Recovering Structured Signals in Noise: Least-Squares Meets Compressed Sensing

Christos Thrampoulidis, Samet Oymak and Babak Hassibi

July 2014

Abstract The typical scenario that arises in most “big data” problems is one where the ambient dimension of the signal is very large (e.g. high resolution images, gene expression data from a DNA microarray, social network data, etc.), yet is such that its desired properties lie in some low dimensional structure (sparsity, low-rankness, clusters, etc.). In the modern viewpoint, the goal is to come up with efficient algorithms to reveal these structures and for which, under suitable conditions, one can give theoretical guarantees. We specifically consider the problem of recovering such a structured signal (sparse, low-rank, block-sparse, etc.) from noisy compressed measurements. A general algorithm for such problems, commonly referred to as generalized LASSO, attempts to solve this problem by minimizing a least-squares cost with an added “structure-inducing” regularization term ($\ell_1$ norm, nuclear norm, mixed $\ell_2/\ell_1$ norm, etc.). While the LASSO algorithm has been around for 20 years and has enjoyed great success in practice, there has been relatively little analysis of its performance. In this chapter, we will provide a full performance analysis and compute, in closed form, the mean-square-error of the reconstructed signal. We will highlight some of the mathematical vignettes necessary for the analysis, make connections to noiseless compressed sensing and proximal denoising, and will emphasize the central role of the “statistical dimension” of a structured signal.
1 Introduction

Consider the standard linear regression setup, in which we are interested in recovering an unknown signal $x_0 \in \mathbb{R}^n$ from a vector $y = Ax_0 + z \in \mathbb{R}^m$ of noisy linear observations.

1.1 Ordinary Least-Squares

The most prominent method for estimating the unknown vector $x_0$ is that of ordinary least-squares (OLS). The OLS estimate $\hat{x}$ of $x_0$ is obtained by minimizing the residual squared error

$$\hat{x} = \arg\min_x \|y - Ax\|_2^2.$$  \hfill (1)

OLS has a long history originating in the early 1800’s due to works by Gauss and Legendre [41, 53], and its behavior is by now very well understood. In particular, in the classical setting $m > n$, assuming $A$ is full column-rank, (1) has a unique solution which is famously given by

$$\hat{x} = (A^T A)^{-1} A^T y.$$  \hfill (2)

The squared error-loss of the OLS estimate in (2) is, thus, expressed as

$$\|\hat{x} - x_0\|_2^2 = z^T A (A^T A)^{-2} A^T z.$$  \hfill (3)

Starting from (3) and imposing certain generic assumptions on the measurement matrix $A$ and/or the noise vector $z$, it is possible to conclude precise and simple formulae characterizing the estimation error $\|\hat{x} - x_0\|_2^2$. As an example, when the entries of $z$ are drawn i.i.d. normal of zero-mean and variance $\sigma^2$, then $\mathbb{E}\|\hat{x} - x_0\|_2^2 = \sigma^2 \text{trace}((A^T A)^{-1})$. Furthermore, when the entries of $A$ are drawn i.i.d. normal of zero-mean and variance $1/m$, $A^T A$ is a Wishart matrix whose asymptotic eigendistribution is well known. Using this, and letting $m,n$ grow, we find that the squared error concentrates around

$$\frac{\|\hat{x} - x_0\|_2^2}{\|z\|_2^2} \approx \frac{n}{m-n}.$$  \hfill (4)

Such expressions serve as valuable insights regarding the behavior of the OLS estimator and are meant to provide guidelines for the effectiveness of the estimator in practical situations.
1.2 Structured Signals and Noisy Compressed Sensing

Gauss and Legendre proposed the OLS method in the context of traditional statistical data analysis, in which, one assumes a large number \( m \) of observations and only a small number \( n \) of well-chosen variables corresponding to the entries of the desired signal \( x_0 \). That trend is rapidly changing in the today’s era of “big-data” and of the collection of massively large data of all kinds. In a number of disciplines and application domains including machine learning, signal processing, social networks and DNA microarrays, we are increasingly confronted with very large data sets where we need to extract some signal-of-interest. In this setting, it is important to have signal recovery algorithms that are computationally efficient and that need not access the entire data directly, i.e. can reliably estimate \( x_0 \) from fewer number of observations \( m \) than the dimension \( n \) of the model to be estimated.

This problem of compressed recovery, is in general ill-posed. Even in the absence of noise, recovering a general signal \( x_0 \) of dimension \( n \) from only \( m < n \) linear observations \( y = Ax_0 \) is futile, since there is typically infinitely many solutions satisfying the measurement equations. Thus, on the face of it the limited availability of data relative to the ambient dimension of the signal to be estimated can lead to the curse of dimensionality [18]. Fortunately, in many applications, the signal of interest lives in a manifold of much lower dimension than that of the original ambient space.

We refer to such signals that are structurally constrained to only have very few degrees of freedom relative to the ambient dimension, as structured signals. The most celebrated example of structure in the compressed sensing literature is sparsity. To appreciate how knowledge of the structure of the unknown signal \( x_0 \) can help alleviate the ill-posed nature of the problem, consider a desired signal \( x_0 \) which is \( k \)-sparse i.e., has only \( k < n \) (often \( k \ll n \)) non-zero entries. Suppose we make \( m \) noisy measurements of \( x_0 \) using the \( m \times n \) measurement matrix \( A \) to obtain \( y = Ax_0 + z \) and further suppose each set of \( m \) columns of \( A \) be linearly independent. Then, as long as \( m > k \), we can always find the sparsest solution to

\[
\hat{x} = \arg\min_x \|y - Ax\|_2^2,
\]

via exhaustive search of \( \binom{n}{k} \) such least-squares problems. Under the same assumptions that led to (4), this gives a normalized squared error

\[
\frac{\|\hat{x} - x_0\|_2^2}{\|z\|_2^2} \approx \frac{k}{m-k}.
\]  

(5)

The catch here, of course, is that the computational complexity of the estimation procedure that was just described is exponential in the ambient dimension \( n \), thus, making it intractable. So, the fundamental question to ask is whether we can estimate the structured (here, sparse) signal \( x_0 \) in a computationally efficient way? And if so, how many measurements \( m \) are needed? How does the squared error behave and how does it compare to (4) and (5)?
1.3 LASSO

LASSO is a method for reconstructing the unknown signal $x_0$ and is particularly useful in the undetermined setting $m < n$ and when one seeks sparse signals. It was originally proposed by Tibshirani in 1996 [61], and consists of solving the following constrained optimization problem

$$\hat{x} = \arg\min_x \|y - Ax\|_2^2 \quad \text{s.t.} \quad \|x\|_1 \leq \|x_0\|_1.$$  

(6)

The program in (6) solves for an estimate $\hat{x}$ that best fits the vector of observations $y$, in the OLS sense, while at the same time is constrained to retain structure similar to that of $x_0$. At this point, recall that the $\ell_1$-norm is typically used in the compressed sensing literature to promote sparse solutions.

Solving (6) requires a priori knowledge of $\|x_0\|_1$. When such knowledge is not available, one can instead solve regularized versions of it, like,

$$\min_x \|y - Ax\|_2^2 + \lambda \|x\|_1,$$  

(7)

or

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \tau \|x\|_1,$$  

(8)

for nonnegative regularizer parameters $\lambda$ and $\tau$. Although very similar in nature, (7) and (8) show in general different statistical behaviors [4, 47]. Lagrange duality ensures that there exist $\lambda$ and $\tau$ such that they both become equivalent to the constrained optimization (6). However, in practice, the challenge lies in tuning the regularizer parameters to achieve good estimation, with as little possible prior knowledge\(^1\). Observe that letting $\lambda = \tau \to 0$, reduces (7) and (8) to the standard $\ell_1$-minimization [12, 19]

$$\min_x \|x\|_1 \quad \text{s.t.} \quad y = Ax.$$  

The problem in (8) is arguably the most encountered in the literature and amongst (6) and (7), it is the one which is widely termed the LASSO estimator.

1.4 Generalized LASSO

The Generalized LASSO algorithm is a generalization of the standard LASSO introduced by Tibshirani, which can be used to enforce other types of structures apart from sparsity. In order to maintain a generic treatment of “structure”, we associate

\(^1\) Assuming that the entries of the noise vector $z$ are i.i.d., it is well-known that a sensible choice of $\tau$ in (8) must scale with the standard deviation $\sigma$ of the noise components [6, 11, 42]. On the other hand, (7) eliminates the need to know or to pre-estimate $\sigma$ [4].
with it a convex function $f(\cdot)$. We will commonly refer to $f(\cdot)$ as the structure-inducing or structure-promoting function. In the case of sparse signals, $f$ is chosen to be the $\ell_1$-norm. When the unknown signal is a low-rank matrix, $f$ is chosen as the nuclear norm (sum of the singular values) and for a block-sparse signal the associated structure-inducing function is the $\ell_{1,2}$-norm. Recently, Chandrasekaran et. al. [14] have proposed a principled way to convert notions of simplicity into such corresponding convex structure-promoting functions.

When a structure-inducing function $f$ is available, we can try estimate $x_0$ by solving generalized versions of (6), (7) and (8) where the $\ell_1$-norm is substituted by $f$. For instance, the generalized version of (8) solves for

$$\min_x \frac{1}{2} \|y - Ax\|_2^2 + \tau f(x).$$

**1.5 The Squared-Error of the Generalized LASSO**

The Generalized LASSO algorithm solves a convex optimization problem and is, thus, a computationally tractable way for estimating the desired signal $x_0$. When viewed as a natural extension of the OLS to the high-dimensional setting $m < n$, it is natural to ask whether we can give performance bounds for the squared error of the Generalized LASSO. While the LASSO method has been around for 20 years and has enjoyed great success in practice, there has been relatively little such analysis of its performance. In particular, most performance bounds derived in the literature are rather loose ([59] and references therein). Here, we are interested in bounds that are sharp and simple, similar to those that characterize the OLS. In particular, under same assumptions on the distribution of the measurement matrix and the noise vector, we ask whether it is possible to derive bounds that resemble (4) and (5)? It turns out, that we can and this chapter is dedicated to providing a full performance analysis and computation of such bounds.

**1.6 Organization**

In Section 2 we review OLS. Section 3 formally introduces the generalized LASSO, and sets the goals of the chapter, while Section 4, reviews the relevant technical background. Sections 5, 6 and 7 are devoted to the analysis of the squared error of the generalized LASSO. We conclude in Section 8 with directions for future work.
2 Least Squares

We start by briefly reviewing the OLS equations and derive performance bounds under the generic assumption that the entries of $A$ are i.i.d. zero-mean normal with variance $1/m$. In Section 2.1 we examine the case where the entries of the noise vector $z$ are also i.i.d. zero-mean normal with variance $\sigma^2$. In Section 2.2 we compute the squared error of OLS for any fixed noise vector $z$, while in Section 2.3 we perform an analysis of the worst-case error of OLS.

Recall that the OLS solves (1). It is clear that when $m < n$, (1) is ill posed. However, when $m > n$ and $A$ has i.i.d. normal entries, $A$ is full column rank with high probability. The solution of (1) is then unique and given by $\hat{x} = (A^T A)^{-1}A^T y$. Recalling that $y = Ax_0 + z$, this gives a squared error-loss as in (3).

2.1 Gaussian Noise

For the purposes of this section, further assume that the entries of $z$ are i.i.d. zero-mean normal with variance $\sigma^2$ and independent of the entries of $A$. In this case, the normalized mean-squared-error takes the form,

$$E\|\hat{x} - x_0\|^2_2 = E[z^T A(A^T A)^{-2} A^T z]$$

$$= \sigma^2 E[tr(A(A^T A)^{-2} A^T)]$$

$$= \sigma^2 E[tr((A^T A)^{-1})].$$

$A^T A$ is a Wishart matrix and the distribution of its inverse is well studied. In particular, when $m > n + 1$, we have $E[(A^T A)^{-1}] = \frac{m}{m-n-1} I_n$ [35]. Hence,

$$E\|\hat{x} - x_0\|^2_2 = m\sigma^2 \frac{n}{(m-1) - n}.$$

Noting that $E\|z\|^2_2 = m\sigma^2$ and letting $m,n$ large enough we conclude with the stronger concentration result on the squared-error of OLS:

$$\frac{\|\hat{x} - x_0\|^2_2}{\|z\|^2_2} \approx \frac{n}{(m-1) - n}. \quad (9)$$

2.2 Fixed Noise

Fix any noise vector $z$, with the only assumption being that it is chosen independently of the measurement matrix $A$. Denote the projection of $z$ onto the range space of $A$ by $\text{Proj}(z, \text{Range}(A)) := A(A^T A)^{-1} A^T z$ and the minimum singular value of $A$ by $\sigma_{\min}(A)$. Then,
\[ \| \hat{x} - x_0 \|_2 \leq \frac{\| A(\hat{x} - x_0) \|_2}{\sigma_{\text{min}}(A)} = \frac{\| \text{Proj}(z, \text{Range}(A)) \|_2}{\sigma_{\text{min}}(A)}. \] (10)

It is well known that, when \( A \) has entries i.i.d. zero-mean normal with variance \( 1/m \), then \( \sigma_{\text{min}}(A) \approx 1 - \sqrt{\frac{m}{n}} \). [63]. Also, since \( z \) is independent of \( A \), and the range space of \( A \) is uniformly random subspace of dimension \( n \) in \( \mathbb{R}^m \), it can be shown that \( \| \text{Proj}(z, \text{Range}(A)) \|_2^2 \approx \frac{n}{m} \| z \|_2^2 \) (e.g. [10, p. 13]). With these, we conclude that with high probability on the draw of \( A \),

\[ \frac{\| \hat{x} - x_0 \|_2^2}{\| z \|_2^2} \lesssim \frac{n}{(\sqrt{m} - \sqrt{n})^2}. \] (11)

### 2.3 Worst-Case Squared-Error

Next, assume no restriction at all on the noise vector \( z \). In particular, this includes the case of adversarial noise, i.e., noise that has information of the sensing matrix \( A \) and can adapt itself accordingly. As expected, this can cause the reconstruction error to be, in general, significantly worse than the guarantees in (9) and (11). In more detail, we can write,

\[ \| \hat{x} - x_0 \|_2 \leq \frac{\| A(\hat{x} - x_0) \|_2}{\sigma_{\text{min}}(A)} = \frac{\| A(A^TA)^{-1}A^Tz \|_2}{\sigma_{\text{min}}(A)} \leq \| z \|_2 \frac{\| A(A^TA)^{-1}A^T \|_2}{\sigma_{\text{min}}(A)} \leq \| z \|_2 \sigma_{\text{min}}^{-1}(A), \]

where \( \| M \|_2 \) denotes the spectral norm of a matrix \( M \) and we used the fact that the spectral norm of a symmetric projection matrix is upper bounded by 1. It is not hard to show that equality in (12) is achieved when \( z \) is equal to the left singular value of \( A \) corresponding to its minimum singular value. Using the fact that \( \sigma_{\text{min}}(A) \approx 1 - \sqrt{\frac{m}{n}} \), we conclude that,

\[ \frac{\| \hat{x} - x_0 \|_2^2}{\| z \|_2^2} \lesssim \frac{m}{(\sqrt{n} - \sqrt{m})^2}. \] (13)

### 3 The Generalized LASSO

In Section 3.1, we formally introduce the generalized LASSO algorithm. We do so by following a “compressed sensing” perspective. In particular, we view the LASSO as an algorithm performing noisy compressed sensing. This motivates the setup of our analysis, such as the assumption on the entries of the sensing matrix \( A \) being i.i.d. normal, which is typical for yielding precise analytical results in the field of
compressed sensing (e.g. [1, 12, 14, 19, 55]). Also, it helps appreciate the fact that our analysis of the error of the LASSO involves and builds upon notions of convex geometry that are also inherent in the classical analysis of noiseless compressed sensing.

Next, in Section 3.2 we restrict attention to the case where the unknown signal is either sparse or low-rank. For those, we present the formulae that characterize the LASSO error under different assumptions on the noise vector. We also contrast them to the ones corresponding to the OLS as they were discussed in Section 2. Generalizations of the formulae to arbitrary convex structure-inducing functions and the formal statement of the results follows in Sections 5-7 after introducing the necessary mathematical background in Section 4.

### 3.1 Noisy Compressed Sensing

The LASSO problem can be viewed as a “merger” of two closely related problems, the problems of noiseless compressed sensing (CS) and that of proximal denoising.

#### 3.1.1 Noiseless compressed sensing

In the noiseless CS problem one wishes to recover $x_0 \in \mathbb{R}^n$ from the random linear measurements $y = Ax_0 \in \mathbb{R}^m$. A common approach is solving the following convex optimization problem

$$
\min_x f(x) \quad \text{subject to} \quad y = Ax.
$$

A critical performance criteria for the problem (14) concerns the minimum number of measurements needed to guarantee successful recovery of $x_0$ [1, 14, 20, 25, 27, 55, 57]. Here, success means that $x_0$ is the unique minimizer of (14), with high probability, over the realizations of the random matrix $A$.

#### 3.1.2 Proximal denoising

The proximal denoising problem tries to estimate $x_0$ from noisy but uncompressed observations $y = x_0 + z$, where the entries of $z$ are i.i.d. zero-mean Gaussian with variance $\sigma^2$. In particular, it solves,

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

A closely related approach to estimate $x_0$, which requires prior knowledge $f(x_0)$ about the signal of interest $x_0$, is solving the constrained denoising problem:
Recovering Structured Signals in Noise: Precise Error Bounds

\[
\min_x \|y - x\|^2 \quad \text{subject to} \quad f(x) \leq f(x_0).
\]

The natural question to be posed in both cases is how well can one estimate \(x_0\) via (15) (or (16)) \([13, 15, 17, 43]\)? The minimizer \(\hat{x}\) of (15) (or (16)) is a function of the noise vector \(z\) and the common measure of performance, is the normalized mean-squared-error which is defined as \(\frac{E[||\hat{x} - x_0||^2]}{\sigma^2}\).

3.1.3 The “merger” LASSO

The Generalized LASSO problem is naturally merging the problems of noiseless CS and proximal denoising. One assumes compressed and noisy measurements \(y = Ax_0 + z \in \mathbb{R}^m\) of an unknown signal \(x_0 \in \mathbb{R}^n\). The following three algorithms are all variants of the LASSO that try to estimate \(x_0\).

\* **C-LASSO**^2:

[\(\hat{x}_c(A, z) = \arg\min_x \|y - Ax\|_2 \quad \text{subject to} \quad f(x) \leq f(x_0)\]. (17)

\* **\(\ell_2\)-LASSO**^3:

[\(\hat{x}_{\ell_2}(\lambda, A, z) = \arg\min_x \left\{ \|y - Ax\|_2 + \frac{\lambda}{\sqrt{m}} f(x) \right\} \). (18)

\* **\(\ell_2^2\)-LASSO**:

[\(\hat{x}_{\ell_2^2}(\tau, A, z) = \arg\min_x \left\{ \frac{1}{2} \|y - Ax\|_2^2 + \frac{\sigma \tau}{\sqrt{m}} f(x) \right\} \). (19)

The compressed nature of observations, poses the question of finding the minimum number of measurements required to recover \(x_0\) 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, a common measure of performance for the LASSO estimate \(\hat{x}\) is defined to be the **normalized squared error** (NSE):

\[
\text{NSE} = \frac{||\hat{x} - x_0||^2}{||z||^2}.
\]

Certainly, the NSE is not the sole measure of performance of the estimator and other measures can be of special interest for different kinds of structures of the unknown vector and for specific applications. As an example, if \(x_0\) is sparse then

---

^2 C-LASSO in (17) stands for “Constrained LASSO”. The algorithm assumes a priori knowledge of \(f(x_0)\).

^3 In the statistics literature the variant of the LASSO algorithm in (7) is mostly known as the “square-root LASSO” [4]. For the purposes of our presentation, we stick to the more compact term “\(\ell_2\)-LASSO” [47].
the “support recovery criteria”, which measures how well the LASSO recovers the subset of nonzero indices of $x_0$ might be of interest [21, 64, 66].

The results presented in this chapter are on the NSE of the generalized LASSO. When restricted to sparse signals, this has been the subject of analysis of lots of research papers (e.g. see [6, 9, 42] and references therein). Yet, those bounds are in general order-wise and do not capture precisely the behavior of the NSE. [3] and [56] manage to accurately characterize the NSE when $z$ is i.i.d. normal, $x_0$ is sparse and $f$ is the $\ell_1$-norm. In [3, 23], Bayati and Montanari are able to prove that the mean-squared-error of the LASSO problem is equivalent to the one achieved by a properly defined “Approximate Message Passing” (AMP) algorithm [39]. 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 [38], Maleki et al. proposes Complex AMP, and characterizes the performance of LASSO for sparse signals with complex entries. On the other hand, Stojnic’s approach [56] relies on results on Gaussian processes [33, 34] to derive sharp bounds for the worst case NSE of the $\ell_1$-constrained LASSO problem in (17). Subsequent work in [47] builds upon the analysis framework introduced in [56] and generalizes the results of the latter to the regularized LASSO and to arbitrary convex functions. In [46] and [60], sharp upper bounds on the NSE of the C-LASSO and the $\ell_2$-LASSO are derived, for arbitrary noise vector provided that it is chosen independently of the sensing matrix. The outcome of those works is a general theory that yields precise bounds for the NSE of the LASSO algorithm for arbitrary convex regularizer $f(\cdot)$. The bounds are simple in nature and offer nice interpretations as generalizations of the classical OLS error bounds discussed in Section 2.

3.2 Motivating Examples

In Section 2 and in particular in equations (9)–(13), we reviewed classical bounds on the normalized square-error of the OLS, which corresponds to the LASSO in the trivial case $f(\cdot) = 0$. How do those result change when a nontrivial convex function $f(\cdot)$ is introduced? What is a precise and simple upper bound on the NSE of the LASSO when the unknown signal is sparse and $f(\cdot) = \| \cdot \|_1$? What if the unknown signal is low-rank and nuclear norm is chosen as the regularizer? Is it possible to generalize such bounds to arbitrary structures and corresponding convex regularizers? In the following paragraphs, we assume that the entries of the sensing matrix $A$ are i.i.d. zero-mean normal with variance $1/m$ and provide answers to those questions. The details and proofs of the statements made are deferred to Sections 5–7.
3.2.1 Sparse Signal Estimation

Assume \( x_0 \in \mathbb{R}^n \) has \( k \) nonzero entries. We estimate \( x_0 \) via the LASSO with \( f \) being the \( \ell_1 \)-norm. First, suppose that the noise vector has i.i.d. zero-mean normal entries with variance \( \sigma^2 \). Then, the NSE of the C-LASSO admits the following sharp upper bound\(^4\), which is attained in the limit as the noise variance \( \sigma^2 \) goes to zero [47, 56]:

\[
\frac{\| \hat{x}_c - x_0 \|_2^2}{\| z \|_2^2} \lesssim \frac{2k(\log \frac{n}{k} + 1)}{m - 2k(\log \frac{n}{k} + 1)}.
\] (20)

Compare this to the formula (9) for the OLS. (20) is obtained from (9) after simply replacing the ambient dimension \( n \) in the latter with \( 2k(\log \frac{n}{k} + 1) \). Also, while (9) requires \( m > n \), (20) relaxes this requirement to \( m > 2k(\log \frac{n}{k} + 1) \). This is to say that any number of measurements greater than \( 2k(\log \frac{n}{k} + 1) \ll n \) are sufficient to guarantee robust recovery. Note that this coincides with the classical phase-transition threshold in the noiseless compressed sensing [14, 55]. If instead of the C-LASSO, one uses the \( \ell_2 \)-LASSO with \( \lambda \geq \sqrt{2\log \frac{n}{k}} \), then

\[
\frac{\| \hat{x}_{\ell_2} - x_0 \|_2^2}{\| z \|_2^2} \lesssim \frac{(\lambda^2 + 3)k}{m - (\lambda^2 + 3)k}.
\] (21)

Again, observe how (21) is obtained from (9) after simply replacing the ambient dimension \( n \) with \( (\lambda^2 + 3)k \). The role of the regularizer parameter \( \lambda \) is explicit in (21). Also, substituting \( \lambda \approx \sqrt{2\log \frac{n}{k}} \) in (21) (almost) recovers (20). This suggests that choosing this value of the regularizer parameter is optimal in that it results in the regularized LASSO (7) to perform as good as the constrained version (6). Note that this value for the optimal regularizer parameter, only depends on the sparsity level \( k \) of the unknown signal \( x_0 \) and not the unknown signal itself.

Next, consider the more general case in which the noise vector \( z \) can be anything but drawn independently of the sensing matrix \( A \). If ones uses the C-LASSO to estimate \( x_0 \), then the estimation error is bounded as follows\(^5\) [46]:

\[
\frac{\| \hat{x}_c - x_0 \|_2^2}{\| z \|_2^2} \lesssim \frac{2k(\log \frac{n}{k} + 1)}{(\sqrt{m} - 2k(\log \frac{n}{k} + 1))^2}.
\] (22)

Accordingly, the \( \ell_2 \)-LASSO for \( \lambda \geq \sqrt{2\log \frac{n}{k}} \) gives [60]:

\[
\frac{\| \hat{x}_{\ell_2} - x_0 \|_2^2}{\| z \|_2^2} \lesssim \frac{(\lambda^2 + 3)k}{(\sqrt{m} - (\lambda^2 + 3)k)^2}.
\] (23)

---

\(^4\) The statements in this sections hold true with high probability in \( A, z \) and under mild assumptions. See Section 5 for the formal statement of the results.

\(^5\) The formula below is subject to some simplifications meant to highlight the essential structure. See Section 6 for the details.
Once more, (22) and (23) resemble the corresponding formula describing OLS in (11). The only difference is that the ambient dimension $n$ is substituted with $2k(\log \frac{n}{k} + 1)$ and $(\lambda^2 + 3)k$, respectively. As a last comment, observe that the bounds in (22) and (23) have squares in the denominators. This is in contrast to equations (20) and (21). A detailed comparison is included in Section 6.1.2.

3.2.2 Low-rank Matrix Estimation

Assume $X_0 \in \mathbb{R}^{\sqrt{n} \times \sqrt{n}}$ is a rank-$r$ matrix, and let $x_0 = \text{vec}(X_0) \in \mathbb{R}^n$ be the vectorization of $X_0$. We use the generalized LASSO with $f(x) = \|\text{vec}^{-1}(x)\|_*$. The nuclear norm of a matrix (i.e. sum of singular values) is known to promote low-rank solutions [30, 50].

As previously, suppose first that $z$ has i.i.d. zero-mean normal entries with variance $\sigma^2$. Then, the NSE of the C-LASSO and that of the $\ell_2$-LASSO for $\lambda \geq 2n^{1/4}$ are bounded as follows:

$$\frac{\|\hat{x}_c - x_0\|_2^2}{\|z\|_2^2} \lesssim \frac{6\sqrt{n}r}{m - 6\sqrt{n}r},$$  \hspace{1cm} (24)

and

$$\frac{\|\hat{x}_c - x_0\|_2^2}{\|z\|_2^2} \lesssim \frac{\lambda^2 r + 2\sqrt{n}(r + 1)}{m - (\lambda^2 r + 2\sqrt{n}(r + 1))}. \hspace{1cm} (25)$$

Just like in the estimation of sparse signals in Section 3.2.1, it is clear from the bounds above that they can be obtained from the OLS bound in (9) after only substituting the dimension of the ambient space $n$ with $6\sqrt{n}r$ and $\lambda^2 r + 2\sqrt{n}(r + 1)$, respectively. And again, $6\sqrt{n}r$ is exactly the phase transition threshold for the noiseless compressed sensing of low-rank matrices [1, 14, 45].

Moving to the case where $z$ is arbitrary but independent of $A$, we find that

$$\frac{\|\hat{x}_c - x_0\|_2^2}{\|z\|_2^2} \lesssim \frac{6\sqrt{n}r}{(\sqrt{m} - 6\sqrt{n}r)^2}, \hspace{1cm} (26)$$

and

$$\frac{\|\hat{x}_{\ell_2} - x_0\|_2^2}{\|z\|_2^2} \lesssim 2\frac{\lambda^2 r + 2\sqrt{n}(r + 1)}{(\sqrt{m} - \sqrt{\lambda^2 r + 2\sqrt{n}(r + 1)})^2}. \hspace{1cm} (27)$$

3.2.3 General Structures

From the discussion in Sections 3.2.1 and 3.2.2, it is becoming clear that the error bounds for the OLS admit nice and simple generalizations to error bounds for the generalized LASSO. What changes in the formulae bounding the NSE of the OLS

---

\footnote{The factor of 2 in (23) is conjectured in [60] that is not essential and, only, appears as an artifact of the proof technique therein. See, also, Section 6.2.1.}
when considering the NSE of the LASSO, is only that the ambient dimension \( n \) is substituted by a specific parameter.

This parameter depends on the particular structure of the unknown signal, but not the signal itself. For example, in the sparse case, it depends only on the sparsity of \( x_0 \), not \( x_0 \) itself, and in the low-rank case, it only depends on the rank of \( X_0 \), not \( X_0 \) itself. Furthermore, it depends on the structure-inducing function \( f(\cdot) \) that is being used. Finally, it is naturally dependent on whether the constrained or the regularized LASSO is being used. In the case of regularized LASSO, it also depends on the value \( \lambda \) of the regularizer parameter. Interestingly, the value of this parameter corresponding to the NSE of the constrained LASSO is exactly the phase-transition threshold of the corresponding noiseless CS problem.

Let us use the following notation, \( D(\text{cone}(\partial f(x_0))) \) and \( D(\lambda \partial f(x_0)) \) to represent this parameter for the cases of the constrained and regularized LASSO, correspondingly. The notation used makes explicit the dependence of the parameter on the quantities just discussed. The formal definition of \( D(\text{cone}(\partial f(x_0))) \) and \( D(\lambda \partial f(x_0)) \) is not at all involved. In fact, they both have a nice and insightful geometric interpretation and are not hard to compute for many practical examples of interest (see the Appendix). We will present all these in Section 4.

To conclude this section, we repeat once more: the classical and well-known error analysis of the NSE of the OLS can be naturally extended to describe the NSE of the generalized LASSO. In particular, when the entries of \( A \) are i.i.d. normal, then an error bound on the NSE of the OLS translates to a bound on the NSE of the generalized (constrained or regularized) LASSO after (almost) only substituting the ambient dimension \( n \) by either \( D(\text{cone}(\partial f(x_0))) \) or \( D(\lambda \partial f(x_0)) \). \( D(\text{cone}(\partial f(x_0))) \) and \( D(\lambda \partial f(x_0)) \) are summary parameters that capture the geometry of the LASSO problem. For instance, for \( f(x) = \ell_1 \) and \( x_0 \) a \( k \)-sparse vector, it will be seen that \( D(\text{cone}(\partial f(x_0))) = 2k \log \frac{n}{k} \). Hence, one obtains (20) and (22) when substituting the \( n \) with \( 2k \log \frac{n}{k} \) in (9) and (11), respectively.

### 4 Background

This section is meant to summarize the background required for the formal statement of the results on the NSE of the generalized LASSO. In Section 4.2 we revise fundamental notions of convex geometry. Next, in Section 4.3 we outline the main tools that underly the analysis. These include a strong probabilistic comparison lemma on Gaussian processes proven by Gordon in [34], as well as some standard but powerful concentration inequalities on the Gaussian measure.
4.1 Notation

For the rest of the paper, let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution of mean $\mu$ and variance $\sigma^2$. Also, to simplify notation, let us write $\| \cdot \|$ instead of $\| \cdot \|_2$. For a vector $\mathbf{g} \in \mathbb{R}^m$ with independent $\mathcal{N}(0, 1)$ entries, we define $\gamma_m := \mathbb{E}[\|\mathbf{g}\|]$. It is well known ([34]) that

$$\gamma_m = \sqrt{2 \Gamma\left(\frac{m+1}{2}\right)} \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{m+1}}$$

and

$$\sqrt{m} \geq \gamma_m \geq m \sqrt{\frac{m}{m+1}}.$$ 

Finally, the Euclidean unit ball and unit sphere are respectively denoted as

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

4.2 Convex Geometry

4.2.1 Subdifferential

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $\mathbf{x}_0 \in \mathbb{R}^n$ be an arbitrary point that is not a minimizer of $f(\cdot)$. The subdifferential of $f(\cdot)$ at $\mathbf{x}_0$ is the set of vectors,

$$\partial f(\mathbf{x}_0) = \{\mathbf{s} \in \mathbb{R}^n \mid f(\mathbf{x}_0 + \mathbf{w}) \geq f(\mathbf{x}_0) + \mathbf{s}^T \mathbf{w}, \forall \mathbf{w} \in \mathbb{R}^n\}.$$ 

$\partial f(\mathbf{x}_0)$ is a convex and closed set [52]. For our purposes, further assume that it is nonempty and bounded [62]. It also, does not contain the origin since we assumed that $\mathbf{x}_0$ is not a minimizer. For any number $\lambda \geq 0$, we denote the scaled (by $\lambda$) subdifferential as $\lambda \partial f(\mathbf{x}_0) = \{\lambda \mathbf{s} \mid \mathbf{s} \in \partial f(\mathbf{x}_0)\}$.

4.2.2 Convex Cones

A convex cone $\mathcal{K} \subset \mathbb{R}^n$ is a set that is convex and closed under multiplication by positive scalars. The polar cone $\mathcal{K}^\circ$ is defined as the closed convex cone

$$\mathcal{K}^\circ := \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p}^T \mathbf{s} \leq 0 \text{ for all } \mathbf{s} \in \mathcal{K}\}.$$ 

For the conic hull of a set $\mathcal{C} \subset \mathbb{R}^n$, we write

$$\text{cone}(\mathcal{C}) := \{\mathbf{s} \mid \mathbf{s} \in \lambda \mathcal{C}, \text{ for some } \lambda \geq 0\}.$$ 

In particular, we denote the convex hull of the subdifferential as $\text{cone}(\partial f(\mathbf{x}_0))$.

Next, we introduce the notion of the tangent cone $\mathcal{T}_f(\mathbf{x}_0)$ of $f$ at $\mathbf{x}_0$. First, define the set of descent directions of $f$ at $\mathbf{x}_0$ as $\mathcal{T}_f(\mathbf{x}_0) = \{\mathbf{v} \mid f(\mathbf{x}_0 + \mathbf{v}) \leq f(\mathbf{x}_0)\}$. The tangent cone $\mathcal{T}_f(\mathbf{x}_0)$ of $f$ at $\mathbf{x}_0$ is defined as

$$\mathcal{T}_f(\mathbf{x}_0) := \text{Cl}(\text{cone}(\mathcal{T}_f(\mathbf{x}_0)))$$,
where \( \text{Cl}(\cdot) \) denotes the closure of a set. Under the assumption that \( \mathbf{x}_0 \) is not a minimizer of \( f(\cdot) \), then the polar of the tangent cone (commonly termed the normal cone [5]) can be equivalently written [52, p. 222] as the conic hull of the subdifferential.

\[
(T_f(x_0))^\circ = \text{cone}(\partial f(x_0)).
\]  

(28)

4.2.3 Gaussian Squared Distance

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

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

The distance of \( \mathbf{v} \) to the set \( \mathcal{C} \) can then be written as

\[
\text{dist}(\mathbf{v}, \mathcal{C}) := \| \mathbf{v} - \text{Proj}(\mathbf{v}, \mathcal{C}) \|.
\]

![Fig. 1: Illustration of the distance of a vector to the dilated subdifferential \( \lambda \partial f(\mathbf{x}_0) \).](image)

**Definition 4.1 (Gaussian squared distance)** Let \( \mathbf{h} \in \mathbb{R}^n \) have i.i.d. \( \mathcal{N}(0, 1) \) entries. The Gaussian squared distance of a set \( \mathcal{C} \subset \mathbb{R}^n \) is defined as

\[
\mathcal{D}(\mathcal{C}) := \mathbb{E}_h \left[ \text{dist}^2(\mathbf{h}, \mathcal{C}) \right].
\]

Accordingly, define the Gaussian projection and correlation as

\[
\mathcal{P}(\mathcal{C}) := \mathbb{E} \left[ \| \text{Proj}(\mathbf{h}, \mathcal{C}) \|^2 \right] \quad \text{and} \quad \mathcal{C}(\mathcal{C}) := \mathbb{E} \left[ (\mathbf{h} - \text{Proj}(\mathbf{h}, \mathcal{C}))^T \text{Proj}(\mathbf{h}, \mathcal{C}) \right].
\]

Of particular interest to us is the Gaussian squared distance to the scaled subdifferential and to the cone of subdifferential

\[
\mathcal{D}(\lambda \partial f(\mathbf{x}_0)) = \mathbb{E} \left[ \text{dist}^2(\mathbf{h}, \lambda \partial f(\mathbf{x}_0)) \right],
\]
D(cone(∂f(x₀))) = E[ dist²(h, cone(λ∂f(x₀))) ].

Following (28), observe that

\[ D(cone(∂f(x₀))) = D((T_0(x₀))^\circ). \]

### 4.2.4 Gaussian Squared Distance to the Scaled Subdifferential

The Gaussian squared distance of the cone of subdifferential \( D(cone(∂f(x₀))) \) appears as a fundamental quantity in the Noiseless CS problem. In particular, [14, 55] prove that \( D(cone(∂f(x₀))) \) number of measurements suffice so that (14) uniquely recovers \( x₀ \). Later on, [1] and [57] showed\(^7\) that this number of measurements is also necessary. Hence, \( D(cone(∂f(x₀))) \) precisely characterizes the phase-transition in the noiseless CS\(^8\).

The Gaussian squared distance of the scaled subdifferential \( D(λ∂f(x₀)) \) naturally arises in the analysis of the normalized mean-square error of the proximal denoising (15). Under the assumption of \( z \) having i.i.d. \( \mathcal{N}(0, σ^2) \) entries, [43] shows that the normalized mean-square error of (15) admits a sharp upper bound (attained for \( σ \to 0 \)) equal to \( D(λ∂f(x₀)) \). For the constrained proximal denoiser in (16) the corresponding upper bound becomes \( D(cone(∂f(x₀))) \).

There is also a deep relation between the two quantities \( D(cone(∂f(x₀))) \) and \( D(λ∂f(x₀)) \), stronger than the obvious fact that

\[ D(cone(∂f(x₀))) ≤ \min_{λ ≥ 0} D(λ∂f(x₀)). \]

In particular, it can be shown [1, 31, 43] that in high-dimensions

\[ D(cone(∂f(x₀))) ≈ \min_{λ ≥ 0} D(λ∂f(x₀)). \quad (29) \]

Moreover, as the next lemma shows, the minimum of \( D(λ∂f(x₀)) \) in (29) is uniquely attained. The lemma also reveals an interesting relation between \( D(λ∂f(x₀)) \) and \( C(λ∂f(x₀)) \). We will make use of this property in the analysis of the NSE of the \( ℓ_2 \)-LASSO in Section 6.

**Lemma 1** ([1]). Suppose \( ∂f(x₀) \) is nonempty and does not contain the origin. Then,

1. \( D(λ∂f(x₀)) \) is a strictly convex function of \( λ ≥ 0 \), and is differentiable for \( λ > 0 \).
2. \( \frac{∂D(λ∂f(x₀))}{∂λ} = -\frac{2}{λ} C(λ∂f(x₀)). \)

In the Appendix we show how to use the definitions and the properties just discussed to compute highly-accurate and closed-form approximations for the quanti-\(^7\) The tools used in [1] and [57] differ. Amelunxen et. al [1] use tools from conic integral geometry, while Stojnic [55] relies on a comparison lemma for Gaussian processes (see Lemma 3).
\(^8\) In [14], \( D(cone(∂f(x₀))) \) shows up indirectly via a closely related notion, that of the “Gaussian width” [34] of the the restricted tangent cone \( T_0(x₀) \cap \mathbb{R}^{n−1} \). In the terminology used in [1, 40], \( D(cone(∂f(x₀))) \) corresponds to the “statistical dimension” of \( T_0(x₀) = (cone(∂f(x₀)))^\circ \).
ties $D(\lambda \partial f(x_0))$ and $D(\text{cone}(\partial f(x_0)))$ for a number of practical convex regularizers and signal structures.

### 4.3 Probability Tools

#### 4.3.1 Gordon’s Lemma

Perhaps the most important technical ingredient underlying the analysis of the NSE of the generalized LASSO is the following lemma proved by Gordon in [34]. Gordon’s Lemma establishes a very useful (probabilistic) inequality for Gaussian processes.

**Lemma 2 (Gordon’s Lemma [34]).** Let $G \in \mathbb{R}^{m \times n}, g \in \mathbb{R}^m, h \in \mathbb{R}^n$ be independent of each other and have independent standard normal entries. Also, let $\mathcal{S} \subset \mathbb{R}^n$ be an arbitrary set and $\psi : \mathcal{S} \rightarrow \mathbb{R}$ be an arbitrary function. Then, for any $c \in \mathbb{R}$,

$$
\Pr \left( \min_{x \in \mathcal{S}} \{ \|Gx\| + \|x\|_g - \psi(x) \} \geq c \right) \geq \Pr \left( \min_{x \in \mathcal{S}} \{ \|x\|_g - h^T x - \psi(x) \} \geq c \right).
$$

(30)

For the analysis of the LASSO, we use a slightly modified version of the original lemma (see [47, Lemma 5.1]).

**Lemma 3 (Modified Gordon’s Lemma).** $G \in \mathbb{R}^{m \times n}, g \in \mathbb{R}^m$ and $h \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, 1)$ entries and be independent of each other. Also, let $\psi(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, and, $\Phi_1 \subset \mathbb{R}^n$ and $\Phi_2 \subset \mathbb{R}^m$ such that either both $\Phi_1$ and $\Phi_2$ are compact or $\Phi_1$ is arbitrary and $\Phi_2$ is a scaled unit sphere. Then, for any $c \in \mathbb{R}$:

$$
\Pr \left( \min_{x \in \Phi_1} \max_{a \in \Phi_2} \{ a^T Gx - \psi(x, a) \} \geq c \right) \geq 2 \Pr \left( \min_{x \in \Phi_1} \max_{a \in \Phi_2} \{ \|x\|_g a - \|a\|_h^T x - \psi(x, a) \} \geq c \right) - 1.
$$

(31)

It is worth mentioning that the “escape through a mesh” lemma, which has been the backbone of the approach introduced by Stojnic [55] (and subsequently refined in [14]) for computing an asymptotic upper bound to the minimum number of measurements required in the Noiseless CS problem, is a corollary of Lemma 2. The “escape through a mesh” lemma gives a bound on the “restricted minimum singular value” of an operator $G$, which is defined as the minimum gain of $G$ restricted to a subset of the unit sphere. The concept is similar to the restricted isometry property and has been topic of several related works, [6, 14, 42, 49].

Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex subset of the unit $\ell_2$-sphere $S^{n-1}$. Then, the minimum singular value of $A \in \mathbb{R}^{m \times n}$ restricted to $\mathcal{C}$ is defined as
\[ \sigma_{\min}(A, \mathcal{C}) = \min_{v \in \mathcal{C}} \| Av \|. \]

Observe that, \[ \sigma_{\min}(A, \mathcal{S}^{n-1}) \] reduces to the standard definition of the minimum singular value of the matrix \( A \).

**Lemma 4 (Restricted eigenvalue or Escape through a mesh).** Let \( G \in \mathbb{R}^{m \times n} \) have i.i.d. standard normal entries and \( \mathcal{K} \subset \mathbb{R}^n \) be a convex cone. Assume \( 0 < t \leq \gamma_m - \sqrt{D(\mathcal{K}^\circ)} \). Then,

\[ \mathbb{P} \left( \min_{v \in \mathcal{K} \cap \mathcal{S}^{n-1}} \| Gv \| \geq \gamma_m - \sqrt{D(\mathcal{K}^\circ)} - t \right) \geq 1 - \exp \left( -\frac{t^2}{2} \right). \]

### 4.3.2 Gaussian Concentration of Lipschitz Functions

The upper bounds on the NSE of the LASSO, similar to the OLS case, are probabilistic in nature. In particular they hold with high probability over the realizations of \( A \) (and occasionally of \( z \)). Key to the derivation of such probabilistic statements is the concentration of measure phenomenon. The lemma below is a well-known and powerful result on the concentration of Lipschitz functions under the Gaussian measure. Recall that a function \( \psi(\cdot) : \mathbb{R}^n \to \mathbb{R} \) is \( L \)-Lipschitz, if for all \( x, y \in \mathbb{R}^n \),

\[ |\psi(x) - \psi(y)| \leq L \|x - y\|. \]

**Lemma 5 (Lipschitz concentration, [37]).** Let \( h \in \mathbb{R}^n \) have i.i.d. \( \mathcal{N}(0, 1) \) entries and \( \psi : \mathbb{R}^n \to \mathbb{R} \) be an \( L \)-Lipschitz function. Then, \( \text{Var}[\psi(h)] \leq L^2 \). Furthermore, for all \( t > 0 \), the events \( \{ \psi(h) - \mathbb{E}[\psi(h)] \geq t \} \) and \( \{ \psi(h) - \mathbb{E}[\psi(h)] \leq -t \} \) hold with probability no greater than \( \exp(-t^2/(2L^2)) \), each.

## 5 The NSE of Generalized LASSO in Gaussian Noise

In this Section we assume that the noise vector \( z \) has entries distributed i.i.d. normal \( \mathcal{N}(0, \sigma^2) \) and derive sharp upper bounds on the NSE of the generalized LASSO. We begin in Section 5.1 with describing the main steps of the required technical analysis. That is only a highlight of the key ideas and insights and we provide specific references in [47] for the details.\(^9\) Next, Sections 5.2, 5.3 and 5.4 are each devoted to upper-bounding the NSE of the C-LASSO, \( \ell_2 \)-LASSO and \( \ell_2^2 \)-LASSO respectively.

---

\(^9\) When referring to [47] keep in mind the following: a) in [47] the entries of \( A \) have variance 1 and not \( 1/m \) as here, b) [47] uses slightly different notation for \( D(\lambda \partial f(x_0)) \) and \( D(\text{cone}(\partial f(x_0))) \) \((\mathcal{D}_f(x_0, \lambda) \) and \( \mathcal{D}_f(x_0, \mathbb{R}^+) \), respectively).
5.1 Technical Framework

Proving the sharp upper bounds for the NSE of the generalized LASSO requires several steps and can be challenging to include all the details in the limited space of a book chapter. However, the main strategy for the proof and the fundamental tools and ideas being used are neat and not hard to appreciate. This will be the purpose of this section. The framework to be presented was first introduced by Stojnic in [56] and subsequently simplified and extended in [47].

The bulk of the results to be presented in this section are for the C-LASSO and the $\ell_2$-LASSO. Based on this and a mapping between the $\ell_2$-LASSO and the $\ell_2$-LASSO, some conclusions can be drawn for the NSE of the $\ell_2$-LASSO, as well. For the purposes of exposition we use the $\ell_2$-LASSO. The analysis for the constrained version C-LASSO is in a large extent similar. In fact, [47] treats those two under a common framework.

5.1.1 First-Order Approximation

Recall the $\ell_2$-LASSO problem introduced in (18):

$$\hat{x}_{\ell_2} = \arg\min_x \| y - Ax \| + \frac{\lambda}{\sqrt{m}} f(x). \tag{31}$$

A key idea behind our approach is using the linearization of the convex structure inducing function $f(\cdot)$ around the vector of interest $x_0$ [7, 52]. From convexity of $f$, for all $x \in \mathbb{R}^n$ and $s \in \partial f(x_0)$, we have $f(x) \geq f(x_0) + s^T(x - x_0)$. In particular,

$$f(x) \geq f(x_0) + \sup_{s \in \partial f(x_0)} s^T(x - x_0) =: \hat{f}(x), \tag{32}$$

and approximate equality holds when $\| x - x_0 \|$ is “small”. Recall that $\partial f(x_0)$ denotes the subdifferential of $f(\cdot)$ at $x_0$ and is always a compact and convex set [52]. We also assume that $x_0$ is not a minimizer of $f(\cdot)$, hence, $\partial f(x_0)$ does not contain the origin.

We substitute $f(\cdot)$ in (31) 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 $y = Ax_0 + z = Ax_0 + \sigma v$, for $v \sim \mathcal{N}(0, I_m)$ and change the optimization variable from $x$ to $w = x - x_0$:

$$\tilde{w}_{\ell_2} = \arg\min_w \left\{ \| Aw - \sigma v \| + \frac{1}{\sqrt{m}} \sup_{s \in \partial \hat{f}(x_0)} s^T w \right\}. \tag{33}$$
We denote $\tilde{w}_2$ the optimal solution of the approximated problem in (33) and $\hat{w}_2 = \hat{x}_2 - x_0$ for the optimal solution of the original problem in (31)\(^\text{10}\). Also, denote the optimal cost achieved in (33) by $\tilde{F}_{\ell_2}(A,v)$, as $\tilde{F}_{\ell_2}(A,v)$. Finally, note that the approximated problem corresponding to C-LASSO can be written as in (33), with only $\lambda \partial f(x_0)$ being substituted by $\text{cone}(\partial f(x_0))$.

Taking advantage of the simple characterization of $\hat{f}(\cdot)$ via the subdifferential $\partial f(x_0)$, we are able to precisely analyze the optimal cost and the normalized squared error of the resulting approximated problem. The approximation is tight when $\|\hat{x}_2 - x_0\| \to 0$ 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.

5.1.2 Importance of $\sigma \to 0$

Our focus is on the precise characterization of the NSE. While we show that the first order characteristics of the function, i.e. $\partial f(x_0)$, suffice to provide sharp and closed-form bounds for small noise level $\sigma$, we believe that higher order terms are required for such precise results when $\sigma$ is arbitrary. On the other hand, empirical observations suggest that the worst case NSE for the LASSO problem is achieved when $\sigma \to 0$. This statement admits a rigorous proof in the case of the C-LASSO. Interestingly, the same phenomena has been observed and proved to be true for related estimation problems. Examples include the proximal denoising problem (15) in [15, 22, 43] and, the LASSO problem with $\ell_1$ penalization [2].

To recap, in what follows we derive formulae that sharply characterize the NSE of the generalized LASSO in the small $\sigma$ regime. For the C-LASSO Section 10 in [47] proves that the corresponding such formula upper bounds the NSE when $\sigma$ is arbitrary. Extended numerical simulations suggest that this is also the case for the regularized LASSO.

5.1.3 Applying Gordon’s Lemma

The (approximated) LASSO problem in (33) is simpler than the original one in (44), yet, still hard to directly analyze. We will apply Gordon’s Lemma 3 to further simplify the problem. This trick is critical and repeatedly used in our analysis. First, write $\sqrt{m}\|Aw - \sigma v\| = \max_{\|a\|=1} a^T [\sqrt{mA}, -v] \begin{bmatrix} w \\ \sigma \sqrt{m} \end{bmatrix}$ and choose the function $\psi(\cdot)$ in Lemma 3 to be $\sup_{s: \lambda \partial f(x_0)} s^T w$. Then, the following result is a simple corollary of applying Gordon’s Lemma to the LASSO objective. Recall that we write $\tilde{F}_{\ell_2}(A,v)$ for the optimal objective value of the approximated LASSO in (33).

\(^{10}\) We follow this conventions throughout: use the symbol “\~{}” over variables that are associated with the approximated problems. To distinguish, use the symbol “\^{}” for the variables associated with the original problem.
Corollary 5.1 (Lower Key Optimization) Let \( g \sim \mathcal{N}(0, I_m) \) and \( h \sim \mathcal{N}(0, I_n) \) be independent of each other. Define the following optimization problem:

\[
L(g, h) = \min_w \left\{ \sqrt{\|w\|^2 + m\sigma^2} \|g\| - h^T w + \max_{s \in \lambda \partial f(x_0)} s^T w \right\}. \tag{34}
\]

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

\[
P\left( \mathcal{F}_{\ell_2}(A, v) \geq c \right) \geq 2P \left( L(g, h) \geq c \sqrt{m} \right) - 1.
\]

Corollary 5.1 establishes a probabilistic connection between the LASSO problem and the minimization (34). The advantage of the latter is that it is much easier to analyze. Intuitively, the main reason for that is that instead of an \( m \times n \) matrix, (34) 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 \( L(g, h) \). Due to its central role in our analysis, we often refer to problem (34) as “key optimization” or “lower key optimization”. The term “lower” is attributed to the fact that analysis of (34) results in a probabilistic lower bound for the optimal cost of the LASSO problem.

5.1.4 Analyzing the Key Optimization

--- Deterministic Analysis: First, we perform the deterministic analysis of \( L(g, h) \) for fixed \( g \in \mathbb{R}^m \) and \( h \in \mathbb{R}^n \). In particular, we reduce the optimization in (34) to a scalar optimization. To see this, perform the optimization over a fixed \( \ell_2 \)-norm of \( w \) to equivalently write

\[
L(g, h) = \min_{\alpha \geq 0} \left\{ \sqrt{\alpha^2 + m\sigma^2} \|g\| - \max_{\|w\| = \alpha} \min_{s \in \lambda \partial f(x_0)} (h - s)^T w \right\}. \tag{35}
\]

The maximin problem that appears in the objective function of the optimization above has a simple solution. It can be shown [47, Lemma E.1] that if \( h \notin \lambda \partial f(x_0) \):

\[
\max_{\|w\| = \alpha} \min_{s \in \lambda \partial f(x_0)} (h - s)^T w = \min_{s \in \lambda \partial f(x_0)} \max_{\|w\| = \alpha} (h - s)^T w
\]

\[
= \alpha \min_{s \in \lambda \partial f(x_0)} \|h - s\|.
\]

This reduces (34) to a scalar optimization problem over \( \alpha \), for which one can compute the optimal value \( \hat{\alpha} \) and the corresponding optimal cost. The result is summarized in Lemma 6 below.

Lemma 6 (Deterministic Result). Let \( \tilde{w}(g, h) \) be a minimizer of the problem in (34). If \( \|g\| > \text{dist}(h, \lambda \partial f(x_0)) > 0 \), then,
a) \[ \| \hat{w}(g, h) \|^2 = m \sigma^2 \frac{\text{dist}^2(h, \lambda \partial f(x_0))}{\| g \|^2 - \text{dist}^2(h, \lambda \partial f(x_0))}, \]

b) \[ \mathcal{L}(g, h) = \sqrt{m} \sigma \sqrt{\| g \|^2 - \text{dist}^2(h, \lambda \partial f(x_0))}. \]

**Probabilistic Analysis:** Of interest is making probabilistic statements about \( \mathcal{L}(g, h) \) and the norm of its minimizer \( \| \hat{w}(g, h) \| \). Lemma 6 provided closed-form deterministic solutions for both of them, which only involve the quantities \( \| g \|^2 \) and \( \text{dist}^2(h, \lambda \partial f(x_0)) \). The \( \ell_2 \)-norm and the distance function to a convex set are 1-Lipschitz functions. Application of Lemma 5, then shows that \( \| g \|^2 \) and \( \text{dist}^2(h, \lambda \partial f(x_0)) \) concentrate nicely around their means \( \mathbb{E} [\| g \|^2] = m \) and \( \mathbb{E} [\text{dist}^2(h, \lambda \partial f(x_0))] = D(\lambda \partial f(x_0)) \), respectively. Combining this with Lemma 6, we conclude with Lemma 7 below.

**Lemma 7 (Probabilistic Result).** Assume that \( (1 - \varepsilon_L)m \geq D(\lambda \partial f(x_0)) \geq \varepsilon_L m \) for some constant \( \varepsilon_L > 0 \). Define\(^{11}\)

\[ \eta = \sqrt{m - D(\lambda \partial f(x_0))} \quad \text{and} \quad \gamma = \frac{D(\lambda \partial f(x_0))}{m - D(\lambda \partial f(x_0))}. \]

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

\[ |\mathcal{L}(g, h) - \sqrt{m} \sigma \eta| \leq \varepsilon \sqrt{m} \sigma \eta, \quad \text{and} \quad \left| \frac{\| \hat{w}(g, h) \|^2}{m \sigma^2} - \gamma \right| \leq \varepsilon \gamma. \]

**Remark:** In Lemma 7, the condition \( (1 - \varepsilon_L)m \geq D(\lambda \partial f(x_0)) \geq \varepsilon_L m \) ensures that \( \| g \| > \text{dist}(h, \lambda \partial f(x_0)) > 0 \) (cf. Lemma 6) with high probability over the realizations of \( g \) and \( h \).

### 5.1.5 From the Key Optimization Back to the LASSO

Before proceeding, let us recap. Application of Gordon’s Lemma to the approximated LASSO problem in (33) introduced the simpler lower key optimization (34). Without much effort, we found in Lemma 7 that its cost \( \mathcal{L}(g, h) \) and the normalized squared norm of its minimizer \( \| \hat{w}(g, h) \|^2 \) concentrate around \( \sqrt{m} \sigma \eta \) and \( \gamma \), respectively. This brings us to the following question:

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

Application of Gordon’s Lemma as performed in Corollary 5.1 when combined with Lemma 7, provides a first answer to this question: \( \mathcal{F}_{\ell_2}(A, v) \) is lower bounded by \( \sigma \eta \) with overwhelming probability. Formally,

\(^{11}\) Observe that the dependence of \( \eta \) and \( \gamma \) on \( \lambda \), \( m \) and \( \partial f(x_0) \), is implicit in this definition.
Lemma 8 (Lower Bound). Assume $(1 - \varepsilon_L)m \geq D(\lambda \hat{\partial} f(x_0)) \geq \varepsilon_L m$ for some constant $\varepsilon_L > 0$ and $m$ is sufficiently large. Then, for any $\varepsilon > 0$, there exists a constant $c > 0$ such that, with probability $1 - \exp(-cm)$,

$$F_{\ell_2}(A,v) \geq (1 - \varepsilon)\sigma\eta.$$

But is that all? It turns out that the connection between the LASSO problem and the simple optimization (34) is much deeper than Lemma 8 predicts. In short, under certain conditions on $\lambda$ and $m$ (similar in nature to those involved in the assumption of Lemma 8), we can prove the following being true:

- Similar to $\mathcal{L}(g,h)$, the optimal cost $F_{\ell_2}(A,v)$ of the approximated $\ell_2$-LASSO concentrates around $\sigma\eta$.
- Similar to $\|\hat{\omega}(g,h)\|^2_{2m \sigma^2}$, the NSE of the approximated $\ell_2$-LASSO $\|\hat{\omega}_{\ell_2}(A,v)\|^2_{2m \sigma^2}$ concentrates around $\gamma$.

In some sense, $\mathcal{L}(g,h)$ “predicts” $F_{\ell_2}(A,v)$ and $\|\hat{\omega}(g,h)\|$ “predicts” $\|\hat{\omega}_{\ell_2}(A,v)\|$.

The main effort in [47, 56] is in proving those claims. The next section contains a synopsis of the main steps of the proof.

5.1.6 Synopsis of the Technical Framework

1. Apply Gordon’s Lemma to the dual of $F_{\ell_2}(A,v)$ to compute a high-probability upper bound for it. Following essentially the same ideas as in Section 5.1.4 it is shown that $F_{\ell_2}(A,v)$ is no larger than $\sigma\eta$ with high probability (see Lemmas 5.2 and 6.2 in [47]).

2. Assume $\|\hat{\omega}_{\ell_2}\|^2_{2m \sigma^2}$ deviates from $\gamma$. Yet another application of Gordon’s Lemma shows that such a deviation would result in a significant increase in the optimal cost, namely $F_{\ell_2}(A,v)$ would be significantly larger than $\sigma\eta$ (see Lemmas 5.2 and 6.3 in [47]).

3. Combining the previous steps, shows that $\|\hat{\omega}_{\ell_2}\|^2_{2m \sigma^2}$ concentrates with high probability around $\gamma$ (see Lemma 6.4 in [47]).

4. The final step requires us to translate this bound on the NSE of the Approximated LASSO to a bound on the NSE of the original one. We choose $\sigma$ small enough such that $\|\hat{\omega}_{\ell_2}\|$ is small and so $f(x_0 + \hat{\omega}_{\ell_2}) \approx f(x_0 + \hat{\omega}_{\ell_2})$. Using this and combining the results of Steps 2 and 3 we show that $\|\hat{\omega}_{\ell_2}\|^2_{2m \sigma^2}$ concentrates with high probability around $\gamma$ (see Section 9.1.2 in [47]).
5.2 C-LASSO

5.2.1 NSE

**Theorem 1** ([47]). Assume there exists a constant \( \varepsilon_L > 0 \) such that, \( (1 - \varepsilon_L)m \geq D(\text{cone}(\partial f(x_0))) \geq \varepsilon_L m \) and \( m \) is sufficiently large. For any \( \varepsilon > 0 \), there exists a constant \( C = C(\varepsilon, \varepsilon_L) \) such that, with probability \( 1 - \exp(-Cm) \),

\[
\frac{\|\hat{x}_c - x_0\|^2}{m\sigma^2} \leq (1 + \varepsilon) \frac{D(\text{cone}(\partial f(x_0)))}{m - D(\text{cone}(\partial f(x_0)))},
\]

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

\[
\left| \frac{\|\hat{x}_c - x_0\|^2}{m\sigma^2} - \frac{m - D(\text{cone}(\partial f(x_0)))}{D(\text{cone}(\partial f(x_0)))} \right| < \varepsilon.
\]

Observe in Theorem 1 that as \( m \) approaches \( D(\text{cone}(\partial f(x_0))) \), the NSE increases and when \( m = D(\text{cone}(\partial f(x_0))) \), NSE = \( \infty \). This behavior is not surprising as when \( m < D(\text{cone}(\partial f(x_0))) \), one cannot even recover \( x_0 \) from noiseless observations via (14) hence it is futile to expect noise robustness.

**Example (sparse signals):** Figure 2 illustrates Theorem 1 when \( x_0 \) is a \( k \)-sparse vector and \( f(\cdot) \) is the \( \ell_1 \) norm. In this case, \( D(\text{cone}(\partial f(x_0))) \) is only a function of \( k \) and \( n \) and can be exactly calculated. [27] (also see Section 8). The dark-blue region corresponds to the unstable region \( m < D(\text{cone}(\partial f(x_0))) \). The dashed gray line obeys \( m = 1.4 \times D(\text{cone}(\partial f(x_0))) \) and yields a constant (worst-case) NSE of 2.5 as sparsity varies. We note that for \( \ell_1 \) minimization, the NSE formula was first proposed by Donoho et al. in [24].

5.2.2 Relation to Proximal Denoising

It is interesting to compare the NSE of the C-LASSO to the MSE risk of the constrained proximal denoiser in (16). Recall from Section 4.2.3 that the normalized MSE of (16) is upper bounded by \( D(\text{cone}(\partial f(x_0))) \) [13, 43]. Furthermore, this bound is attained asymptotically as \( \sigma \to 0 \). From Theorem 1 we find that the corresponding quantity \( \|\hat{x}_c - x_0\|^2/\sigma^2 \) is upper bounded by

\[
D(\text{cone}(\partial f(x_0))) \frac{m}{m - D(\text{cone}(\partial f(x_0)))},
\]

and is again attained asymptotically as \( \sigma \to 0 \). We conclude that the NSE of the LASSO problem is amplified compared to the corresponding quantity of proximal denoising by a factor of \( \frac{m}{m - D(\text{cone}(\partial f(x_0)))} > 1 \). This factor can be interpreted as the penalty paid in the estimation error for observing noisy linear measurements of the unknown signal instead of just noisy measurements of the signal itself.
Fig. 2: NSE heatmap for $\ell_1$ minimization based on Theorem 1. The $x$ and $y$ axes are the sparsity and measurements normalized by the ambient dimension. To obtain the figure, we plotted the heatmap of the function $-\log m^{-1}D(\partial f(x_0))$ (clipped to ensure the values are between $[-10, 5]$).

### 5.3 $\ell_2$-LASSO

Characterization of the NSE of the $\ell_2$-LASSO is more involved than that of the NSE of the C-LASSO. For this problem, choice of $\lambda$ naturally plays a critical role. We state the main result in Section 5.3.1 and discuss it in detail in Section 5.3.2-5.3.4.

#### 5.3.1 NSE

**Definition 5.1 ($\mathcal{R}_{ON}$)** Suppose $m > \min_{\lambda \geq 0} D(\lambda \partial f(x_0))$. Define $\mathcal{R}_{ON}$ as follows,

$$\mathcal{R}_{ON} = \{ \lambda > 0 \mid m - D(\lambda \partial f(x_0)) > \max\{0, C(\lambda \partial f(x_0))\} \}.$$ 

**Theorem 2 ([47])** Assume there exists a constant $\varepsilon_L > 0$ such that $(1 - \varepsilon_L)m \geq \max\{D(\lambda \partial f(x_0)), D(\lambda \partial f(x_0)) + C(\lambda \partial f(x_0))\}$ and $D(\lambda \partial f(x_0)) \geq \varepsilon_Lm$. Further, assume that $m$ is sufficiently large. Then, for any $\varepsilon > 0$, there exist a constant $C = C(\varepsilon, \varepsilon_L)$ and a deterministic number $\sigma_0 > 0$ (i.e. independent of $A, \mathbf{x}$) such that, whenever $\sigma \leq \sigma_0$, with probability $1 - \exp(-C \min\{m, \frac{\sigma^2}{m}\})$,

$$\frac{\|\hat{x}_L - x_0\|^2}{m\sigma^2} \times \frac{m - D(\lambda \partial f(x_0))}{D(\lambda \partial f(x_0))} < 1.$$
5.3.2 Regions Of Operation

First, we identify the regime in which the $\ell_2$-LASSO can robustly recover $x_0$. In this direction, the number of measurements should be large enough to guarantee at least noiseless recovery in (14), which is the case when $m > D(\text{cone}(\partial f(x_0)))$ [1, 14]. To translate this requirement in terms of noiseless recovery in (14), which is the case when $\lambda_{best}$ is optimal in the sense that the NSE is minimized for this particular choice of the penalty parameter (see Section 5.3.4). This also explains the term “best” we associate with it.

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

1. $\lambda_{best}$: $\lambda_{best}$ is optimal in the sense that the NSE is minimized for this particular choice of the penalty parameter (see Section 5.3.4). This also explains the term “best” we associate with it.

2. $\lambda_{\text{max}}$: Over $\lambda \geq \lambda_{best}$, the equation $m = D(\lambda \partial f(x_0))$ has a unique solution. We denote this solution by $\lambda_{\text{max}}$. For values of $\lambda$ larger than $\lambda_{\text{max}}$, we have $m \leq D(\lambda \partial f(x_0))$.

3. $\lambda_{\text{crit}}$: Over $0 \leq \lambda \leq \lambda_{\text{best}}$, if $m \leq n$, the equation $m - D(\lambda \partial f(x_0)) = C(\lambda \partial f(x_0))$ has a unique solution which we denote $\lambda_{\text{crit}}$. 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 $\ell_2$-LASSO, as follows,

1. $\mathcal{R}_{ON} = \{ \lambda \in \mathbb{R}^+ | \lambda_{\text{crit}} < \lambda < \lambda_{\text{max}} \}$.
2. $\mathcal{R}_{OFF} = \{ \lambda \in \mathbb{R}^+ | \lambda \leq \lambda_{\text{crit}} \}$.
3. $\mathcal{R}_{\infty} = \{ \lambda \in \mathbb{R}^+ | \lambda \geq \lambda_{\text{max}} \}$.

See Figure 5 for an illustration of the definitions above and Section 8 in [47] for the detailed proofs of the statements.

5.3.3 Characterizing the NSE in each Region

Theorem 2 upper bounds the NSE of the $\ell_2$-LASSO in $\mathcal{R}_{ON}$. Here, we also briefly discuss on some observations that can be made regarding $\mathcal{R}_{OFF}$ and $\mathcal{R}_{\infty}$:

- $\mathcal{R}_{OFF}$: For $\lambda \in \mathcal{R}_{OFF}$, extended empirical observations suggest that the LASSO estimate $\hat{x}_{\ell_1}$ satisfies $y = \lambda \hat{x}_{\ell_1}$ and the optimization (18) reduces to the standard $\ell_1$ minimization (14) of noiseless CS. Indeed, Lemma 9.2 in [47] proves that this reduction is indeed true for sufficiently small values of $\lambda$. Proving the validity of the claim for the entire region $\mathcal{R}_{OFF}$ would show that when $\sigma \to 0$, the NSE is $D(\lambda_{\text{crit}} \cdot \partial f(x_0))/(m - D(\lambda_{\text{crit}} \cdot \partial f(x_0)))$, for all $\lambda \in \mathcal{R}_{OFF}$.

- $\mathcal{R}_{ON}$: Begin with observing that $\mathcal{R}_{ON}$ is a nonempty and open interval. In particular, $\lambda_{\text{best}} \in \mathcal{R}_{ON}$ since $m > D(\lambda_{\text{best}} \cdot \partial f(x_0))$. Theorem 2 proves that for all $\lambda \in \mathcal{R}_{ON}$ and for $\sigma$ sufficiently small,

$$\frac{\|\hat{x}_{\ell_2} - x_0\|}{m\sigma^2} \approx \frac{D(\lambda \partial f(x_0))}{m - D(\lambda \partial f(x_0))},$$

(38)
Fig. 3: Regions of operation of the \( \ell_2 \)-LASSO.

Fig. 4: We consider the \( \ell_1 \)-penalized \( \ell_2 \)-LASSO problem for a \( k \) sparse signal in \( \mathbb{R}^n \). For \( \frac{k}{n} = 0.1 \) and \( \frac{m}{n} = 0.5 \), we have \( \lambda_{\text{crit}} \approx 0.76 \), \( \lambda_{\text{best}} \approx 1.14 \), \( \lambda_{\text{max}} \approx 1.97 \).

Also, empirical observations suggest that (38) holds for arbitrary \( \sigma \) when \( \approx \) replaced with \( \lesssim \). Finally, we should note that the NSE formula
\[
\frac{D(\lambda \partial f(x_0))}{m - D(\lambda \partial f(x_0))}
\]
is a convex function of \( \lambda \) over \( \mathcal{R}_{\text{ON}} \).

- \( \mathcal{R}_{\infty} \): Empirically, we observe that the stable recovery of \( x_0 \) is not possible for \( \lambda \in \mathcal{R}_{\infty} \).
5.3.4 Optimal Tuning of the Penalty Parameter

It is not hard to see that the formula in (38) is strictly increasing in $D(\lambda \partial f(x_0))$. Thus, when $\sigma \to 0$, the NSE achieves its minimum value when the penalty parameter is set to $\lambda_{\text{best}}$. Recall from (29) that $D(\lambda_{\text{best}} \cdot \partial f(x_0)) \approx D(\text{cone}(\partial f(x_0)))$ and compare the formulae in Theorems 1 and 2, to conclude that the C-LASSO and $\ell_2$-LASSO can be related by choosing $\lambda = \lambda_{\text{best}}$. In particular, we have,

$$\|\hat{x}_{\ell_2}(\lambda_{\text{best}}) - x_0\|_2^2 \approx \frac{D(\lambda_{\text{best}} \cdot \partial f(x_0))}{m} \approx \frac{D(\text{cone}(\partial f(x_0)))}{m} \approx \frac{\|\hat{x}_c - x_0\|_2^2}{m \sigma^2}.$$  

It is important to note that deriving $\lambda_{\text{best}}$ does not require knowledge of any properties (e.g. variance) of the noise vector neither does it require knowledge of the unknown signal $x_0$ itself. All it requires is knowledge of the particular structure of the unknown signal. For example, in the $\ell_1$-case, $\lambda_{\text{best}}$ depends only on the sparsity of $x_0$, not $x_0$ itself, and in the nuclear norm case, it only depends on the rank of $x_0$, not $x_0$ itself.

5.4 $\ell_2^2$-LASSO

5.4.1 Connection to $\ell_2$-LASSO

We propose a mapping between the penalty parameters $\lambda$ of the $\ell_2$-LASSO program (18) and $\tau$ of the $\ell_2^2$-LASSO program (19), for which the NSE of the two problems behaves the same. The mapping function is defined as follows.

**Definition 5.2 (Mapping Function)** For any $\lambda \in \mathcal{R}_{\text{ON}}$, define

$$\text{map}(\lambda) = \lambda \frac{m - D(\lambda \partial f(x_0)) - C(\lambda \partial f(x_0))}{\sqrt{m - D(\lambda \partial f(x_0))}}.$$  

Observe that $\text{map}(\lambda)$ is well-defined over the region $\mathcal{R}_{\text{ON}}$, since $m > D(\lambda \partial f(x_0))$ and $m - D(\lambda \partial f(x_0)) > C(\lambda \partial f(x_0))$ for all $\lambda \in \mathcal{R}_{\text{ON}}$. It can be proven that $\text{map}(\cdot)$ defines a bijective mapping from $\mathcal{R}_{\text{ON}}$ to $\mathbb{R}^+$ [47, Theorem 3.3].

**Theorem 3 (Properties of map(·)).** Assume $m > \min_{\lambda \geq 0} D(\lambda \partial f(x_0))$. The function $\text{map}(\cdot) : \mathcal{R}_{\text{ON}} \to \mathbb{R}^+$ is strictly increasing and continuous. Thus, its inverse function $\text{map}^{-1}(\cdot) : \mathbb{R}^+ \to \mathcal{R}_{\text{ON}}$ is well defined.

Some other useful properties of the mapping function include the following:

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

28 Christos Thrampoulidis, Samet Oymak and Babak Hassibi
5.4.2 Proposed Formula

The mapping function can potentially be used to translate results on the NSE of the $\ell_2$-LASSO over $\mathcal{R}_{ON}$ to corresponding results on the $\ell_2^2$-LASSO for $\tau \in \mathbb{R}^+$. Assume $m > D(\lambda_{\text{best}} \cdot \partial f(x_0))$. It is conjectured in [47] that for any $\tau > 0$,

$$\frac{D(\text{map}^{-1}(\tau) \cdot \partial f(x_0))}{m - D(\text{map}^{-1}(\tau) \cdot \partial f(x_0))}$$

accurately characterizes the NSE $||\tilde{x}_{\ell_2^2} - x_0||^2/(m\sigma^2)$ for sufficiently small $\sigma$, and upper bounds it for arbitrary $\sigma$. Extended numerical simulations (see Section 13 in [47]) support the conjecture.

Fig. 5: Consider the exact same setup of Figure 4. We plot the NSE of the $\ell_2^2$-LASSO as a function of the regularizer parameter.

This formula would suggest a simple recipe for computing the optimal value of the penalty parameter, which we call $\tau_{\text{best}}$. Recall that $\lambda_{\text{best}}$ minimizes the error in the $\ell_2$-LASSO. Then, the proposed mapping between the two problems, suggests that $\tau_{\text{best}} = \text{map}(\lambda_{\text{best}})$. To evaluate $\text{map}(\lambda_{\text{best}})$ we make use of Lemma 1 and the fact that $\frac{dD(\lambda \partial f(x_0))}{d\lambda} = -\frac{2}{\lambda} C(\lambda \partial f(x_0))$ for all $\lambda \geq 0$. Combine this with the fact that $\lambda_{\text{best}}$ is the unique minimizer of $D(\lambda \partial f(x_0))$, to show that $C(\lambda_{\text{best}} \cdot \partial f(x_0)) = 0$, and to conclude with,

$$\tau_{\text{best}} = \lambda_{\text{best}} \sqrt{m - D(\lambda_{\text{best}} \cdot \partial f(x_0))} \approx \lambda_{\text{best}} \sqrt{m - D(\text{cone}(\partial f(x_0)))}.$$
6 The NSE of Generalized LASSO with Arbitrary Fixed Noise

Here, we relax the assumption of Section 5 that the entries of $z$ are i.i.d. normal. Instead, assume that the noise vector $z$ is arbitrary, but still independent of the sensing matrix $A$. Under this assumption, we derive simple and non-asymptotic upper bounds on the NSE of the C-LASSO and of the $\ell_2$-LASSO. Those upper bounds can be interpreted as generalizations of the bound on the error of the OLS as was discussed in Section 2.2. Compared to the bounds of Section 5, the bounds derived here, not only hold under more general assumption on the noise vector, but they are also non-asymptotic.

6.1 C-LASSO

Recall the generalized C-LASSO in (6). Section 6.1.1 introduces an upper bound on its NSE for arbitrary fixed noise vector that is independent of $A$. In Section 6.1.2 we compare this bound to the result of Theorem 1 and Section 6.1.3 provides an overview of the proof technique.

6.1.1 NSE

Theorem 4. [46] Assume $m \geq 2$ and $0 < t \leq \sqrt{m - 1} - \sqrt{D(\text{cone}(\partial f(x_0)))}$. Then, with probability, $1 - 6\exp(-t^2/26)$,

$$\frac{\|\hat{x}_c - x_0\|}{\|z\|} \leq \frac{\sqrt{m}}{\sqrt{m - 1}} \frac{\sqrt{D(\text{cone}(\partial f(x_0)))} + t}{\sqrt{m - 1} - \sqrt{D(\text{cone}(\partial f(x_0)))} - t}.$$

6.1.2 Comparison to Theorem 1

It is interesting to see how the bound of Theorem 4 compares to the result of Theorem 1 in the case $z \sim \mathcal{N}(0, \sigma^2 I_m)$. Of course, when this is the case the bound of Theorem 1 is tight and our intention is to see how loose is the bound of Theorem 4. Essentially, the only difference appears in the denominators of the two bounds: $\sqrt{m - D(\text{cone}(\partial f(x_0)))} \geq \sqrt{m} - \sqrt{D(\text{cone}(\partial f(x_0)))}$ for all regimes of $0 \leq D(\text{cone}(\partial f(x_0))) < m$. The contrast becomes significant when $m \approx D(\text{cone}(\partial f(x_0)))$. In particular, setting $m = (1 + \varepsilon)^2 D(\text{cone}(\partial f(x_0)))$, we have,

$$\frac{\sqrt{m - D(\text{cone}(\partial f(x_0)))}}{\sqrt{m} - D(\text{cone}(\partial f(x_0)))} = \frac{\sqrt{2\varepsilon + \varepsilon^2}}{\varepsilon} = \sqrt{\frac{2}{\varepsilon} + 1}.$$

12 Precisely: assuming $m \approx m - 1$ and ignoring the $t$'s in the bound of Theorem 4.
Thus, when \( \varepsilon \) is large, the bound of Theorem 4 is arbitrarily tight. On the other hand, when \( \varepsilon \) is small, it can be arbitrarily worse. Simulation results (see Figure 6) verify that the error bound of Theorem 4 becomes sharp as the number of measurements \( m \) increases. Besides, even if tighter, the bound of Theorem 1 requires stronger assumptions namely, an i.i.d. Gaussian noise vector \( z \) and an asymptotic setting where \( m \) and \( D(\text{cone}(\partial f(x_0))) \) are large enough.

![Fig. 6: NSE of the C-LASSO with \( \ell_1 \). The dimension of the \( x_0 \) is \( n = 500 \) and its sparsity is \( k = 5 \). The number of measurements \( m \) varies from 0 to 360. We plot the empirical NSE assuming \( z \sim \mathcal{N}(0, \sigma^2 I) \) for several values of \( \sigma \). The solid black line corresponds to the bound of Theorem 4. The dashed line corresponds to the phase transition line of noiseless CS \( D(\text{cone}(\partial f(x_0))) \).](image)

**6.1.3 Proof Overview**

We only provide an overview of the proof. The details can be found in [46]. We begin with introducing some useful notation. \( A \mathcal{T}_f(x_0) \) will denote the cone obtained by multiplying elements of \( \mathcal{T}_f(x_0) \) by \( A \), i.e.,

\[
A \mathcal{T}_f(x_0) = \{ Av \in \mathbb{R}^m \mid v \in \mathcal{T}_f(x_0) \}.
\]

The lemma below derives a deterministic upper bound on the squared error of the C-LASSO. It is interesting to compare this to the corresponding bound (10) for the OLS. Recall the notions of “tangent cone” and “restricted minimum singular value” introduced in Section 4.

**Lemma 9 (Deterministic error bound).**
$$\|\hat{x}_c - x_0\| \leq \frac{\|\text{Proj}(z, A T f(x_0))\|}{\sigma_{\text{min}}(A, T f(x_0) \cap S^{n-1})}.$$ 

**Proof.** From first-order optimality conditions (e.g. [52, p. 270-271]),

$$\langle A^T (A \hat{x}_c - y), \hat{x}_c - x_0 \rangle \leq 0.$$ 

Writing $y = Ax_0 + z$, and rearranging terms, we find that,

$$\|A(\hat{x}_c - x_0)\| \leq \sup_{v \in A T f(x_0) \cap S^{n-1}} \langle z, A(\hat{x}_c - x_0) \rangle \leq 0. \tag{39}$$

(39) follows since $\hat{x}_c - x_0 \in T f(x_0)$. For (40), we applied Moreau’s decomposition Theorem [52, Theorem 31.5]. To conclude with the desired result it remains to invoke the definition of the restricted singular values $\sigma_{\text{min}}(A, T f(x_0) \cap S^{n-1})$.

To prove Theorem 4 we will translate the deterministic bound of Lemma 9 to a probabilistic one. For this, we need a high-probability lower bound for $\sigma_{\text{min}}(A, T f(x_0) \cap S^{n-1})$ and a high-probability upper bound for $\|\text{Proj}(z, A T f(x_0))\|$. The former is given by a direct application of the “escape through a mesh” Lemma 4. The latter requires some more effort to derive. The result is summarized in Lemma 10 below and is the main technical contribution of [46]. The proof makes use of Gordon’s Lemma 3 and can be found in [46].

**Lemma 10 (Restricted correlation).** Let $\mathcal{K} \in \mathbb{R}^n$ be a convex and closed cone, $G \in \mathbb{R}^{m \times n}$ have independent standard normal entries, $m \geq 2$ and $z \in \mathbb{R}^m$ be arbitrary and independent of $G$. For any $t > 0$, pick $\alpha \geq \sqrt{D(\mathcal{K}) + t \gamma_{m-1}} \|z\|$. Then,

$$\sup_{v \in \mathcal{K} \cap S^{n-1}} \{z^T Gv - \alpha \|Gv\|\} \leq 0, \tag{41}$$

with probability $1 - 5 \exp\left(-\frac{t^2}{2\gamma}ight)$.

We may now complete the proof of Theorem 4.

**Proof (of Theorem 4).** Suppose $0 \leq t < \gamma_m - D(\text{cone}(\partial f(x_0)))$. First apply Lemma 4 with $G = \sqrt{m} A$ and $\mathcal{K} = T f(x_0)$. Then, with probability $1 - \exp\left(-\frac{t^2}{2\gamma}\right)$,

$$\sigma_{\text{min}}(A, T f(x_0) \cap S^{n-1}) \geq \frac{\gamma_m - \sqrt{D(\text{cone}(\partial f(x_0)))}}{\sqrt{m}} - t. \tag{42}$$
Next, apply Lemma 10 with $G = \sqrt{m} A$ and $\mathcal{K} = \mathcal{F}(x_0)$. With probability $1 - 5\exp(-t^2/32),$

$$\|\text{Proj}(z, A\mathcal{F}(x_0))\| = z^T \text{Proj}(z, A\mathcal{F}(x_0)) \leq \sup_{v \in \mathcal{F}(x_0) \cap S_{n-1}} \frac{z^T A v}{\sqrt{D(\lambda \partial f(x_0)) + t}} \leq \frac{\sqrt{D(\lambda \partial f(x_0)) + t}}{\sqrt{m - 1}} \|z\|. \quad (43)$$

Theorem 4, now follows after substituting (42) and (43) in Lemma 9 and using the following: $\gamma_m \gamma_{m-1} = m - 1$ and $\gamma_{m-1} \leq m - 1$.

### 6.2 $\ell_2$-LASSO

Recall the generalized $\ell_2$-LASSO in (7). Section 6.2.1 derives an upper bound on its NSE for arbitrary fixed noise vector that is independent of $A$. In Section 6.2.2 we compare this bound to the result of Theorem 2 and Section 6.2.3 provides an overview of the proof.

#### 6.2.1 NSE

**Theorem 5 ([60]).** Assume $m \geq 2$. Fix the regularizer parameter in (7) to be $\lambda \geq 0$ and let $\hat{x}_\ell^2$ be a minimizer of (7). Then, for any $0 < t \leq (\sqrt{m - 1} - \sqrt{D(\lambda \partial f(x_0))}),$ with probability $1 - 5\exp(-t^2/32),$

$$\|\hat{x} - x_0\| \leq 2\|z\| \frac{\sqrt{D(\lambda \partial f(x_0)) + t}}{\sqrt{m - 1} - \sqrt{D(\lambda \partial f(x_0))} - t}.$$  

Theorem 5 provides a simple, general, non-asymptotic and (rather) sharp upper bound on the error of the regularized lasso estimator (7), which also takes into account the specific choice of the regularizer parameter $\lambda \geq 0$. It is non-asymptotic and is applicable in any regime of $m$, $\lambda$ and $D(\lambda \partial f(x_0))$. Also, the constants involved in it are small making it rather tight\(^{13}\).

For the bound of Theorem 5 to be at all meaningful, we require $m > \min_{\lambda \geq 0} D(\lambda \partial f(x_0)) = D(\lambda_{\text{best}} \partial f(x_0)).$ Recall that this translates to the number of measurements being large enough to at least guarantee noiseless recovery. Also, similar to the discussion in Section 5.3.2 there exists\(^{14}\) a unique $\lambda_{\text{max}}$ satisfying $\lambda_{\text{max}} > \lambda_{\text{best}}$ and $\sqrt{D(\lambda_{\text{max}} \partial f(x_0))} = \sqrt{m - 1}$, and, when $m \leq n$, there exists

\(^{13}\) It is conjectured in [60] and supported by simulations (e.g. Figure 7) that the factor of 2 in Theorem 5 is an artifact of the proof technique and not essential.

\(^{14}\) For proofs of those claims, see Section 8 and in particular Lemma 8.1 in [47].
unique $\lambda_{\min} < \lambda_{\text{best}}$ satisfying $\sqrt{D(\lambda_{\min} \partial f(x_0)))} = \sqrt{m - 1}$. From this, it follows that $\sqrt{m - 1} > \sqrt{D(\lambda \partial f(x_0)))}$ if and only if $\lambda \in (\lambda_{\min}, \lambda_{\max})$. This is a superset of $\mathcal{R}_{\ON}$ (recall the definition in Section 5.3.2) and is exactly the range of values of the regularizer parameter $\lambda$ for which the bound of Theorem 5 is meaningful.

As a superset of $\mathcal{R}_{\ON}$, $(\lambda_{\min}, \lambda_{\max})$ contains $\lambda_{\text{best}}$, for which, the bound of Theorem 5 achieves its minimum value since it is strictly increasing in $D(\lambda \partial f(x_0))$.

Recall from Section 5.3.4 that deriving $\lambda_{\text{best}}$ does not require knowledge of any properties (e.g. variance) of the noise vector neither does it require knowledge of the unknown signal $x_0$ itself.

As a final remark, comparing Theorem 5 to Theorem 4 reveals the similar nature of the two results. Apart from a factor of 2, the upper bound on the error of the regularized lasso (7) for fixed $\lambda$, is essentially the same as the upper bound on the error of the constrained lasso (6), with $D(\lambda \partial f(x_0))$ replaced by $D(\lambda \partial f(x_0))$. This when combined with (29) suggests that setting $\lambda = \lambda_{\text{best}}$ in (7) achieves performance almost as good as that of (6).

6.2.2 Comparison to Theorem 2

To start with, Theorem 5 is advantageous to Theorem 2 in that it holds in more general setting than standard Gaussian noise and, also, characterizes a superset of $\mathcal{R}_{\ON}$. Furthermore, it is non-asymptotic, while Theorem 2 requires $m, D(\lambda \partial f(x_0))$ to be large enough. On the other side, when $z \sim \mathcal{N}(0, \sigma^2 I_m)$, then, Theorem 2 offers clearly a tighter bound on the NSE. Yet, apart from a factor of 2, this bound only differs from the bound of Theorem 5 in the denominator, where instead of $\sqrt{m - 1 - \sqrt{D(\lambda \partial f(x_0))}}$ we have the larger quantity $\sqrt{m - D(\lambda \partial f(x_0))}$. This difference becomes insignificant and indicates that our bound is rather tight when $m$ is large. Finally, the bound of Theorem 2 is only conjectured in [47] to upper bound the estimation error for arbitrary values of the noise variance $\sigma^2$. In contrast, Theorem 5 is a fully rigorous upper bound on the estimation error of (7).

6.2.3 Proof Overview

It is convenient to rewrite the generalized $\ell_2$-LASSO in terms of the error vector $w = x - x_0$ as follows:

$$
\min_w \|Aw - z\| + \frac{\lambda}{\sqrt{m}} (f(x_0 + w) - f(x_0)).
$$

Denote the solution of (44) by $\hat{w}$. Then, $\hat{w} = \hat{x} - x_0$ and we want to bound $\|\hat{w}\|$. To simplify notation, for the rest of the proof, we denote the value of that desired upper bound as
Recovering Structured Signals in Noise: Precise Error Bounds

Fig. 7: Figure 7 illustrates the bound of Theorem 5, which is given in red for $n = 340, m = 140$, $k = 10$ and for $A$ having $\mathcal{N}(0, \frac{1}{m})$ entries. The upper bound of Theorem 2, which is asymptotic in $m$ and only applies to i.i.d. Gaussian $z$, is given in black. In our simulations, we assume $x_0$ is a random unit norm vector over its support and consider both i.i.d. $\mathcal{N}(0, \sigma^2)$, as well as, non-Gaussian noise vectors $z$. We have plotted the realizations of the normalized error for different values of $\lambda$ and $\sigma$. As noted, the bound of Theorem 2 is occasionally violated since it requires very large $m$, as well as, i.i.d. Gaussian noise. On the other hand, the bound of Theorem 5 always holds.

$\ell(t) := 2\|z\| \frac{\sqrt{D(\lambda \partial f(x_0))} + t}{\sqrt{m-1} - \sqrt{D(\lambda \partial f(x_0))} - t}$. \hspace{1cm} (45)

It is easy to see that the optimal value of the minimization in (44) is no greater than $\|z\|$. Observe that $w = 0$ achieves this value. However, Lemma 11 below shows that if we constrain the minimization in (44) to be only over vectors $w$ whose norm is greater than $\ell(t)$, then the resulting optimal value is (with high probability on the measurement matrix $A$) strictly greater than $\|z\|$. Combining those facts yields the desired result, namely $\|\hat{w}\| \leq \ell(t)$. The fundamental technical tool in the proof of Lemma 11 is (not surprisingly at this point) Gordon’s Lemma 3.

Lemma 11. Fix some $\lambda \geq 0$ and $0 < t \leq (\sqrt{m-1} - \sqrt{D(\lambda \partial f(x_0))})$. Let $\ell(t)$ be defined as in (45). Then, with probability $1 - 5\exp(-t^2/32)$, we have,

$$\min_{\|w\| \geq \ell(t)} \{ \|Aw - z\| + \lambda \sqrt{m} (f(x_0 + w) - f(x_0)) \} > \|z\|. \hspace{1cm} (46)$$

Proof. Fix $\lambda$ and $t$, as in the statement of the lemma. From the convexity of $f(\cdot)$, $f(x_0 + w) - f(x_0) \geq \max_{s \in \partial f(x_0)} s^T w$. Hence, it suffices to prove that w.h.p. over $A$, 
\[
\min_{\|w\| \geq \ell(t)} \{\sqrt{m}\|Aw - z\| + \max_{s \in \lambda \partial_f(x_0)} s^T w\} > \sqrt{m}\|z\|.
\]

We begin with applying Gordon’s Lemma 3 to the optimization problem in the expression above. Define, \(z = \sqrt{m}z\), rewrite \(\|Aw - z\|\) as \(\max_{|a| = 1} \{a^T Aw - a^T z\}\) and, then, apply Lemma 3 with \(G = \sqrt{m}A\). \(\mathcal{S} = \{w \mid \|w\| \geq \ell(t)\}\) and \(\psi(w, a) = -a^T z + \max_{s \in \lambda \partial_f(x_0)} s^T w\). This leads to the following statement:

\[
P(46) \text{ is true } \geq 2 \cdot P(\mathcal{L}(t; g, h) > \|z\|) - 1,
\]

where, \(\mathcal{L}(t; g, h)\) is defined as

\[
\min_{\|w\| \geq \ell(t)} \max_{|a| = 1} \{\|w\|g^{T} a - \min_{s \in \lambda \partial_f(x_0)} (h - s)^T w\}. \tag{47}
\]

In the remaining, we analyze the simpler optimization problem defined in (47), and prove that \(\mathcal{L}(t; g, h) > \|z\|\) holds with probability \(1 - \frac{3}{2} \exp(-t^2/32)\). We begin with simplifying the expression for \(\mathcal{L}(t; g, h)\), as follows:

\[
\mathcal{L}(t; g, h) = \min_{\|w\| \geq \ell(t)} \{\|w\|g - z\} - \min_{s \in \lambda \partial_f(x_0)} (h - s)^T w
\geq \min_{\alpha \geq \ell(t)} \{\|\alpha g - z\| - \alpha \text{dist}(h, \lambda \partial f(x_0))\}
= \min_{\alpha \geq \ell(t)} \left\{ \alpha^2 \|g\|^2 + \|z\|^2 - 2\alpha g^T z - \alpha \text{dist}(h, \lambda \partial f(x_0)) \right\}.
\]

The first equality above follows after performing the trivial maximization over \(a\) in (47). For the inequality that follows, we apply the minimax inequality [52, Lemma 36.1] and use the second equality in (35). Next, we show that \(\mathcal{L}(t; g, h)\) is strictly greater than \(\|z\|\) with the desired high probability over realizations of \(g\) and \(h\). Consider the event \(\mathcal{E}_i\) of \(g\) and \(h\) satisfying all three conditions listed below,

\[
1. \|g\| \geq \gamma_m - t/4, \tag{48a}
2. \text{dist}(h, \lambda \partial f(x_0)) \leq \sqrt{D(\lambda \partial f(x_0))} + t/4, \tag{48b}
3. g^T z \leq (t/4) \|z\|. \tag{48c}
\]

The conditions in (48) hold with high probability. In particular, the first two hold with probability no less than \(1 - \exp(-t^2/32)\). This is because the \(\ell_2\)-norm and the distance function to a convex set are both 1-Lipschitz functions and, thus, Lemma 5 applies. The third condition holds with probability at least \(1 - (1/2) \exp(-t^2/32)\), since \(g^T z\) is statistically identical to \(\mathcal{N}(0, \|z\|^2)\). Union bounding yields,

\[
P(\mathcal{E}_i) \geq 1 - (5/2) \exp(-t^2/32). \tag{49}
\]

Furthermore, it can be shown (see Lemma 4.2 in [60]) that if \(g\) and \(h\) are such that \(\mathcal{E}_i\) is satisfied, then \(\mathcal{L}(t; g, h) > \|z\|\). This, when combined with (49) shows that \(P(\mathcal{L}(t; g, h) > \|z\|) \geq 1 - (5/2) \exp(-t^2/32)\), completing the proof of Lemma 11.
7 The Worst-Case NSE of Generalized LASSO

Here, we assume no restriction at all on the distribution of the noise vector $z$. In particular, this includes the case of adversarial noise, i.e., noise that has information on $A$ and can adapt itself accordingly. We compute the resulting worst-case NSE of the C-LASSO in the next section.

7.1 C-LASSO

7.1.1 NSE

**Theorem 6.** Assume $0 < t \leq \sqrt{m} - \sqrt{\mathcal{D}(\lambda \partial f(x_0))}$. Then, with probability $1 - \exp(-t^2/2)$,

$$\frac{\|\hat{x}_c - x_0\|}{\|z\|} \leq \frac{\sqrt{m}}{\gamma_m - \sqrt{\mathcal{D}(\text{cone}(\partial f(x_0)))}} - t.$$

Recall in the statement of Theorem 6 that $\gamma_m = E[\|g\|]$, with $g \sim \mathcal{N}(0, I_m)$. For large $m$, $\gamma_m \approx \sqrt{m}$ and according to Theorem 6 the worst-case NSE of the C-LASSO can be as large as 1. Contrast this to Theorem 4 and the case where $z$ is not allowed to depend on $A$. There, the NSE is approximately $\mathcal{D}(\text{cone}(\partial f(x_0))) / m$ for large $m$.

7.1.2 Proof

The proof of Theorem 6 follows easily from Lemma 9:

$$\|\hat{x}_c - x_0\| \leq \frac{\|A(\hat{x}_c - x_0)\|}{\sigma_{\min}(A, \mathcal{g})} \leq \frac{\|z\|}{\sigma_{\min}(A, \mathcal{g})} \leq \frac{\sqrt{m}}{\gamma_m - \sqrt{\mathcal{D}(\text{cone}(\partial f(x_0)))}} - t.$$

Apply (42) to the above, to conclude with the desired upper bound.

8 Conclusion

Precise undersampling theorems were first studied by Donoho and Tanner [25, 26]. They characterized the phase transitions for $\ell_1$ minimization in the noiseless setup. Since then, there have been major improvements in several directions. The theory was initially limited to sparse signal recovery and noiseless setup. With recent advances, we can precisely analyze arbitrary convex functions under noise.

We have presented a detailed analysis of the normalized squared error of the generalized LASSO under the assumption that the entries of the sensing matrix be-
ing i.i.d. normal. The derived formulae are precise and simple. They only involve two geometric measures of signal complexity, the Gaussian squared distance to the scaled subdifferential and to the cone of subdifferential, which also appear in the most recent literature of noiseless compressed sensing [1, 14, 55, 57]. Moreover, they admit insightful interpretations when seen as generalizations of the corresponding formulae for the classical ordinary least-squares estimation.

There are several remaining directions for future work. Some of these involve improving and adding on the Theorems presented here. For instance, the multiplicative factor of 2 in Theorem 5 seems to be unnecessary, urging for an improvement. Similar in nature are the problems of establishing the conjecture of [47] on the NSE of the $\ell_2$-LASSO (see Section 5.4.2) and studying the worst-case NSE of the $\ell_2$-LASSO. A potentially more challenging direction for future work involves studying different measurement ensembles and the universality phenomenon. It is now widely accepted that, whether the sensing matrix is i.i.d. Gaussian or i.i.d. Bernoulli, this does not change the phase transition and stability characteristics of the linear inverse problems [16, 28]. A provable extension of the results of this chapter to other measurement ensembles would, thus, be of great interest.

Appendix

The upper bounds on the NSE of the generalized LASSO presented in Sections 5-7 are in terms of the summary parameters $D(\lambda \partial f(x_0))$ and $D(\text{cone}(\partial f(x_0)))$. While the bounds are simple, concise and nicely resemble the corresponding ones in the case of OLS, it may appear to the reader that the formulae are rather abstract, because of the presence of $D(\text{cone}(\partial f(x_0)))$ and $D(\lambda \partial f(x_0))$.

However, as discussed here, for a large number of widely-used convex regularizers $f(\cdot)$, one can calculate (tight) upper bounds or even explicit formulae for these quantities. For example, for the estimation of a $k$-sparse signal $x_0$ with $f(\cdot) = \|\cdot\|_1$, it has been shown that $D(\text{cone}(\partial f(x_0))) \lesssim 2k(\log \frac{n}{k} + 1)$. Substituting this into Theorems 1 and 4 results in the “closed-form” upper bounds given in (20) and (22), i.e. ones expressed only in terms of $m$, $n$ and $k$. Analogous results have been derived [14, 31, 45, 54] for other well-known signal models as well, including low rankness and block-sparsity. The first column of Table 1 summarizes some of the results for $D(\text{cone}(\partial f(x_0)))$ found in the literature [14, 31]. The second column provides closed form results on $D(\lambda \partial f(x_0))$ when $\lambda$ is sufficiently large [47]. Note that, by setting $\lambda$ to its lower bound in the second row, one approximately obtains the corresponding result in the first row. This should not be surprising due to (29). Also, this

\footnote{We say $x_0 \in \mathbb{R}^n$ is block-sparse if it can be grouped into $t$ known blocks of size $b = n/t$ each 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 $\ell_2$ norms of the blocks, [29, 48, 54, 58]. In particular, denoting the subvector corresponding to $i$th block of a vector $x$ by $x_i$, the $\ell_{1,2}$ norm is defined as $\|x\|_{1,2} = \sum_{i=1}^t \|x_i\|_2$.}
value of $\lambda$ is a good proxy for the optimal regularizer $\lambda_{\text{best}}$ of the $\ell_2$-LASSO as was discussed in Sections 5.3.4 and 6.2.1.

Table 1: Closed form upper bounds for $D(\text{cone}(\partial f(x_0)))$ and $D(\lambda \partial f(x_0))$.

<table>
<thead>
<tr>
<th>Case</th>
<th>$D(\text{cone}(\partial f(x_0)))$</th>
<th>$D(\lambda \partial f(x_0))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$-sparse, $x_0 \in \mathbb{R}^n$</td>
<td>$2k(\log \frac{n}{\lambda} + 1)$</td>
<td>$(\lambda^2 + 3)k$ for $\lambda \geq \sqrt{2\log \frac{n}{\lambda}}$</td>
</tr>
<tr>
<td>$k$-block sparse, $x_0 \in \mathbb{R}^{n^k}$</td>
<td>$6\sqrt{nr}$</td>
<td>$\lambda^2 r + 2\sqrt{\pi(r+1)}$ for $\lambda \geq 2n^{1/4}$</td>
</tr>
<tr>
<td>Rank $r$, $x_0 \in \mathbb{R}^{\sqrt{\pi r}, \pi}$</td>
<td>$4k(\log \frac{r}{4} + b)$</td>
<td>$(\lambda^2 + b + 2)k$ for $\lambda \geq \sqrt{5} + \sqrt{2\log \frac{n}{\lambda}}$</td>
</tr>
</tbody>
</table>

We refer the reader to [1, 14, 31, 47] for the details and state-of-the-art bounds on $D(\text{cone}(\partial f(x_0)))$ and $D(\lambda \partial f(x_0))$. Identifying the subdifferential $\partial f(x_0)$ and calculating $D(\lambda \partial f(x_0))$ for all $\lambda \geq 0$, are the critical steps. Once those are available, computing $\min_{\lambda \geq 0} D(\lambda \partial f(x_0))$ provides upper approximation formulae for $D(\text{cone}(\partial f(x_0)))$. This idea was first introduced by Stojnic [55] and was subsequently refined and generalized in [14]. Most recently [1, 31] proved (29), thus showing that the resulting approximation on $D(\text{cone}(\partial f(x_0)))$ is in fact highly-accurate. Section 4 of [1] is an excellent reference for further details and the notation used there is closer to ours.

We should emphasize that examples of regularizers are not limited to the ones discussed here and presented in Table 1. There are increasingly more signal classes that exhibit low-dimensionality and to which the theorems of Sections 5-7 would apply. Some of these are as follows.

- Non-negativity constraint: $x_0$ has non-negative entries, [26].
- Low-rank plus sparse matrices: $x_0$ can be represented as sum of a low-rank and a sparse matrix, [65].
- Signals with sparse gradient: Rather than $x_0$ itself, its gradient $d_{x_0}(i) = x_0(i) - x_0(i-1)$ is sparse, [8].
- Low-rank tensors: $x_0$ is a tensor and its unfoldings are low-rank matrices, [32, 36].
- Simultaneously sparse and low-rank matrices: For instance, $x_0 = ss^T$ for a sparse vector $s$, [44, 51].

Establishing new and tighter analytic bounds for $D(\lambda \partial f(x_0))$ and $D(\text{cone}(\partial f(x_0)))$ for more regularizers $f$ is certainly an interesting direction for future research. In the case where such analytic bounds do not already exist in literature or are hard to derive, one can numerically estimate $D(\lambda \partial f(x_0))$ and $D(\text{cone}(\partial f(x_0)))$ once there is an available characterization of the subdifferential $\partial f(x_0)$. Using the concentration property of $\text{dist}^2(h, \lambda \partial f(x_0))$ around $D(\lambda \partial f(x_0))$, when $h \sim \mathcal{N}(0, I_n)$, we can compute $D(\lambda \partial f(x_0))$, as follows:

1. draw a vector $h \sim \mathcal{N}(0, I_n)$,
2. return the solution of the convex program \( \min_{s \in \partial f(x_0)} \| h - \lambda s \|^2 \).

Computing \( D(\text{cone}(\partial f(x_0))) \) can be built on the same recipe by recognizing
\( \text{dist}^2(h, \text{cone}(\partial f(x_0))) \) as \( \min_{\lambda \geq 0, s \in \partial f(x_0)} \| h - \lambda s \|^2 \).

To sum up, any bound on \( D(\lambda \partial f(x_0)) \) and \( D(\text{cone}(\partial f(x_0))) \) translates, through
Theorems 1–6, into corresponding upper bounds on the NSE of the generalized LASSO. For purposes of illustration and completeness, we review next the details of computing \( D(\lambda \partial f(x_0)) \) and \( D(\text{cone}(\partial f(x_0))) \) for the celebrated case where \( x_0 \) is sparse and the \( \ell_1 \)-norm is used as the regularizer.

**Sparse signals**

Suppose \( x_0 \) is a \( k \)-sparse signal and \( f(\cdot) = \| \cdot \|_1 \). Denote by \( S \) the support set of \( x_0 \), and by \( S^c \) its complement. The subdifferential at \( x_0 \) is \([52]\),
\[
\partial f(x_0) = \{ s \in \mathbb{R}^n \mid \| s \|_\infty \leq 1 \text{ and } s_i = \text{sign}(x_0)_i, \forall i \in S \}.
\]

Let \( h \in \mathbb{R}^n \) have i.i.d. \( \mathcal{N}(0, 1) \) entries and define
\[
\text{shrink}(\chi, \lambda) = \begin{cases} 
\chi - \lambda, & \chi > \lambda, \\
0, & -\lambda \leq \chi \leq \lambda, \\
\chi + \lambda, & \chi < -\lambda.
\end{cases}
\]

Then, \( D(\lambda \partial f(x_0)) \) is equal to \([1, 14]\)
\[
D(\lambda \partial f(x_0)) = \mathbb{E}[\text{dist}^2(h, \lambda \partial f(x_0))] = \sum_{i \in S} \mathbb{E}[(h_i - \lambda \text{sign}((x_0)_i))^2] + \sum_{i \in S^c} \mathbb{E}[\text{shrink}^2(h_i, \lambda)] = k(1 + \lambda^2) + (n - k) \sqrt{2 \pi} \left[ (1 + \lambda^2) \int_{-\lambda}^{\lambda} e^{-t^2/2} \, dt - \lambda \exp(-\lambda^2/2) \right].
\]

Note that \( D(\lambda \partial f(x_0)) \) depends only on \( n, \lambda \) and \( k = |S| \), and not explicitly on \( S \) itself (which is not known). Substituting the expression in \((50)\) in place of the \( D(\lambda \partial f(x_0)) \) in Theorems 2 and 5, yields explicit expressions for the corresponding upper bounds in terms of \( n, m, k \text{ and } \lambda \).

We can obtain an even simpler upper bound on \( D(\lambda \partial f(x_0)) \) which does not involve error functions as we show below. Denote \( Q(t) = \frac{1}{\sqrt{2 \pi}} \int_t^{\infty} e^{-\tau^2/2} \, d\tau \) the complementary c.d.f. of a standard normal random variable. Then,
\[
\frac{1}{2} \mathbb{E}[\text{shrink}^2(h, \lambda)] = \int_{\lambda}^{\infty} (t - \lambda)^2 d(-Q(t)) \\
= -\left[(t - \lambda)^2 Q(t)\right]_{\lambda}^{\infty} + 2 \int_{\lambda}^{\infty} (t - \lambda) Q(t) dt \\
\leq \int_{\lambda}^{\infty} (t - \lambda) e^{-r^2/2} dt \\
\leq e^{-\lambda^2/2} - \frac{\lambda^2}{\lambda^2 + 1} e^{-\lambda^2/2} \\
= \frac{1}{\lambda^2 + 1} e^{-\lambda^2/2}.
\]

(51) and (52) follow from standard upper and lower tail bounds on normal random variables, namely \(\frac{1}{\sqrt{2\pi}} \frac{1}{t^{1/2}} e^{-t^2/2} \leq Q(t) \leq \frac{1}{2} e^{-t^2/2}\). From this, we find that

\[
D(\lambda \partial f(x_0)) \leq k(1 + \lambda^2) + (n - k) \frac{2}{\lambda^2 + 1} e^{-\lambda^2/2}.
\]

Letting \(\lambda \geq \sqrt{2\log(n/k)}\) in the above expression, recovers the corresponding entry in Table 1:

\[
D(\lambda \partial f(x_0)) \leq (\lambda^2 \alpha) k, \text{ when } \lambda \geq \sqrt{2\log(n/k)}.
\]

Substituting (53) in Theorems 2 and 5 recovers the bounds in (21) and (23), respectively.

Setting \(\lambda = \sqrt{2\log(n/k)}\) in (53) provides an approximation to \(D(\text{cone}(\partial f(x_0)))\).

In particular, \(D(\text{cone}(\partial f(x_0))) \leq 2k(\log(n/k) + 3/2)\). [14] obtains an even tighter bound \(D(\text{cone}(\partial f(x_0))) \leq 2k(\log(n/k) + 3/4)\) starting again from (50), but using different tail bounds for Gaussians. We refer the reader to Proposition 3.10 in [14] for the exact details.

**Acknowledgements**

We would like to thank Professor G. Kutyniok for the invitation to contribute to the Springer/Birkhaeuser book entitled "Compressed Sensing and its Applications". Also, the authors gratefully acknowledge the anonymous reviewers for their attention and their helpful comments.

**References**


