\mathbf{b}\| - \|\mathbf{b}\|\|\boldsymbol{\mu}\|\\ &= (1 - \|\boldsymbol{\mu}\|)\,\|\mathbf{b}\|.\end{aligned}$$

Thus,

$$\min_{\mathbf{b}} \left\{ \|\mathbf{b}\| + \boldsymbol{\mu}^T \mathbf{b} \right\} = \begin{cases} 0 & , \|\boldsymbol{\mu}\| \leq 1, \\ -\infty & , o.w.. \end{cases}$$

Combining this with (D.1) we conclude that the dual problem of the problem in (5.7) is the following:

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

We equivalently rewrite the dual problem in the format of a minimization problem as follows:

$$- \min_{\|\boldsymbol{\mu}\| \leq 1} \max_{\mathbf{w}} \left\{ \boldsymbol{\mu}^T \left(\mathbf{A}\mathbf{w} - \sigma \mathbf{v} \right) - p(\mathbf{w}) \right\}. \tag{D.2}$$

If $p(\mathbf{w})$ is a finite convex function from $\mathbb{R}^n \to \mathbb{R}$, the problem in (5.7) is convex and satisfies Slater's conditions. When $p(\mathbf{w})$ is the indicator function of a convex set $\{\mathbf{w} | g(\mathbf{w}) \leq 0\}$, the problem can be viewed as $\min_{g(\mathbf{w}) \leq 0, \mathbf{b}} \left\{ \|\mathbf{b}\| + \boldsymbol{\mu}^T \left(\mathbf{b} - \mathbf{A}\mathbf{w} + \sigma \mathbf{v} \right) \right\}$. 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 (5.7) and (D.2) have the same optimal cost $\mathcal{F}(\mathbf{A}, \mathbf{v})$.

E. Proofs for Section 6

E.1. Proof of Lemma 6.1

We prove the statements of the Lemma in the order that they appear.

E.1.1 Scalarization

The first statement of Lemma 6.1 claims that the optimization problem in (6.4) can be reduced into a one dimensional optimization problem. To see this begin by evaluating the optimization over $\mathbf{w}$ for fixed $\|\mathbf{w}\|$:

$$\begin{aligned}
\hat{\mathcal{L}}(\mathbf{g}, \mathbf{h}) &= \min_{\mathbf{w}} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2} \|\mathbf{g}\| - \mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} \\
&= \min_{\substack{\mathbf{w}: \|\mathbf{w}\| = \alpha \\ \alpha \geq 0}} \left\{ \sqrt{\|\mathbf{w}\|^2 + \sigma^2} \|\mathbf{g}\| - \mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} \\
&= \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + \sigma^2} \|\mathbf{g}\| + \min_{\mathbf{w}: \|\mathbf{w}\| = \alpha} \left\{ -\mathbf{h}^T \mathbf{w} + \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} \right\} \\
&= \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + \sigma^2} \|\mathbf{g}\| - \max_{\mathbf{w}: \|\mathbf{w}\| = \alpha} \left\{ \mathbf{h}^T \mathbf{w} - \min_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} \right\} \\
&= \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + \sigma^2} \|\mathbf{g}\| - \max_{\mathbf{w}: \|\mathbf{w}\| = \alpha} \min_{\mathbf{s} \in \mathcal{C}} \left\{ (\mathbf{h} - \mathbf{s})^T \mathbf{w} \right\} \right\}
\end{aligned} \tag{E.1}$$

To further simplify (E.1), we use the following key observation as summarized in the Lemma below.

Lemma E.1. *Let $\mathcal{C} \in \mathbb{R}^n$ be a nonempty convex set in $\mathbb{R}^n$, $\mathbf{h} \in \mathbb{R}^n$ and $\alpha \geq 0$. Then,*

$$\max_{\mathbf{w}: \|\mathbf{w}\| = \alpha} \min_{\mathbf{s} \in \mathcal{C}} \left\{ (\mathbf{h} - \mathbf{s})^T \mathbf{w} \right\} = \min_{\mathbf{s} \in \mathcal{C}} \max_{\mathbf{w}: \|\mathbf{w}\| = \alpha} \left\{ (\mathbf{h} - \mathbf{s})^T \mathbf{w} \right\}.$$

Thus,

$$\max_{\mathbf{w}: \|\mathbf{w}\| = \alpha} \min_{\mathbf{s} \in \mathcal{C}} \left\{ (\mathbf{h} - \mathbf{s})^T \mathbf{w} \right\} = \alpha \cdot dist(\mathbf{h}, \mathcal{C}),$$

and the optimum is attained at $\mathbf{w}^* = \alpha \cdot \frac{\Pi(\mathbf{h}, \mathcal{C})}{dist(\mathbf{h}, \mathcal{C})}$.

Proof. First notice that

$$\min_{\mathbf{s}\in\mathcal{C}} \max_{\mathbf{w}:\|\mathbf{w}\|=\alpha} (\mathbf{h}-\mathbf{s})^T\mathbf{w} = \min_{\mathbf{s}\in\mathcal{C}} \alpha\|\mathbf{h}-\mathbf{s}\| = \alpha\cdot\text{dist}(\mathbf{h},\mathcal{C}).$$

Furthermore, MinMax is never less than MaxMin [83]. Thus,

$$\max_{\mathbf{w}:\|\mathbf{w}\|=\alpha} \min_{\mathbf{s}\in\mathcal{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} \;\leq\; \min_{\mathbf{s}\in\mathcal{C}} \max_{\mathbf{w}:\|\mathbf{w}\|=\alpha} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} = \alpha\cdot\text{dist}(\mathbf{h},\mathcal{C}).$$

It suffices to prove that

$$\max_{\mathbf{w}:\|\mathbf{w}\|=\alpha} \min_{\mathbf{s}\in\mathcal{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} \;\geq\; \alpha\cdot\text{dist}(\mathbf{h},\mathcal{C}).$$

Consider $\mathbf{w}^* = \alpha\cdot\frac{\Pi(\mathbf{h},\mathcal{C})}{\text{dist}(\mathbf{h},\mathcal{C})}$. Clearly,

$$\max_{\mathbf{w}:\|\mathbf{w}\|=\alpha} \min_{\mathbf{s}\in\mathcal{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}\right\} \;\geq\; \min_{\mathbf{s}\in\mathcal{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}^*\right\}.$$

But,

$$\begin{aligned}
\min_{\mathbf{s}\in\mathcal{C}} \left\{(\mathbf{h}-\mathbf{s})^T\mathbf{w}^*\right\} &= \frac{\alpha}{\text{dist}(\mathbf{h},\mathcal{C})}\cdot\left(\mathbf{h}^T\Pi(\mathbf{h},\mathcal{C}) - \max_{\mathbf{s}\in\mathcal{C}} \mathbf{s}^T\Pi(\mathbf{h},\mathcal{C})\right) &\text{(E.2)}\\
&= \frac{\alpha}{\text{dist}(\mathbf{h},\mathcal{C})}\cdot\left(\mathbf{h}^T\Pi(\mathbf{h},\mathcal{C}) - \text{Proj}(\mathbf{h},\mathcal{C})^T\Pi(\mathbf{h},\mathcal{C})\right) &\text{(E.3)}\\
&= \alpha\cdot\text{dist}(\mathbf{h},\mathcal{C}),
\end{aligned}$$

where (E.3) follows from Fact A.2. This completes the proof of the Lemma. □

Applying the result of Lemma E.1 to (E.1), we conclude that

$$\begin{aligned}
\hat{\mathcal{L}}(\mathbf{g},\mathbf{h}) &= \min_{\mathbf{w}} \left\{\sqrt{\|\mathbf{w}\|^2+\sigma^2}\|\mathbf{g}\| - \mathbf{h}^T\mathbf{w} + \max_{\mathbf{s}\in\mathcal{C}} \mathbf{s}^T\mathbf{w}\right\}\\
&= \min_{\alpha\geq 0} \left\{\sqrt{\alpha^2+\sigma^2}\|\mathbf{g}\| - \alpha\cdot\text{dist}(\mathbf{h},\mathcal{C})\right\} &\text{(E.4)}
\end{aligned}$$

E.1.2 Deterministic Result

The optimization problem in (E.4) 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

$$\|\mathbf{g}\|^2 > \text{dist}(\mathbf{h},\mathcal{C})^2, \tag{E.5}$$

it only takes a few simple calculations to prove the second statement of Lemma 6.1, i.e.

$$(\alpha^*)^2 = \|\mathbf{w}^*_{low}(\mathbf{g},\mathbf{h})\|^2 = \sigma^2\frac{\text{dist}^2(\mathbf{h},\mathcal{C})}{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h},\mathcal{C})}$$

and,

$$\hat{\mathcal{L}}(\mathbf{g},\mathbf{h}) = \sigma\sqrt{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h},\mathcal{C})}. \tag{E.6}$$

E.1.3 Probabilistic Result

Next, we prove the high probability lower bound for $\hat{\mathcal{L}}(\mathbf{g},\mathbf{h})$ implied by the last statement of Lemma 6.1. To do this, we will make use of concentration results for specific functions of Gaussian vectors as they are stated in Lemma B.3. Setting $t = \delta\sqrt{m}$ in Lemma B.3, with probability $1 - 8\exp(-c_0\delta^2 m)$,

$$\begin{aligned}
&|\|\mathbf{g}\|^2 - m| \leq 2\delta m + \delta^2 m + 1,\\
&|\text{dist}^2(\mathbf{h},\mathcal{C}) - \mathbf{D}(\mathcal{C})| \leq 2\delta\sqrt{\mathbf{D}(\mathcal{C})m} + \delta^2 m + 1 \leq 2\delta m + \delta^2 m + 1.
\end{aligned}$$

Combining these and using the assumption that $m \geq \mathbf{D}(\mathcal{C}) + \epsilon_L m$, we find that

$$\begin{aligned}\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \mathcal{C}) &\geq m - \mathbf{D}(\mathcal{C}) - [(2\delta^2 + 4\delta)m + 2] \\ &\geq m - \mathbf{D}(\mathcal{C}) - [(2\delta^2 + 4\delta)\frac{m - \mathbf{D}(\mathcal{C})}{\epsilon_L} + 2] \\ &\geq (m - \mathbf{D}(\mathcal{C}))[1 - \frac{(2\delta^2 + 4\delta)}{\epsilon_L}] - 2,\end{aligned}$$

with the same probability. Choose ϵ' so that $\sqrt{1-\epsilon'} = 1 - \epsilon$. Also, choose δ such that $\frac{(2\delta^2+4\delta)}{\epsilon_L} < \frac{\epsilon'}{2}$ and m sufficiently large to ensure $\epsilon_L \epsilon' m > 4$. Combined,

$$\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \mathcal{C}) \geq (m - \mathbf{D}(\mathcal{C}))(1 - \frac{\epsilon'}{2}) - 2 \geq (m - \mathbf{D}(\mathcal{C}))(1 - \epsilon'), \tag{E.7}$$

with probability $1 - 8\exp(-c_0\delta^2 m)$. Since the right hand side in (E.7) is positive, it follows from the second statement of Lemma 6.1 that

$$\hat{\mathcal{L}}(\mathbf{g}, \mathbf{h}) \geq \sigma\sqrt{(m - \mathbf{D}(\mathcal{C}))(1 - \epsilon')} = \sigma(1 - \epsilon)\sqrt{m - \mathbf{D}(\mathcal{C})},$$

with the same probability. This concludes the proof.

E.2. Proof of Lemma 6.2

E.2.1 Scalarization

We have

$$\begin{aligned}\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h}) &= -\min_{\|\mu\| \leq 1} \max_{\|\mathbf{w}\| = C_{up}} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + \|\boldsymbol{\mu}\| \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} \\ &= -\min_{\|\mu\| \leq 1} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + \max_{\|\mathbf{w}\| = C_{up}} \left\{ \|\boldsymbol{\mu}\| \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} \right\}. \end{aligned} \tag{E.8}$$

Notice that

$$\begin{aligned}\max_{\|\mathbf{w}\| = C_{up}} \left\{ \|\boldsymbol{\mu}\| \mathbf{h}^T \mathbf{w} - \max_{\mathbf{s} \in \mathcal{C}} \mathbf{s}^T \mathbf{w} \right\} &= \max_{\|\mathbf{w}\| = C_{up}} \min_{\mathbf{s} \in \mathcal{C}} (\|\boldsymbol{\mu}\| \mathbf{h} - \mathbf{s})^T \mathbf{w} \\ &= C_{up} \text{dist}(\|\boldsymbol{\mu}\| \mathbf{h}, \mathcal{C}). \end{aligned} \tag{E.9}$$

where (E.9) follows directly from Lemma E.1. Combine (E.8) and (E.9) to conclude that

$$\begin{aligned}\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h}) &= -\min_{\|\mu\| \leq 1} \left\{ \sqrt{C_{up}^2 + \sigma^2}\, \mathbf{g}^T \boldsymbol{\mu} + C_{up} \text{dist}(\|\boldsymbol{\mu}\| \mathbf{h}, \mathcal{C}) \right\} \\ &= -\min_{0 \leq \alpha \leq 1} \left\{ -\alpha \cdot \sqrt{C_{up}^2 + \sigma^2}\, \|\mathbf{g}\| + C_{up} \text{dist}(\alpha \mathbf{h}, \mathcal{C}) \right\}. \end{aligned} \tag{E.10}$$

E.2.2 Deterministic Result

For convenience denote the objective function of problem (E.10) as

$$\phi(\alpha) = C_{up} \text{dist}(\alpha \mathbf{h}, \mathcal{C}) - \alpha \sqrt{C_{up}^2 + \sigma^2} \|\mathbf{g}\|.$$

Notice that $\phi(\cdot)$ is convex. By way of justification, $\text{dist}(\alpha \mathbf{h}, \mathcal{C})$ is a convex function for $\alpha \geq 0$ [86], and $\alpha\sqrt{C^2 + \sigma^2}\|\mathbf{g}\|$ 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 E.2. *Let C be a nonempty closed and convex set and $\mathbf{h} \notin C$. Then*

$$\lim_{\epsilon \to 0} \frac{dist(\mathbf{h} + \epsilon \mathbf{h}, C) - dist(\mathbf{h}, C)}{\epsilon} = \langle \mathbf{h}, \frac{\Pi(\mathbf{h}, C)}{\|\Pi(\mathbf{h}, C)\|} \rangle,$$

Proof. Let H be a hyperplane of $\mathcal{C}$ at $\text{Proj}(\mathbf{h}, \mathcal{C})$ orthogonal to $\Pi(\mathbf{h}, C)$. Using the second statement of Fact A.2, H is a supporting hyperplane and $\mathbf{h}$ and C lie on different half planes induced by H (also see [83]). Also, observe that $\Pi(\mathbf{h}, \mathcal{C}) = \Pi(\mathbf{h}, H)$ and $\text{Proj}(\mathbf{h}, \mathcal{C}) = \text{Proj}(\mathbf{h}, H)$. Choose $\epsilon > 0$ sufficiently small such that $(1+\epsilon)\mathbf{h}$ lies on the same half-plane as $\mathbf{h}$. We then have,

$$\|\Pi((1+\epsilon)\mathbf{h}, \mathcal{C})\| \geq \|\Pi((1+\epsilon)\mathbf{h}, H)\| = \|\Pi(\mathbf{h}, \mathcal{C})\| + \left\langle \epsilon \mathbf{h}, \frac{\Pi(\mathbf{h}, \mathcal{C})}{\|\Pi(\mathbf{h}, \mathcal{C})\|} \right\rangle. \tag{E.11}$$

Denote the $n-1$ dimensional subspace that is orthogonal to $\Pi(\mathbf{h}, H)$ and parallel to H by H_0. Decomposing $\epsilon \mathbf{h}$ to its orthonormal components along $\Pi(\mathbf{h}, H)$ and H_0, we have

$$\|\Pi((1+\epsilon)\mathbf{h}, C)\|^2 \leq \|(1+\epsilon)\mathbf{h} - \text{Proj}(\mathbf{h}, C)\|^2 = \left(\|\Pi(\mathbf{h}, C)\| + \left\langle \epsilon \mathbf{h}, \frac{\Pi(\mathbf{h}, C)}{\|\Pi(\mathbf{h}, C)\|} \right\rangle \right)^2 + \epsilon^2 \|\text{Proj}(\mathbf{h}, H_0)\|^2. \tag{E.12}$$

Take square roots in both sides of (E.12) 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 (E.11) and let $\epsilon \to 0$ to conclude the proof. □

Since $\mathbf{h} \notin \mathcal{C}$, it follows from Lemma E.2, that $\text{dist}(\alpha \mathbf{h}, \mathcal{C})$ is differentiable as a function of α at $\alpha = 1$, implying the same result for $\phi(\alpha)$. In fact, we have

$$\phi'(1) = C_{up} \text{dist}(\mathbf{h}, \mathcal{C}) + C_{up} \frac{\langle \Pi(\mathbf{h}, \mathcal{C}), \text{Proj}(\mathbf{h}, \mathcal{C}) \rangle}{\text{dist}(\mathbf{h}, \mathcal{C})} - \sqrt{C_{up}^2 + \sigma^2} \|\mathbf{g}\| < 0,$$

where the negativity follows from assumption (6.6). To conclude the proof, we make use of the following simple lemma.

Lemma E.3. *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 E.3 for the convex function $\phi(\cdot)$ at $\alpha = 1$, gives that $\phi(\alpha) \geq \phi(1)$ for all $\alpha \in [0, 1]$. Therefore, $\alpha^* = 1$.

E.2.3 Probabilistic Result

We consider the setting where m is sufficiently large and,

$$(1 - \epsilon_L)m \geq \max\left(\mathbf{D}(\mathcal{C}) + \mathbf{C}(\mathcal{C}), \mathbf{D}(\mathcal{C})\right), \quad \mathbf{D}(\mathcal{C}) \geq \epsilon_L m \tag{E.13}$$

Choose $C_{up} = \sigma \sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}}$ which would give $C_{up}^2 + \sigma^2 = \sigma^2 \frac{m}{m - \mathbf{D}(\mathcal{C})}$. Hence, the assumption (6.6) in the second statement of Lemma 6.2 can be rewritten as,

$$\sqrt{m} \|\mathbf{g}\| \text{dist}(\mathbf{h}, \mathcal{C}) > \sqrt{\mathbf{D}(\mathcal{C})} (\text{dist}(\mathbf{h}, \mathcal{C})^2 + \text{corr}(\mathbf{h}, \mathcal{C})). \tag{E.14}$$

The proof technique is as follows. We first show that (E.14) (and thus (6.6)) holds with high probability. Also, that $\mathbf{h} \notin \mathcal{C}$ with high probability. Then, as a last step we make use of the second statement of Lemma 6.2 to compute the lower bound on $\hat{\mathcal{U}}$.

• (6.6) *holds with high probability:*

Using standard concentration arguments (see Lemma B.2), we have

$$\sqrt{m}\|\mathbf{g}\|\text{dist}(\mathbf{h},\mathcal{C}) \geq \sqrt{m}(\sqrt{m-1}-t)(\sqrt{\mathbf{D}(\mathcal{C})-1}-t)$$

with probability $1-4\exp\left(\frac{-t^2}{2}\right)$. Choose a sufficiently small constant $\delta > 0$ and set $t = \delta\sqrt{\mathbf{D}(\mathcal{C})}$ to ensure,

$$\sqrt{m}\|\mathbf{g}\|\text{dist}(\mathbf{h},\mathcal{C}) \geq (1-\frac{\epsilon_L}{2})m\sqrt{\mathbf{D}(\mathcal{C})} \tag{E.15}$$

with probability $1-\exp(-\mathcal{O}(m))$, where we used $(1-\epsilon_L) \geq \mathbf{D}(\mathcal{C}) \geq \epsilon_L m$. In particular, for sufficiently large $\mathbf{D}(\mathcal{C})$ we need $(1-\delta)^2 > 1-\frac{\epsilon_L}{2}$.

Equation (E.15) establishes a high probability lower bound for the expression at the left hand side of (E.14). Next, we show that the expression at the right hand side of (E.14) is upper bounded with high probability by the same quantity.

Case 1: If $\mathcal{C}$ is a cone, $\text{corr}(\mathbf{h},\mathcal{C}) = 0$ and using Lemma B.3 $\text{dist}(\mathbf{h},\mathcal{C})^2 \leq \mathbf{D}(\mathcal{C}) + 2t\sqrt{\mathbf{D}(\mathcal{C})} + t^2 \leq (1-\epsilon_L)m + 2t\sqrt{m} + t^2$ with probability $1-2\exp(-\frac{t^2}{2})$. Hence, we can choose $t = \epsilon\sqrt{m}$ for a small constant $\epsilon > 0$ to ensure, $\text{dist}(\mathbf{h},\mathcal{C})^2 < (1-\frac{\epsilon_L}{2})m$ with probability $1-\exp(-\mathcal{O}(m))$. This gives (E.14) in combination with (E.15).

Case 2: Otherwise, from Lemma B.4, we have that $\mathbf{P}(\mathcal{C}) \leq 2(n+\mathbf{D}(\mathcal{C}))$ and from (E.13), $m \geq \mathbf{D}(\mathcal{C})$. Then, applying Lemma B.3, we have

$$\begin{aligned}
\text{dist}(\mathbf{h},\mathcal{C})^2 + \text{corr}(\mathbf{h},\mathcal{C}) &\leq \mathbf{D}(\mathcal{C}) + \mathbf{C}(\mathcal{C}) + 3t\underbrace{\sqrt{\mathbf{D}(\mathcal{C})}}_{\leq\sqrt{m}} + t\underbrace{\sqrt{\mathbf{P}(\mathcal{C})}}_{\leq\sqrt{2(n+m)}} + 2(t^2+1) \\
&\leq \mathbf{D}(\mathcal{C}) + \mathbf{C}(\mathcal{C}) + 3t\sqrt{m} + t\sqrt{2(n+m)} + 2(t^2+1) \\
&\leq (1-\epsilon_L)m + 3t\sqrt{m} + t\sqrt{2(n+m)} + 2(t^2+1).
\end{aligned}$$

with probability $1-4\exp\left(\frac{-t^2}{2}\right)$. Therefore, with the same probability,

$$\sqrt{\mathbf{D}(\mathcal{C})}(\text{dist}(\mathbf{h},\mathcal{C})^2 + \text{corr}(\mathbf{h},\mathcal{C})) \leq (1-\epsilon_L)m\sqrt{\mathbf{D}(\mathcal{C})} + 3t\sqrt{m}\sqrt{\mathbf{D}(\mathcal{C})} + t\sqrt{2(n+m)}\sqrt{\mathbf{D}(\mathcal{C})} + 2(t^2+1)\sqrt{\mathbf{D}(\mathcal{C})} \tag{E.16}$$

Comparing the right hand sides of inequalities E.15 and E.16 , we need to ensure that,

$$3t\sqrt{m}\sqrt{\mathbf{D}(\mathcal{C})} + t\sqrt{2(n+m)}\sqrt{\mathbf{D}(\mathcal{C})} + 2(t^2+1)\sqrt{\mathbf{D}(\mathcal{C})} \leq \frac{\epsilon_L}{2}m\sqrt{\mathbf{D}(\mathcal{C})} \iff 3t\sqrt{m} + t\sqrt{2(n+m)} + 2(t^2+1) \leq \frac{\epsilon_L}{2}m. \tag{E.17}$$

Choose $t = \epsilon\min\{\sqrt{m}, \frac{m}{\sqrt{n}}\}$ for sufficiently small ϵ such that (E.17) and (E.14) then hold with probability $1-\exp\left(-\mathcal{O}\left(\min\{\frac{m^2}{n}, m\}\right)\right)$.

Combining Case 1 and Case 2, (E.14) holds with probability $1-\exp\left(-\mathcal{O}\left(\gamma(m,n)\right)\right)$ where $\gamma(m,n) = m$ when $\mathcal{C}$ is cone and $\gamma(m,n) = \min\{\frac{m^2}{n}, m\}$ otherwise.

• $\mathbf{h} \notin \mathcal{C}$ *with high probability:*

Apply Lemma B.2 on $\text{dist}(\mathbf{h},\mathcal{C})$ with $t = \epsilon\sqrt{\mathbf{D}(\mathcal{C})}$ to show that $\text{dist}(\mathbf{h},\mathcal{C})$ is strictly positive. This proves that $\mathbf{h} \notin \mathcal{C}$, with probability $1-\exp(-\mathcal{O}(\mathbf{D}(\mathcal{C})))$=$1-\exp(-\mathcal{O}(m))$.

• *High probability lower bound for* $\hat{\mathcal{U}}$*:*

Thus far we have proved that assumptions $\mathbf{h} \notin \mathcal{C}$ and (6.6) of the second statement in Lemma 6.2 hold with the desired probability. Therefore, (6.7) holds with the same high probability, namely,

$$\hat{\mathcal{U}}(\mathbf{g},\mathbf{h}) = \frac{\sigma}{\sqrt{m-\mathbf{D}(\mathcal{C})}}\left(\sqrt{m}\|\mathbf{g}\| - \sqrt{\mathbf{D}(\mathcal{C})}\text{dist}(\mathbf{h},\mathcal{C})\right) \tag{E.18}$$

We will use similar concentration arguments as above to upper bound the right hand side of (E.18). For any $t > 0$:

$$\sqrt{m}\|\mathbf{g}\| \leq m + t\sqrt{m}$$
$$\sqrt{\mathbf{D}(\mathcal{C})}\text{dist}(\mathbf{h}, \mathcal{C}) \geq \sqrt{\mathbf{D}(\mathcal{C})}(\sqrt{\mathbf{D}(\mathcal{C}) - 1} - t)$$

with probability $1 - 4\exp(-\frac{t^2}{2})$. Thus,

$$\sqrt{m}\|\mathbf{g}\| - \sqrt{\mathbf{D}(\mathcal{C})}\text{dist}(\mathbf{h}, \mathcal{C}) \leq m - \mathbf{D}(\mathcal{C}) + t(\sqrt{m} + \sqrt{\mathbf{D}(\mathcal{C})}) + 1. \tag{E.19}$$

For a given constant $\epsilon > 0$, substitute (E.19) in (E.18) and choose $t = \epsilon'\sqrt{m}$ (for some sufficiently small constant $\epsilon' > 0$), to ensure that,

$$\hat{\mathcal{U}}(\mathbf{g}, \mathbf{h}) \leq (1 + \epsilon)\sigma\sqrt{m - \mathbf{D}(\mathbf{x}_0, \lambda)}$$

with probability $1 - 4\exp\left(\frac{-\epsilon'^2 m}{2}\right)$. Combining this with the high probability events of all previous steps, we obtain the desired result.

E.3. Proof of Lemma 6.3

E.3.1 Scalarization

The reduction of $\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h})$ to an one-dimensional optimization problem follows identically the steps as in the proof for $\hat{\mathcal{L}}(\mathbf{g}, \mathbf{h})$ in Section E.1.1.

E.3.2 Deterministic Result

From the first statement of Lemma 6.3,

$$\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) = \min_{\alpha \in S_{dev}} \left\{ \underbrace{\sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\| - \alpha \cdot \text{dist}(\mathbf{h}, \mathcal{C})}_{:= L(\alpha)} \right\}, \tag{E.20}$$

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 F.1) to prove that $L(\cdot)$:

- is a strictly convex function,
- attains its minimum at

$$\alpha^*(\mathbf{g}, \mathbf{h}) = \frac{\sigma \cdot \text{dist}(\mathbf{h}, \mathcal{C})}{\sqrt{\|\mathbf{g}\|^2 - \text{dist}^2(\mathbf{h}, \mathcal{C})}}.$$

The minimization of $L(\alpha)$ in (E.20) is restricted to the set S_{dev}. Also, by assumption (6.9), $\alpha^*(\mathbf{g}, \mathbf{h}) \notin S_{dev}$. 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 $(1 \pm \delta_{dev})C_{dev}$ [83]. Thus, $\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) = L((1 \pm \delta_{dev})C_{dev})$, which completes the proof.

E.3.3 Probabilistic Result

Choose $C_{dev} = \sigma\sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}}$ and consider the regime where $(1 - \epsilon_L)m > \mathbf{D}(\mathcal{C}) > \epsilon_L m$ for some constant $\epsilon_L > 0$. $\delta_{dev} > 0$ is also a constant.

- **Mapping $\hat{\mathcal{L}}_{dev}$ to Lemma F.1:** It is helpful for the purposes of the presentation to consider the function

$$L(x) := L(x; a, b) = \sqrt{x^2 + \sigma^2}a - xb, \tag{E.21}$$

over $x \geq 0$, and a, b are positive parameters. Substituting a, b, x with $\|\mathbf{g}\|, \text{dist}(\mathbf{h}, \mathcal{C}), \alpha$, we can map $L(x; a, b)$ to our function of interest,

$$L(\alpha; \|\mathbf{g}\|, \text{dist}(\mathbf{h}, \mathcal{C})) = \sqrt{\alpha^2 + \sigma^2}\|\mathbf{g}\| - \alpha\text{dist}(\mathbf{h}, \mathcal{C}).$$

In Lemma F.1 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 $\|\mathbf{g}\|, \mathrm{dist}(\mathbf{h}, \mathcal{C})$ which are probabilistic quantities and a, b correspond to $\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}$, i.e. the approximate means of the former ones.

In what follows, we refer continuously to statements of Lemma F.1 and use them to complete the proof of the "Probabilistic result" of Lemma 6.3. 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 F.1 that,

$$L\left(x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}); \sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}\right) = \sigma\sqrt{m - \mathbf{D}(\mathcal{C})}, \tag{E.22}$$

and

$$x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}) = \sigma\sqrt{\frac{\mathbf{D}(\mathcal{C})}{m - \mathbf{D}(\mathcal{C})}} = C_{dev}, \tag{E.23}$$

• **Verifying assumption (6.9):** Going back to the proof, we begin by proving that assumption (6.9) of the second statement of Lemma 6.3 is valid with high probability. Observe that from the definition of S_{dev} and (E.23), assumption (6.9) can be equivalently written as

$$\left|\frac{x^*(\|\mathbf{g}\|, \mathrm{dist}(\mathbf{h}, \mathcal{C}))}{x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})})} - 1\right| \leq \delta_{dev}. \tag{E.24}$$

On the other hand, from the third statement of Lemma F.1 there exists sufficiently small constant $\epsilon_1 > 0$ such that (E.24) is true for all $\mathbf{g}$ and $\mathbf{h}$ satisfying

$$|\|\mathbf{g}\| - \sqrt{m}| \leq \epsilon_1\sqrt{m} \quad \text{and} \quad |\mathrm{dist}(\mathbf{h}, \mathcal{C}) - \sqrt{\mathbf{D}(\mathcal{C})}| \leq \epsilon_1\sqrt{m}. \tag{E.25}$$

Furthermore, for large enough $\mathbf{D}(\mathcal{C})$ and from basic concentration arguments (see Lemma B.2), $\mathbf{g}$ and $\mathbf{h}$ satisfy (E.25) with probability $1 - 2\exp(-\frac{\epsilon_1^2 m}{2})$. This proves that assumption (6.9) holds with the same high probability.

• **Lower bounding $\hat{\mathcal{L}}_{dev}$:** From the deterministic result of Lemma 6.3, once (6.9) is satisfied then

$$\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) = L\left((1 \pm \delta_{dev})C_{dev}; \|\mathbf{g}\|, \mathrm{dist}(\mathbf{h}, \mathcal{C})\right). \tag{E.26}$$

Thus, to prove (6.10) we will show that there exists $t > 0$ such that

$$L\left((1 \pm \delta_{dev})C_{dev}; \|\mathbf{g}\|, \mathrm{dist}(\mathbf{h}, \mathcal{C})\right) \geq (1 + t)\sigma\sqrt{m - \mathbf{D}(\mathcal{C})}, \tag{E.27}$$

with high probability. Equivalently, using (E.22), it suffices to show that there exists a constant $t > 0$ such that

$$L\left((1 \pm \delta_{dev})x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}); \|\mathbf{g}\|, \mathrm{dist}(\mathbf{h}, \mathcal{C})\right) - L\left(x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}); \sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}\right) \geq t\sigma\sqrt{m}, \tag{E.28}$$

with high probability. Applying the sixth statement of Lemma F.1 with $\gamma \leftarrow \delta_{dev}$, for any constant $\delta_{dev} > 0$, there exists constants t, ϵ_2 such that (E.28) holds for all $\mathbf{g}$ and $\mathbf{h}$ satisfying

$$|\|\mathbf{g}\| - \sqrt{m}| \leq \epsilon_2\sqrt{m} \quad \text{and} \quad |\mathrm{dist}(\mathbf{h}, \mathcal{C}) - \sqrt{\mathbf{D}(\mathcal{C})}| \leq \epsilon_2\sqrt{m},$$

which holds with probability $1 - 2\exp(-\frac{\epsilon_2^2 m}{2})$ for sufficiently large $\mathbf{D}(\mathcal{C})$. Thus, (E.28) is true with the same high probability.

Union bounding over the events that (E.24) and (E.28) are true, we end up with the desired result. The reason is that with high probability (E.26) and (E.28) hold, i.e.,

$$\hat{\mathcal{L}}_{dev}(\mathbf{g}, \mathbf{h}) = L\left((1 \pm \delta_{dev})C_{dev}; \|\mathbf{g}\|, \mathrm{dist}(\mathbf{h}, \mathcal{C})\right) \geq L\left(x^*(\sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}); \sqrt{m}, \sqrt{\mathbf{D}(\mathcal{C})}\right) + t\sigma\sqrt{m} = \sigma\sqrt{m - \mathbf{D}(\mathcal{C})} + t\sigma\sqrt{m}.$$

F. Deviation Analysis: Key Lemma

Lemma F.1. *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 $(1-\epsilon)a > b > \epsilon a$ for some constant $\epsilon > 0$. 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 $\epsilon_1 > 0$, such that*

$$\left|\frac{x^*(a', b')}{x^*(a, b)} - 1\right| \leq \eta,$$

   for all a', b' satisfying $|a' - a| < \epsilon_1 a$ and $|b' - b| < \epsilon_1 a$.

4. *There exists positive constant $\eta > 0$, such that, for sufficiently small constant $\epsilon_1 > 0$,*

$$\left|L(x^*(a, b); a', b') - L(x^*(a, b), a, b)\right| \leq \eta\epsilon_1\sigma a,$$

   for all a', b' satisfying $|a' - a| < \epsilon_1 a$ and $|b' - b| < \epsilon_1 a$.

5. *For any constant $\gamma > 0$, there exists a constant $\epsilon_2 > 0$ such that for sufficiently small constant $\epsilon_1 > 0$,*

$$L(x; a', b') - L(x^*(a, b); a', b') \geq \epsilon_2\sigma a,$$

   for all x, a' and b' satisfying $|x - x^*(a, b)| > \gamma x^*(a, b)$, $|a' - a| < \epsilon_1 a$ and $|b' - b| < \epsilon_1 a$.

6. *For any constant $\gamma > 0$, there exists a constant $\epsilon_2 > 0$ such that for sufficiently small constant $\epsilon_1 > 0$,*

$$L(x; a', b') - L(x^*(a, b); a, b) \geq \epsilon_2\sigma a,$$

   for all x, a' and b' satisfying $|x - x^*(a, b)| > \gamma x^*(a, b)$, $|a' - a| < \epsilon_1 a$ and $|b' - b| < \epsilon_1 a$.

7. *Given $c_{low} > 0$, consider the restricted optimization, $\min_{x \geq c_{low}} L(x; a, b)$. We have,*

$$\lim_{c_{low} \to \infty} \min_{x \geq c_{low}} L(x; a, b) \to \infty \tag{F.1}$$

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 $|a'-a| < \epsilon_1 a, |b'-b| < \epsilon_1 a$. Therefore,

$$\frac{b-\epsilon_1 a}{\sqrt{(a+\epsilon_1 a)^2-(b-\epsilon_1 a)^2}} \leq \frac{b'}{\sqrt{a'^2-b'^2}} \leq \frac{b+\epsilon_1 a}{\sqrt{(a-\epsilon_1 a)^2-(b+\epsilon_1 a)^2}}.$$

Now, for any constant $\delta > 0$, we can choose ϵ_1 sufficiently small such that both $b-\epsilon_1 a$ and $b+\epsilon_1 a$ lie in the interval $(1\pm\delta)b$. Similarly, $(a\pm\epsilon_1 a)^2-(b\mp\epsilon_1 a)^2$ can be also chosen to lie in the interval $(1\pm\delta)(a^2-b^2)$. Combining, we obtain,

$$\left|\frac{b'}{\sqrt{a'^2-b'^2}} - \frac{b}{\sqrt{a^2-b^2}}\right| < \eta(\delta)\frac{b}{\sqrt{a^2-b^2}},$$

as desired.

Fourth statement: For $|a-a'| < \epsilon_1 a$ and $|b-b'| < \epsilon_1 a$, we have,

$$|L(x^*(a,b);a',b') - L(x^*(a,b);a,b)| = \frac{\sigma}{\sqrt{a^2-b^2}}|(aa'-bb')-(a^2-b^2)| \leq \epsilon_1\sigma\frac{|a^2+ab|}{a^2-b^2}.$$

By assumption, $(1-\epsilon)a > b > \epsilon a$. Thus,

$$\epsilon_1\sigma\frac{|a^2+ab|}{a^2-b^2} \leq \epsilon_1\sigma\frac{2a^2}{2\epsilon a^2} = \frac{\epsilon_1\sigma}{\epsilon}.$$

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 $|a'-a| < \epsilon_1 a$ and $|b'-b| < \epsilon_1 a$. 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,

$$L((1\pm\gamma)x^*(a,b);a',b') = \frac{\sigma}{\sqrt{a^2-b^2}}(\sqrt{a^2+(\pm 2\gamma+\gamma^2)b^2}a' - (1\pm\gamma)bb'), \tag{F.2}$$

and

$$L(x^*(a,b);a',b') = \frac{\sigma}{\sqrt{a^2-b^2}}(aa'-bb'). \tag{F.3}$$

In the following discussion, without loss of generality, we consider only the "$+\gamma$" case in (F.2) since the exact same argument works for the "$-\gamma$" case as well.

Subtracting (F.3) from (F.2) and discarding the constant in front, we will focus on the following quantity,

$$\begin{aligned}\text{diff}(\gamma) &= (\sqrt{a^2+(2\gamma+\gamma^2)b^2}a' - (1+\gamma)bb') - (aa'-bb') \\ &= (\underbrace{\sqrt{a^2+(2\gamma+\gamma^2)b^2}}_{:=g(\gamma)} - a)a' - \gamma bb'.\end{aligned} \tag{F.4}$$

To find a lower bound for $g(\gamma)$, write

$$\begin{aligned}g(\gamma) &= \sqrt{a^2+(2\gamma+\gamma^2)b^2} \\ &= \sqrt{(a+\gamma\frac{b^2}{a})^2+\gamma^2(b^2-\frac{b^4}{a^2})} \\ &\geq (a+\gamma\frac{b^2}{a}) + \frac{\gamma^2(b^2-\frac{b^4}{a^2})}{4(a+\gamma\frac{b^2}{a})},\end{aligned} \tag{F.5}$$

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 (F.5) 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 (F.4) , we find that,

$$\text{diff}(\gamma) \geq \gamma(\frac{b^2}{a}a' - bb') + \gamma^2\frac{a^2b^2-b^4}{8a^3}a'. \tag{F.6}$$

Consider the second term on the right hand side of the inequality in (F.6). Choosing $\epsilon_1 < 1/2$, we ensure, $a' \geq a/2$, and thus,

$$\gamma^2\frac{a^2b^2-b^4}{8a^3}a' \geq \gamma^2\frac{a^2b^2-b^4}{16a^2} \geq \gamma^2\frac{\epsilon a^2b^2}{16a^2} = \gamma^2\epsilon\frac{b^2}{16}. \tag{F.7}$$

Next, consider the other term in (F.6). 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,

$$\left|\frac{b^2}{a}(a'-a)\right| + |b(b'-b)| < \gamma\epsilon\frac{b^2}{32}. \tag{F.8}$$

Combining (F.6), (F.7) and (F.8), we conclude that there exists sufficiently small constant $\epsilon_1 > 0$ such that,

$$\text{diff}(\gamma) \geq \gamma^2\epsilon\frac{b^2}{32}. \tag{F.9}$$

Multiplying with $\frac{\sigma}{\sqrt{a^2-b^2}}$, we end up with the desired result since $\frac{b^2}{\sqrt{a^2-b^2}} \geq \frac{\epsilon^2}{\sqrt{1-\epsilon^2}}a$.

Sixth statement: The last statement can be deduced from the fourth and fifth statements. Given $\gamma > 0$, choose $\epsilon_1 > 0$ sufficiently small to ensure,

$$L(x;a',b') - L(x^*(a,b),a',b') \geq \epsilon_2\sigma a \tag{F.10}$$

and

$$|L(x^*(a,b);a,b) - L(x^*(a,b),a',b')| \geq \eta\epsilon_1\sigma a \tag{F.11}$$

Using the triangle inequality,

$$\begin{aligned} L(x;a',b') - L(x^*(a,b),a,b) &\geq L(x;a',b') - L(x^*(a,b),a',b') - |L(x^*(a,b),a',b') - L(x^*(a,b),a,b)| \\ &\geq (\epsilon_2 - \eta\epsilon_1)\sigma a. \end{aligned} \tag{F.12}$$

Choosing ϵ_1 to further satisfy $\eta\epsilon_1 < \frac{\epsilon_2}{2}$, (F.12) is guaranteed to be larger than $\frac{\epsilon_2}{2}\sigma a$ which gives the desired result.

Seventh statement: To show this, we may use $a > b$ and simply write,

$$L(x;a,b) \geq (a-b)x \implies \lim_{c_{low}\to\infty}\min_{x\geq c_{low}} L(x;a,b) \geq \lim_{c_{low}\to\infty}(a-b)c_{low} = \infty \tag{F.13}$$

□

G. Proof of Lemma 8.1

Proof of the Lemma requires some work. We prove the statements in the specific order that they appear.

Statement 1: We have

$$\begin{aligned} n = \mathbb{E}\left[\|\mathbf{h}\|^2\right] = \mathbb{E}\left[\|\text{Proj}_\lambda(\mathbf{h}) + \mathbf{h} - \text{Proj}_\lambda(\mathbf{h})\|^2\right] &= \mathbb{E}[\|\text{Proj}_\lambda(\mathbf{h})\|^2] + \mathbb{E}[\|\Pi_\lambda(\mathbf{h})\|^2] + 2\mathbb{E}[\langle\Pi_\lambda(\mathbf{h}), \text{Proj}_\lambda(\mathbf{h})\rangle] \\ &= \mathbf{P}_f(\mathbf{x}_0,\lambda) + \mathbf{D}_f(\mathbf{x}_0,\lambda) + 2\mathbf{C}_f(\mathbf{x}_0,\lambda). \end{aligned}$$

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 $r = \inf_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \|\mathbf{s}\|$. Then, for any $\lambda \geq 0$, $\|\text{Proj}_\lambda(\mathbf{v})\| \geq \lambda\|\mathbf{s}\|$, which implies $\mathbf{P}_f(\mathbf{x}_0, \lambda) \geq \lambda^2\|\mathbf{s}\|^2$. Letting $\lambda \to \infty$, we find $\mathbf{P}_f(\mathbf{x}_0, \lambda) \to \infty$.

Similarly, for any $\mathbf{h}$, application of the triangle inequality gives

$$\|\Pi_\lambda(\mathbf{h})\| \geq \lambda r - \|\mathbf{h}\| \implies \|\Pi_\lambda(\mathbf{h})\|^2 \geq \lambda^2 r^2 - 2\lambda r\|\mathbf{h}\|.$$

Let $\mathbf{h} \sim \mathcal{N}(0, I)$ and take expectations in both sides of the inequality above. Recalling that $\mathbb{E}[\|\mathbf{h}\|] \leq \sqrt{n}$, and letting $\lambda \to \infty$, we find $\mathbf{D}_f(\mathbf{x}_0, \lambda) \to \infty$.

Finally, since $\mathbf{D}_f(\mathbf{x}_0, \lambda) + \mathbf{P}_f(\mathbf{x}_0, \lambda) + 2\mathbf{C}_f(\mathbf{x}_0, \lambda) = n$, $\mathbf{C}_f(\mathbf{x}_0, \lambda) \to -\infty$ as $\lambda \to \infty$. This completes the proof.
Statement 4: Continuity of $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ follows from Lemma $B.2$ in Amelunxen et al. [31]. We will now show continuity of $\mathbf{P}_f(\mathbf{x}_0, \lambda)$ and continuity of $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ will follow from the fact that $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ is a continuous function of $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{P}_f(\mathbf{x}_0, \lambda)$.

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,

$$\|\text{Proj}_\lambda(\mathbf{v}_1) - \text{Proj}_\lambda(\mathbf{v}_2)\| \leq \|\mathbf{v}_1 - \mathbf{v}_2\| \tag{G.1}$$

Consequently, given $\lambda_1, \lambda_2 > 0$,

$$\begin{aligned}
\|\text{Proj}_{\lambda_1}(\mathbf{v}) - \text{Proj}_{\lambda_2}(\mathbf{v})\| &= \|\lambda_1\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1}) - \lambda_2\text{Proj}_1(\frac{\mathbf{v}}{\lambda_2})\| &\text{(G.2)}\\
&\leq |\lambda_1 - \lambda_2|\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \|\lambda_2(\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1}) - \text{Proj}_1(\frac{\mathbf{v}}{\lambda_2}))\| &\text{(G.3)}\\
&\leq |\lambda_1 - \lambda_2|\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \lambda_2\|\mathbf{v}\|\frac{|\lambda_1 - \lambda_2|}{\lambda_1\lambda_2} &\text{(G.4)}\\
&= |\lambda_1 - \lambda_2|(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1}) &\text{(G.5)}
\end{aligned}$$

Hence, setting $\lambda_2 = \lambda_1 + \epsilon$,

$$\|\text{Proj}_{\lambda_2}(\mathbf{v})\|^2 \leq [\|\text{Proj}_{\lambda_1}(\mathbf{v})\| + \epsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1})]^2 \tag{G.6}$$

which implies,

$$\|\text{Proj}_{\lambda_2}(\mathbf{v})\|^2 - \|\text{Proj}_{\lambda_1}(\mathbf{v})\|^2 \leq 2\epsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1})\|\text{Proj}_{\lambda_1}(\mathbf{v})\| + \epsilon^2(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1}) \tag{G.7}$$

Similarly, using $\|\text{Proj}_{\lambda_2}(\mathbf{v})\| \geq \|\text{Proj}_{\lambda_1}(\mathbf{v})\| - \epsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1})$, we find,

$$\|\text{Proj}_{\lambda_1}(\mathbf{v})\|^2 - \|\text{Proj}_{\lambda_2}(\mathbf{v})\|^2 \leq 2\epsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1})\|\text{Proj}_{\lambda_1}(\mathbf{v})\|\lambda_1) \tag{G.8}$$

Combining these, we always have,

$$|\|\text{Proj}_{\lambda_2}(\mathbf{v})\|^2 - \|\text{Proj}_{\lambda_1}(\mathbf{v})\|^2| \leq 2\epsilon(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1})\|\text{Proj}_{\lambda_1}(\mathbf{v})\| + \epsilon^2(\|\text{Proj}_1(\frac{\mathbf{v}}{\lambda_1})\| + \frac{\|\mathbf{v}\|}{\lambda_1}) \tag{G.9}$$

Now, letting $\mathbf{v} \sim \mathcal{N}(0, I)$ and taking the expectation of both sides and letting $\epsilon \to 0$, we conclude with the continuity of $\mathbf{P}_f(\mathbf{x}_0, \lambda)$ for $\lambda > 0$.

To show continuity at 0, observe that, for any $\lambda > 0$, we have, $\|\text{Proj}_\lambda(\mathbf{v})\| \leq R\lambda$ where $R = \sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \|\mathbf{s}\|$. Hence,

$$|\mathbf{P}_f(\mathbf{x}_0, \lambda) - \mathbf{P}_f(\mathbf{x}_0, 0)| = \mathbf{P}_f(\mathbf{x}_0, \lambda) \leq R^2\lambda^2 \tag{G.10}$$

As $\lambda \to 0$, $\mathbf{P}_f(\mathbf{x}_0, \lambda) = 0$.
Statement 5: For a proof see Lemma $B.2$ in [31].

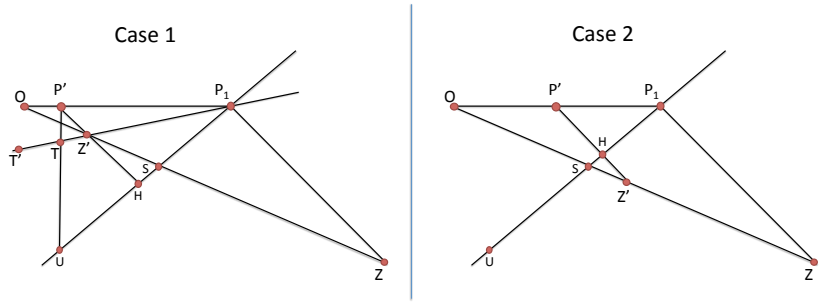


Figure 9: Possible configurations of the points in Lemma G.1 when $Z\hat{P}_1O$ is wide angle.

Statement 6: Based on Lemma G.1, given vector $\mathbf{v}$, set $\mathcal{C}$ and scalar $1 \geq c > 0$, we have,

$$\frac{\|\text{Proj}(c\mathbf{v}, \mathcal{C})\|}{c} \geq \|\text{Proj}(\mathbf{v}, \mathcal{C})\| \tag{G.11}$$

Given $\lambda_1 > \lambda_2 > 0$, this gives,

$$\|\text{Proj}(\mathbf{v}, \lambda_1 \partial f(\mathbf{x}_0))\| = \lambda_1 \|\text{Proj}(\frac{\mathbf{v}}{\lambda_1}, \partial f(\mathbf{x}_0))\| \geq \lambda_1 \frac{\lambda_2}{\lambda_1} \|\text{Proj}(\frac{\mathbf{v}}{\lambda_2}, \partial f(\mathbf{x}_0))\| = \|\text{Proj}(\mathbf{v}, \lambda_2 \partial f(\mathbf{x}_0))\| \tag{G.12}$$

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 $\mathbf{D}_f(\mathbf{x}_0, \lambda) \geq 0$ 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 G.1. *Let $\mathcal{C}$ be a convex and closed set in $\mathbb{R}^n$. Let $\mathbf{z}$ and $0 < \alpha < 1$ be arbitrary, let $\mathbf{p}_1 = Proj(\mathbf{z}, \mathcal{C})$, $\mathbf{p}_2 = Proj(\alpha\mathbf{z}, \mathcal{C})$. Then,*

$$\|\mathbf{p}_1\| \leq \frac{\|\mathbf{p}_2\|}{\alpha} \tag{G.13}$$

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 $\mathcal{C}'$ by projecting the set $\mathcal{C}$ to the $2D$ plane induced by the points Z, P_1 and O. Now, let $\mathbf{p}_2' = \text{Proj}(\alpha\mathbf{z}, \mathcal{C}')$. Due to the projection, we still have: $\|\mathbf{z} - \mathbf{p}_2'\| \leq \|\mathbf{z} - \mathbf{p}_2\|$ and $\|\mathbf{p}_2'\| \leq \|\mathbf{p}_2\|$. We wish to prove that $\|\mathbf{p}_2'\| \geq \|\alpha\mathbf{p}_1\|$. Figures 9 and 10 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 $\mathcal{C}$ 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 9.

Case 1 (S lies on $Z'Z$): Consider the lefthand side of Figure 9. 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 $UTZ'H$ 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:

$$P_2'\hat{Z}'P_1 \geq U\hat{Z}'P_1 > U\hat{T}P_1 > U\hat{P}'P_1 = \frac{\pi}{2} \tag{G.14}$$

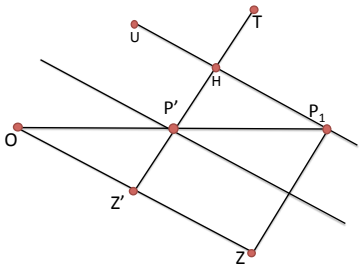


Figure 10: Lemma G.1 when $Z\hat{P}_1O$ is acute or right angle.

Case 2 (S lies on OZ'): Consider the righthand side of Figure 9. 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 10. 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 $\|\mathbf{p}_2\| \geq \|\mathbf{p}_2'\| \geq \alpha\|\mathbf{p}_1\|$ as desired.

□

Statement 7: For a proof see Lemma $B.2$ in [31].

Statement 8: From Statement 7, $\mathbf{C}_f(\mathbf{x}_0,\lambda) = -\frac{\lambda}{2}\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda}$. Also from Statement 5, $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ is strictly convex. Thus, $\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda} \leq 0$ for all $\lambda \in [0,\lambda_{\text{best}}]$ which yields $\mathbf{C}_f(\mathbf{x}_0,\lambda) \geq 0$ for all $\lambda \in [0,\lambda_{\text{best}}]$. Similarly, $\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda} \geq 0$ for all $\lambda \in [\lambda_{\text{best}},\infty)$ which yields $\mathbf{C}_f(\mathbf{x}_0,\lambda) \leq 0$ for all $\lambda \in [\lambda_{\text{best}},\infty)$. Finally, λ_{best} minimizes $\mathbf{D}_f(\mathbf{x}_0,\lambda)$. Hence $\frac{d\mathbf{D}_f(\mathbf{x}_0,\lambda)}{d\lambda}|_{\lambda=\lambda_{\text{best}}} = 0$ which yields $\mathbf{C}_f(\mathbf{x}_0,\lambda_{\text{best}}) = 0$.

Statement 9: We prove that for any $0 \leq \lambda_1 < \lambda_2 \leq \lambda_{\text{best}}$,

$$\mathbf{D}_f(\mathbf{x}_0,\lambda_1) + \mathbf{C}_f(\mathbf{x}_0,\lambda_1) > \mathbf{D}_f(\mathbf{x}_0,\lambda_2) + \mathbf{C}_f(\mathbf{x}_0,\lambda_2). \tag{G.15}$$

From Statement 5, $\mathbf{D}_f(\mathbf{x}_0,\lambda)$ is strictly decreasing for $\lambda \in [0,\lambda_{\text{best}}]$. Thus,

$$\mathbf{D}_f(\mathbf{x}_0,\lambda_1) > \mathbf{D}_f(\mathbf{x}_0,\lambda_2). \tag{G.16}$$

Furthermore, from Statement 6, $\mathbf{P}_f(\mathbf{x}_0,\lambda)$ is an increasing function of λ. Thus,

$$\mathbf{D}_f(\mathbf{x}_0,\lambda_1) + 2\mathbf{C}_f(\mathbf{x}_0,\lambda_1) \geq \mathbf{D}_f(\mathbf{x}_0,\lambda_2) + 2\mathbf{C}_f(\mathbf{x}_0,\lambda_2). \tag{G.17}$$

where we have used Statement 1. Combining (G.16) and (G.17), we conclude with (G.15), as desired.

H. Explicit formulas for well-known functions

H.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 $f(\cdot) = \|\cdot\|_1$,

- $\frac{\mathbf{D}_f(\mathbf{x}_0,\lambda)}{n} = (1+\lambda^2)(1-(1-\beta)\text{erf}(\frac{\lambda}{\sqrt{2}})) - \sqrt{\frac{2}{\pi}}(1-\beta)\lambda\exp(-\frac{\lambda^2}{2})$
- $\frac{\mathbf{P}_f(\mathbf{x}_0,\lambda)}{n} = \beta\lambda^2 + (1-\beta)[\text{erf}(\frac{\lambda}{\sqrt{2}}) + \lambda^2\text{erfc}(\frac{\lambda}{\sqrt{2}}) - \sqrt{\frac{2}{\pi}}\lambda\exp(-\frac{\lambda^2}{2})]$
- $\frac{\mathbf{C}_f(\mathbf{x}_0,\lambda)}{n} = -\lambda^2\beta + (1-\beta)[\sqrt{\frac{2}{\pi}}\lambda\exp(-\frac{\lambda^2}{2}) - \lambda^2\text{erfc}(\frac{\lambda}{\sqrt{2}})]$

These are not difficult to obtain. For example, to find $\mathbf{D}_f(\mathbf{x}_0, \lambda)$, 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,

$$(\Pi(\mathbf{g}, \lambda \partial f(\mathbf{x}_0)))_i = \begin{cases} \mathbf{g}(i) - \lambda \cdot \operatorname{sgn}(\mathbf{x}_0(i)) & \text{if } \mathbf{x}_0(i) \neq 0, \\ \operatorname{shrink}_\lambda(\mathbf{g}(i)) & \text{otherwise.} \end{cases}$$

where $\operatorname{shrink}_\lambda(\mathbf{g}(i))$ is the soft thresholding operator defined as,

$$\operatorname{shrink}_\lambda(x) = \begin{cases} x - \lambda & \text{if} \quad x > \lambda, \\ 0 & \text{if} \quad |x| \leq \lambda, \\ x + \lambda & \text{if} \quad x < -\lambda. \end{cases}$$

Consequently, we obtain our formulas after taking the expectation of $\mathbf{g}(i) - \lambda \cdot \operatorname{sgn}(\mathbf{x}_0(i))$ and $\operatorname{shrink}_\lambda(\mathbf{g}(i))$. For more details on these formulas, the reader is referred to [26, 28, 29, 40] which calculate the phase transitions of ℓ_1 minimization.

H.1.1 Closed form bound

We will now find a closed form bound on $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ for the same sparse signal $\mathbf{x}_0$. In particular, we will show that $\mathbf{D}_f(\mathbf{x}_0, \lambda) \leq (\lambda^2 + 2)k$ 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, $\mathbb{E}[(\mathbf{g}_i - \lambda \cdot \operatorname{sgn}(\mathbf{x}_0(i)))^2] = \lambda^2 + 1$

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) = \sum \mathbb{E}[(\mathbf{g}(i) - \lambda \cdot \operatorname{sgn}(\mathbf{x}_0(i)))^2] + (n-k)\mathbb{E}[\operatorname{shrink}_\lambda(\mathbf{g}(i))^2] \tag{H.1}$$

The sum on the left hand side is simply $(\lambda^2 + 1)k$. The interesting term is $\operatorname{shrink}_\lambda(\mathbf{g}(i))$. To calculate this, we will use the following lemma.

Lemma H.1. *Let x be a nonnegative random variable. Assume, there exists $c > 0$ such that for all $t > 0$,*

$$\mathbb{P}(x \geq c + t) \leq \exp(-\frac{t^2}{2}) \tag{H.2}$$

For any $a \geq 0$, we have,

$$\mathbb{E}[\mathit{shrink}_{a+c}(x)^2] \leq \frac{2}{a^2+1} \exp(-\frac{a^2}{2}). \tag{H.3}$$

Proof. Let $Q(t) = \mathbb{P}(x \geq t)$.

$$\begin{aligned} \mathbb{E}[\operatorname{shrink}_{a+c}(x)^2] &= \int_{a+c}^{\infty} (x - a - c)^2 d(-Q(x)) && \text{(H.4)} \\ &\leq -[Q(x)(x-a-c)^2]_{a+c}^{\infty} + \int_{a+c}^{\infty} Q(x) d(x-a-c)^2 = \int_{a+c}^{\infty} Q(x) d(x-a-c)^2 && \text{(H.5)} \\ &\leq \int_{a+c}^{\infty} 2(x-a-c)Q(x)d(x-a-c) \leq 2\int_{a+c}^{\infty} (x-a-c)\exp(-\frac{(x-c)^2}{2}) d(x-a-c) \\ &\leq 2\int_{a}^{\infty} (u-a)\exp(-\frac{u^2}{2})du \leq 2\exp(-\frac{a^2}{2}) - 2a\frac{a}{a^2+1}\exp(-\frac{a^2}{2}) = \frac{2}{a^2+1}\exp(-\frac{a^2}{2}) && \text{(H.6)} \end{aligned}$$

(H.5) follows from integration by parts and (H.6) 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 $\mathbb{E}[\operatorname{shrink}_\lambda(g)^2]$ 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 H.1 with $c = 0$ and $a = \lambda$ yields, $\mathbb{E}[|\operatorname{shrink}_\lambda(g)|^2] \leq \frac{2}{\lambda^2+1}\exp(-\frac{\lambda^2}{2})$. Combining this with (H.1), we find,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) \leq (\lambda^2+1)k + \frac{2n}{\lambda^2+1}\exp(-\frac{\lambda^2}{2}) \tag{H.7}$$

For $\lambda \geq \sqrt{2\log\frac{n}{k}}$, $\exp(-\frac{\lambda^2}{2}) \leq \frac{k}{n}$. Hence, we obtain,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) \leq (\lambda^2+1)k + \frac{2k}{\lambda^2+1} \leq (\lambda^2+3)k \tag{H.8}$$

H.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,

$$S_{\mathbf{X}_0} = \{\mathbf{M} \big| (\mathbf{I}-\mathbf{U}\mathbf{U}^T)\mathbf{M}(\mathbf{I}-\mathbf{V}\mathbf{V}^T) = 0\} \tag{H.9}$$

The subdifferential of nuclear norm is given as,

$$\partial\|\mathbf{X}_0\|_\star = \{\mathbf{S} \in \mathbb{R}^{d\times d} \big| \text{Proj}(\mathbf{S}, S_{\mathbf{X}_0}) = \mathbf{U}\mathbf{V}^T, \text{ and } \|\text{Proj}(\mathbf{S}, \bar{S}_{\mathbf{X}_0})\|_2 \leq 1\} \tag{H.10}$$

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 [45,55,56], $\Pi(\mathbf{G}, \lambda\partial f(\mathbf{x}_0))$ effectively behaves as singular value soft thresholding. In particular, we have,

$$\Pi(\mathbf{G}, \lambda\partial f(\mathbf{x}_0)) = (\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}) - \lambda\mathbf{U}\mathbf{V}^T) + \sum_{i=1}^{n-r} \text{shrink}_\lambda(\sigma_{\mathbf{G},i})\mathbf{u}_{\mathbf{G},i}\mathbf{v}_{\mathbf{G},i}^T \tag{H.11}$$

where $\text{Proj}(\mathbf{G}, \bar{S}_{\mathbf{X}_0})$ 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 quarter circle law when singular values are normalized by $\sqrt{d}$.

$$\psi(x) = \begin{cases} \frac{1}{\pi}\sqrt{4-x^2} \text{ if } 0 \leq x \leq 2 \\ 0 \text{ else} \end{cases} \tag{H.12}$$

Based on ψ, define the quantities related to the moments of tail of ψ. Namely,

$$\Psi_i(x) = \int_x^\infty x^i\psi(x)dx \tag{H.13}$$

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,

$$v = \frac{\lambda}{2\sqrt{1-\beta}} \tag{H.14}$$

- $\frac{\mathbf{D}_f(\mathbf{x}_0, \lambda\sqrt{d})}{n} = [2\beta - \beta^2 + \beta\lambda^2] + [(1-\beta)\lambda^2\Psi_0(v) + (1-\beta)^2\Psi_2(v) - 2(1-\beta)^{3/2}\lambda\Psi_1(v)]$
- $\frac{\mathbf{P}_f(\mathbf{x}_0, \lambda\sqrt{d})}{n} = \beta\lambda^2 + (1-\beta)\lambda^2\Psi_0(v) + (1-\beta)^2(1-\Psi_2(v))$
- $\frac{\mathbf{C}_f(\mathbf{x}_0, \lambda\sqrt{d})}{n} = -\lambda^2\beta - (1-\beta)\lambda^2\Psi_0(v) + (1-\beta)^{3/2}\lambda\Psi_1(v)$

H.2.1 Closed form bounds

Our approach will exactly follow the proof of Proposition 3.11 in [25]. Given $\mathbf{G}$ with i.i.d. standard normal entries, the spectral norm of the off-support term $\text{Proj}(\mathbf{G}, \bar{S}_{\mathbf{X}_0})$ satisfies,

$$\mathbb{P}(\|\text{Proj}(\mathbf{G}, \bar{S}_{\mathbf{X}_0})\|_2 \geq 2\sqrt{d-r} + t) \leq \exp(-\frac{t^2}{2}) \tag{H.15}$$

It follows that all singular values of $\text{Proj}(\mathbf{G}, \bar{S}_{\mathbf{X}_0})$ satisfies the same inequality as well. Consequently, for any singular value and for $\lambda \geq 2\sqrt{d-r}$, applying Lemma H.1, we may write,

$$\mathbb{E}[\text{shrink}_\lambda(\sigma_{\mathbf{G},i})^2] \leq \frac{2}{(\lambda - 2\sqrt{d-r})^2 + 1} \exp(-\frac{(\lambda - 2\sqrt{d-r})^2}{2}) \leq 2 \tag{H.16}$$

It follows that,

$$\sum_{i=1}^{d-r} \mathbb{E}[\text{shrink}_\lambda(\sigma_{\mathbf{G},i})^2] \leq 2(d-r) \tag{H.17}$$

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,

$$\|\text{Proj}(\mathbf{G}, S_{\mathbf{X}_0}) - \lambda \mathbf{U}\mathbf{V}^T\|_F^2 = \lambda^2 r + |S_{\mathbf{X}_0}| = \lambda^2 r + 2dr - r^2 \tag{H.18}$$

Combining, we find,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) \leq \lambda^2 r + 2dr - r^2 + 2d - 2r \leq (\lambda^2 + 2d)r + 2d \tag{H.19}$$

H.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, [46, 48, 49]. In particular, denoting the subvector corresponding to i'th block of $\mathbf{x}$ by $\mathbf{x}_i$

$$\|\mathbf{x}\|_{1,2} = \sum_{i=1}^{t} \|\mathbf{x}_i\| \tag{H.20}$$

To calculate $\mathbf{D}_f(\mathbf{x}_0, \lambda), \mathbf{C}_f(\mathbf{x}_0, \lambda), \mathbf{P}_f(\mathbf{x}_0, \lambda)$ 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 [48, 49] and is given as,

$$\Pi(\mathbf{g}, \lambda\partial\|\mathbf{x}_0\|_{1,2}) = \begin{cases} \mathbf{g}_i - \lambda \frac{\mathbf{x}_{0,i}}{\|\mathbf{x}_{0,i}\|} & \text{if} \quad \mathbf{x}_{0,i} \neq 0 \\ \text{vshrink}_\lambda(\mathbf{g}_i) & \text{else} \end{cases} \tag{H.21}$$

where the vector shrinkage vshrink_λ is defined as,

$$\text{vshrink}_\lambda(\mathbf{v}) = \begin{cases} \mathbf{v}(1 - \frac{\lambda}{\|\mathbf{v}\|}) & \text{if} \quad \|\mathbf{v}\| > \lambda \\ 0 \quad \text{if} \quad \|\mathbf{v}\| \leq \lambda \end{cases} \tag{H.22}$$

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 [49]). Similar to previous section, define the tail function of a χ^2-distribution with b degrees of freedom as,

$$\Psi_i(x) = \int_x^\infty x^i \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} \exp(-\frac{x}{2}) dx \tag{H.23}$$

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 $\mathbf{D}_f(\mathbf{x}_0, \lambda), \mathbf{P}_f(\mathbf{x}_0, \lambda)$ and $\mathbf{C}_f(\mathbf{x}_0, \lambda)$ as follows.

- $\mathbf{D}_f(\mathbf{x}_0, \lambda) = k(b + \lambda^2) + [\Psi_1(\lambda^2) + \Psi_0(\lambda^2)\lambda^2 - 2\Psi_{\frac{1}{2}}(\lambda^2)\lambda](t-k)$
- $\mathbf{P}_f(\mathbf{x}_0, \lambda) = \lambda^2 k + [(\Psi_1(0) - \Psi_1(\lambda^2)) + \lambda^2\Psi_0(\lambda^2)](t-k)$
- $\mathbf{C}_f(\mathbf{x}_0, \lambda) = -\lambda^2 k + [\lambda\Psi_{\frac{1}{2}}(\lambda^2) - \lambda^2\Psi_0(\lambda^2)](t-k)$

H.3.1 Closed form bound

Similar to Proposition 3 of [32], we will make use of the following bound for a x distributed with χ^2-distribution with b degrees of freedom.

$$\mathbb{P}(\sqrt{x} \geq \sqrt{b} + t) \leq \exp(-\frac{t^2}{2}) \quad \text{for all } t > 0 \tag{H.24}$$

Now, the total contribution of nonzero blocks to $\mathbf{D}_f(\mathbf{x}_0, \lambda)$ 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 H.1, with $c = \sqrt{b}$ and $a = \lambda - \sqrt{b}$ and using the tail bound (H.24), we obtain,

$$\mathbb{E}[\|\text{vshrink}_\lambda(\mathbf{g}_i)\|^2] \leq \frac{2}{(\lambda - \sqrt{b})^2 + 1} \exp(-\frac{(\lambda - \sqrt{b})^2}{2}) \tag{H.25}$$

Combining everything,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) \leq k(\lambda^2 + b) + \frac{2t}{(\lambda - \sqrt{b})^2 + 1} \exp(-\frac{(\lambda - \sqrt{b})^2}{2}) \tag{H.26}$$

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,

$$\mathbf{D}_f(\mathbf{x}_0, \lambda) \leq k(\lambda^2 + b) + \frac{2k}{(\lambda - \sqrt{b})^2 + 1} \leq k(\lambda^2 + b + 2) \tag{H.27}$$

I. 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 $\tau > 0$. To state the results, we will first define the Gaussian width which has been the topic of closely related papers [25, 31, 41, 72].

Definition I.1. *Let $S \subseteq \mathbb{R}^n$. Let $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$. Then, the Gaussian width of S is given as,*

$$\omega(S) = \mathbb{E}[\sup_{\mathbf{v} \in S} \langle \mathbf{v}, \mathbf{g} \rangle] \tag{I.1}$$

Let us also state a standard result on the Gaussian width and cones that can be found in [31, 32].

Proposition I.1.

The following lemma provides a Gaussian width characterization of "widening of a tangent cone".

Lemma I.1. *Assume $f(\cdot)$ is a convex function and $\mathbf{x}_0$ is not a minimizer of $f(\cdot)$. Given $\epsilon_0 > 0$, consider the ϵ_0-widened tangent cone defined as,*

$$\mathcal{T}_f(\mathbf{x}_0, \epsilon_0) = Cl(\{\alpha \cdot \mathbf{w} | f(\mathbf{x}_0 + \mathbf{w}) \leq f(\mathbf{x}_0) + \epsilon_0 \|\mathbf{w}\|,\ \alpha \geq 0\}) \tag{I.2}$$

Let $R_{min} = \min_{\mathbf{s} \in \partial f(\mathbf{x}_0)} \|\mathbf{s}\|$ and $\mathcal{B}^{n-1}$ be the unit ℓ_2-ball in $\mathbb{R}^n$. Then,

$$\omega(\mathcal{T}_f(\mathbf{x}_0, \epsilon_0) \cap \mathcal{B}^{n-1}) \leq \omega(\mathcal{T}_f(\mathbf{x}_0) \cap \mathcal{B}^{n-1}) + \frac{\epsilon_0 \sqrt{n}}{R_{min}} \tag{I.3}$$

Proof. Let $\mathbf{w} \in \mathcal{T}_f(\mathbf{x}_0, \epsilon_0)$. Write $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$ via Moreau's decomposition theorem (Fact A.1) where $\mathbf{w}_1 \in \mathcal{T}_f(\mathbf{x}_0)$ 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 $\mathcal{T}_f(\mathbf{x}_0)^* = \text{cone}(\partial f(\mathbf{x}_0))$. To find a bound on $\mathcal{T}_f(\mathbf{x}_0, \epsilon_0)$ in terms of $\mathcal{T}_f(\mathbf{x}_0)$, 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 $1 > \epsilon > 0$, $\epsilon\epsilon_0\|\mathbf{w}\| \geq f(\epsilon\mathbf{w} + \mathbf{x}_0) - f(\mathbf{x}_0)$. Now, using Proposition 9.2 with $\delta \to 0$, we obtain,

$$\begin{aligned}\epsilon_0\|\mathbf{w}\| \geq \lim_{\epsilon\to 0}\frac{f(\epsilon\mathbf{w}+\mathbf{x}_0)-f(\mathbf{x}_0)}{\epsilon} &= \sup_{\mathbf{s}\in\partial f(\mathbf{x}_0)} \mathbf{w}^T\mathbf{s} \\ &\geq \mathbf{w}^T\mathbf{s}(\mathbf{w}_2) = \mathbf{w}_1^T\mathbf{s}(\mathbf{w}_2) + \mathbf{w}_2^T\mathbf{s}(\mathbf{w}_2) \\ &= \|\mathbf{w}_2\|\|\mathbf{s}(\mathbf{w}_2)\| \geq \|\mathbf{w}_2\|R_{min} \end{aligned} \tag{I.4}$$

This gives, $\frac{\|\mathbf{w}_2\|}{\|\mathbf{w}\|} \leq \frac{\epsilon_0}{R_{min}}$. Equivalently, for a unit size $\mathbf{w}$, $\|\mathbf{w}_2\| \leq \frac{\epsilon_0}{R_{min}}$.

What remains is to estimate the Gaussian width of $\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}$. Let $\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)$. $\mathbf{w}_1,\mathbf{w}_2$ still denote the projection of $\mathbf{w}$ onto $\mathcal{T}_f(\mathbf{x}_0)$ and cone$(\partial f(\mathbf{x}_0))$ respectively.

$$\omega(\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}) = \mathbb{E}[\sup_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \mathbf{w}^T\mathbf{g}] \tag{I.5}$$

$$\leq \mathbb{E}[\sup_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \mathbf{w}_1^T\mathbf{g}] + \mathbb{E}[\sup_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \mathbf{w}_2^T\mathbf{g}] \tag{I.6}$$

Observe that, for $\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}$, $\|\mathbf{w}_2\| \leq \frac{\epsilon_0}{R_{min}}$,

$$\mathbb{E}[\sup_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \mathbf{w}_2^T\mathbf{g}] \leq \mathbb{E}[\sup_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \|\mathbf{w}_2\|\|\mathbf{g}\|] \leq \frac{\epsilon_0}{R_{min}}\mathbb{E}\|\mathbf{g}\| \leq \frac{\epsilon_0\sqrt{n}}{R_{min}} \tag{I.7}$$

For $\mathbf{w}_1$, we have $\mathbf{w}_1\in\mathcal{T}_f(\mathbf{x}_0)$ and $\|\mathbf{w}_1\| \leq \|\mathbf{w}\| \leq 1$ which gives,

$$\mathbb{E}[\sup_{\mathbf{w}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \mathbf{w}_1^T\mathbf{g}] \leq \mathbb{E}[\sup_{\mathbf{w}'\in\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{B}^{n-1}} \mathbf{w}'^T\mathbf{g}] = \omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{B}^{n-1}) \tag{I.8}$$

Combining these individual bounds, we find,

$$\omega(\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}) \leq \omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{B}^{n-1}) + \frac{\epsilon_0\sqrt{n}}{R_{min}} \tag{I.9}$$

□

Lemma I.2. *Let $\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)$ denote the widened cone defined in (I.2) and consider the exact same setup in Lemma I.1. Fix $\epsilon_1 > 0$. Let $\mathbf{A}\in\mathbb{R}^{m\times n}$ have i.i.d. standard normal entries. Then, whenever,*

$$\gamma(m,f,\epsilon_0,\epsilon_1) := \sqrt{m-1} - \sqrt{\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)} - \frac{\epsilon_0\sqrt{n}}{R_{min}} - \epsilon_1 > 0 \tag{I.10}$$

we have,

$$\mathbb{P}(\min_{\mathbf{v}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \|\mathbf{A}\mathbf{v}\| \geq \epsilon_1) \geq 1 - 2\exp(-\frac{1}{2}\gamma(m,f,\epsilon_0,\epsilon_1)^2) \tag{I.11}$$

Proof. Our proof will follow the same lines as the proof of Corollary 3.3 of Chandrasekaran et al. [25]. For this proof, we will make use of the following lemma of Gordon [72] (Corollary 1.2).

Proposition I.2. *Let $\mathcal{C}\in\mathbb{R}^n$ be a closed and convex subset of $\mathcal{B}^{n-1}$. Then,*

$$\mathbb{E}[\min_{\mathbf{v}\in\mathcal{C}} \|\mathbf{A}\mathbf{v}\|] \geq \sqrt{m-1} - \omega(\mathcal{C}) \tag{I.12}$$

Pick $\mathcal{C} = \mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}$ in the above proposition. Combined with Lemma I.1, this gives,

$$\mathbb{E}[\min_{\mathbf{v}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}} \|\mathbf{A}\mathbf{v}\|] \geq \sqrt{m-1} - \omega(\mathcal{T}_f(\mathbf{x}_0)) - \frac{\epsilon_0\sqrt{n}}{R_{min}} \tag{I.13}$$

Following [25], the function $\min_{\mathbf{v}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}}\|\mathbf{A}\mathbf{v}\|$ is 1-Lipschitz function of $\mathbf{A}$ in Frobenius norm. Using Lemma A.4, for ϵ_1 smaller than the right hand side of (I.13), we find,

$$\mathbb{P}(\min_{\mathbf{v}\in\mathcal{T}_f(\mathbf{x}_0,\epsilon_0)\cap\mathcal{B}^{n-1}}\|\mathbf{A}\mathbf{v}\|\geq\epsilon_1)\geq 1-2\exp(-\frac{1}{2}(\sqrt{m-1}-\omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{B}^{n-1})-\frac{\epsilon_0\sqrt{n}}{R_{min}}-\epsilon_1)^2) \tag{I.14}$$

To conclude, we will use $\omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{B}^{n-1})\leq\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+)$. To see this, applying Moreau's decomposition theorem (Fact A.1), observe that for a closed and convex cone $\mathcal{K}$ and an arbitrary vector $\mathbf{g}$,

$$\|\text{Proj}(\mathbf{g},\mathcal{K})\|=\sup_{\mathbf{v}\in\mathcal{K}\cap\mathcal{B}^{n-1}}\mathbf{v}^T\mathbf{g} \tag{I.15}$$

Picking $\mathcal{K}=\mathcal{T}_f(\mathbf{x}_0)$ and $\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)$,

$$\omega(\mathcal{T}_f(\mathbf{x}_0)\cap\mathcal{B}^{n-1})=\mathbb{E}[\|\text{Proj}(\mathbf{g},\mathcal{T}_f(\mathbf{x}_0))\|]\leq\sqrt{\mathbb{E}[\|\text{Proj}(\mathbf{g},\mathcal{T}_f(\mathbf{x}_0))\|^2]}=\mathbf{D}_f(\mathbf{x}_0,\mathbb{R}^+) \tag{I.16}$$

□