# Primer on $\ell^1$ minimization

**Convex Methods Reading Group**

Badri Narayan Bhaskar

*Presented on* March 23, 2012

## 1 Introduction

Many applications require an estimation of a high dimensional parameter vector from a relatively few independent looks. Even in the simple case of a linear model, which we will study, it is a hopeless task, as it involves solving an underdetermined system. However, we can rescue this situation with the assumption of *sparsity,* i.e., the assumption that the high dimensional parameter vector is non-zero only for a few parameters.

To make this concrete, let us consider a linear model. The target $\beta^0 \in \mathbb{R}^p$ is an unknown high dimensional parameter vector that we wish to estimate, and we make $n$ observations $\{y_i\}_{i=1}^n$ from our linear model of the form $y_i = \langle x_i, \beta^0 \rangle + \varepsilon_i$, where $\varepsilon_i$ is zero mean noise. We can also write this in matrix form as $y = X\beta^0 + \varepsilon$ where $\beta^0 \in \mathbb{R}^p, X \in \mathbb{R}^{n\times p}$ and $y, \varepsilon \in \mathbb{R}^n$.

When $n \ll p$, as mentioned before, there is no hope for recovering $\beta^0$ from $\{(x_i, y_i)\}_{i=1}^n$ , as this would entail solving a underdetermined system. But in many problems, only a few of the parameters (say $s$ of them) are actually important or active, and we have more observations than the number of active parameters. In other words, we may assume that $\beta^0$ is *sparse* with $S = \operatorname{supp}\{\beta^0\}$ of size $s = |S| \ll n$. We say $\beta^0$ is $s$-sparse if it has a support of size atmost $s$.

If an oracle were to provide us $S$, the solution is simple. Otherwise, we may have to search through $\sum_{k=0}^{s} \binom{p}{k}$ combinations for the support. In fact, there can be no polynomial time algorithm to do this algorithm for all instances of $\beta, X$ and $y$, since it can be shown that sparse recovery reduces to Subset-Sum and sparse approximation is also NP hard[14]. Since we cannot solve every single instance of this problem in polynomial time, we attempt to find instances or conditions on $X, \beta$ for which we have a tractable algorithm.

Following statistics terminology, we will call $\beta^0$ the target, $X$ the design matrix, and $y$ the observation vector. In this note, we see how to set up a convex relaxaton (called the $\ell^1$ heuristic) to provide a tractable algorithm for estimating $\beta^0$ from $y$ and $X$ in a large number of cases. We will separately explore the condition for exact recovery of $\beta^0$ in the presence of noise and the statistical properties of the estimate in the presence of noise.

## 2 Noiseless Case – Basis Pursuit

Suppose $\beta^0 \in \mathbb{R}^p$ is an unknown target and the design matrix $X \in \mathbb{R}^{n\times p}$ is known. If we observe $y = X\beta^0$, the noiseless sparse recovery problem involves finding a $s$-sparse $\beta^0$ from $y$. The problem is well posed if there is only a unique $\beta$ that is $s$-sparse that satisfies $y = X\beta$. In that case, $\beta^0$ would be the be solution of the (non-convex) optimization problem

$$\begin{aligned} &\underset{\beta}{\text{minimize}} && \|\beta\|_0 \\ &\text{subject to} && y = X\beta. \end{aligned} \tag{1}$$

where $\|\beta\|_0$ sometimes called the $\ell^0$ norm (although it is not really a norm) is the number of nonzero elements in $\beta$. As discussed in the introduction, this is a NP-hard combinatorial optimization problem and is therefore not tractable. So, we will look at a tractable $\ell^1$ norm minimization technique called *Basis Pursuit* and characterize when it works.

## 2.1 Unique Recovery

Suppose we observe $y = X\beta^0$. The first order question is when the optimization problem (1) has $\beta^0$ as the unique solution. Our first proposition shows that as long as $n \geq 2s$, under the mild assumption that every $2s$ columns of $X$ are independent, $\beta^0$ is the unique solution. This ensures that irrespective of the location of the nonzero elements of $\beta^0$, each observation provides an *independent look* at $\beta^0$. The proposition also

**Propositon 1** (Unique Recovery Condition[10])**.** *If every $2s$ columns of $X$ are independent and $n \geq 2s$, every $s$-sparse $\beta^0$ is the unique solution of* (1) *with $y = X\beta^0$. Conversely, if $n < 2s$, one can always find $s$-sparse $\beta^0$ so that* (1) *does not have a unique solution.*

*Proof.* In order that every $2s$ columns of $X$ are independent, note that we need $n \geq 2s$. Now, there cannot be two $s$-sparse vectors $\beta_1$ and $\beta_2$ with $X\beta_1 = X\beta_2$ since there are no $2s$-sparse vectors in the nullspace. If $n < 2s$, every $2s$ columns are linearly dependent and consequently there is $2s$-sparse vector in the nullspace of $X$. Let $\beta$ be any such $2s$ sparse vector in the nullspace. Then, we can write $\beta = \beta_1 + \beta_2$ for two $s$-sparse vectors $\beta_1$ and $\beta_2$ with disjoint supports. We cannot have unique recovery in this case since (1) cannot distinguish $\beta^1$ and $\beta^2$. □

Now that we know a condition for wellposedness of (1), we shall introduce the $\ell^1$ heuristic for solving (1). We will assume that in the following that the unique recovery condition holds.

## 2.2 Basis Pursuit

Instead of (1), the authors in [5] propose solving the *basis pursuit* problem given by

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \|\beta\|_1 \\ \text{subject to} \quad & X\beta = y. \end{aligned} \tag{2}$$

This is a convex relaxation because we replace the non convex function $\|\cdot\|_0$ by its convex envelope on the unit ball. Before we algebraically characterize when the solution of (2) coincides with (1), we will see a geometric intuition for why this might work.

## 2.3 Geometry of L1 Minimization

The solution to (2) is the point of contact of the smallest expansion of the $\ell^1$ norm ball $C = \{\beta \mid \|\beta\|_1 < 1\}$ [1] that touches the hyperplane $\{\beta \mid X\beta = y\}$, specified by the linear constraints. The solution is always on a face of the cross-polytope and with high probability most hyperplanes touch a low dimensional face which contains sparse vectors.

We can also look at the geometric picture on the quotient polytope $Q = XC$. A vector $\hat{\beta}$ solves (2) if and only if $X\hat{\beta}$ is an exposed face of the quotient polytope $Q$. Thus, if $\ell^1$ minimization works, then $X\beta$ should be a low dimensional face whenever $\beta$ is sparse. See [9] for a proof of optimality of (1), that uses the geometric connection between faces of quotient polytopes and solutions of (1)

[1] also called the cross-polytope

## 2.4 Descent Directions and Nullspace Property

The constraint $X\beta = y$ in (1) can be written as $X\left(\beta - \beta^0\right) = 0$, or $\beta \in \beta^0 + \ker(X)$, where $\ker(X)$ denotes the nullspace of $X$. Thus, we can reformulate (2) like this:

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \|\beta\|_1 \\ \text{subject to} \quad & \beta \in \beta^0 + \ker(X) \,. \end{aligned} \tag{3}$$

For any direction $d \in \ker(X)$ , $\beta^0 + d$ is a feasible solution and hence, $\beta^0$ is the unique solution if and only if every direction $d$ in $\ker(X)$ decreases the $\ell^1$ norm at $\beta^0$. Denote the set of descent directions at $\beta^0$ by

$$\begin{aligned} D_0\left(\beta^0\right) &= \left\{\beta - \beta^0 : \|\beta\|_1 < \left\|\beta^0\right\|_1\right\} \\ &= \left\{d \,|\, \left\|\beta^0 + d\right\|_1 < \left\|\beta^0\right\|_1\right\}. \end{aligned}$$

In summary, we have the following elementary proposition:

**Propositon 2** (Exact Recovery)**.** *The unique optimal solution to Basis pursuit with $y = X\beta^0$ will recover $\beta^0$ as the solution, if and only if the set of descent directions $D(\beta^0)$ at $\beta^0$ does not meet the nullspace* $\ker(X)$ *of the design matrix.*

For our analysis of the robust version, we will also need the set of approximate descent directions. Define

$$\begin{aligned} D_\gamma\left(\beta^0\right) &= \left\{\beta - \beta^0 \;\middle|\; \|\beta\|_1 < \left\|\beta^0\right\|_1\right\} \\ &= \left\{d \;\middle|\; \left\|\beta^0 + d\right\|_1 < \left\|\beta^0\right\|_1\right\}. \end{aligned}$$

### 2.4.1 Descent Cone

Denote the cone of descent directions $\operatorname{cone}\left(D\left(\beta^0\right)\right)$ by $C_0\left(\beta^0\right)$ . Therefore,

$$\begin{aligned} C_0\left(\beta^0\right) &= \operatorname{cone}\left\{d \;\middle|\; \left\|\beta^0 + d\right\|_1 \leq \left\|\beta^0\right\|_1\right\} \\ &= \left\{d \;\middle|\; \exists\alpha > 0 \;\text{ such that }\; \left\|\beta^0 + \alpha d\right\|_1 \leq \left\|\beta^0\right\|_1\right\}. \end{aligned} \tag{4}$$

Then, a sufficient condition for exact recovery is that the descent cone $C_0\left(\beta^0\right)$ does not meet the nullspace $\ker(X)$ of the design matrix.

*Remark* 1. If $X$ is a random matrix, then we care about the probability that the random subspace $\ker(X)$ misses the cone $C(\beta^0)$. An estimate of this probability is given by Gordon's escape through the mesh theorem [12] in terms of the Gaussian width of $C\left(\beta^0\right) \cap S^{n-1}$.

## 2.5 Nullspace Property

A necessary and sufficient condition for basis pursuit to recover *all vectors with support $S$*, is that the nullspace $\ker(X)$ must not meet $D(\beta^0)$ for any $\beta^0$. Denote by $\Sigma_S$ is the set of all $S$-sparse vectors. We can show that elements in the set $\bigcup_{\beta^0 \in \Sigma_S} D(\beta^0)$ are characterized by concentration of $\ell^1$ norm on small supports:

**Lemma 1.** *$d \in D(\beta^0)$ for some $\beta^0 \in \Sigma_S$ if and only if $\|d_{S^c}\|_1 \leq \|d_S\|_1$ .*

*Proof.* Suppose $d \in D(\beta^0)$ for some $\beta^0 \in \Sigma_S$. Then, introducing the notation $z_S = z \circ 1_S$, we have

$$\left\|\beta^0 + d_S\right\|_1 + \|d_{S^c}\|_1 \leq \left\|\beta^0\right\|_1$$

whence, by an application of triangle inequality $\left\|\beta^0\right\|_1 - \left\|\beta^0 + d_S\right\|_1 \leq \|d_S\|_1$, we get $\|d_{S^c}\|_1 \leq \|d_S\|_1$ . Conversely, if $\|d_{S^c}\|_1 \leq \|d_S\|_1$ , then, for $\beta^0 = -d_S$, we have,

$$\left\|\beta^0 + d_S\right\|_1 \leq \left\|\beta^0\right\|_1$$

This completes the proof. □

Using this lemma, and Proposition 2, we have shown the following useful theorem

**Theorem 1** (Nullspace Property [6] [2]). *Basis pursuit will recover every vector supported on $S$ if and only if the nullspace* $\ker(X)$ *does not meet the set* $C_0'(S) = \{d \mid \|d_{S^c}\|_1 \le \|d_S\|_1\}$.

*Remark* 2. Since the set $C_0'(S)$ is a cone, we have by Lemma 1 that $C_0'(S) = \bigcup_{\beta^0 \in \Sigma_S} C(\beta^0)$.

### 2.6 Connection to Uncertainty Principles

Nullspace property requires that no vector in the nullspace be concentrated on a small set. If we view the design matrix as producing coefficients on another basis, then the uncertainty principles can guarantee that no vector in the nullspace of the design matrix is also concentrated on a small set. For instance, suppose $X \in \mathbb{R}^{n\times p}$ is a partial DFT matrix so that $X\beta$ for $\beta \in \mathbb{R}^p$ yields the first $n$ DFT coefficients, $d \in \text{nullspace}\,(X)$ when the first $n$ fourier coefficients of $d$ are zero. Since $d$ has many zeros in its spectrum, it cannot be too concentrated on a set of small support. This means that the energy of $d$ (in terms of the $\ell^1$ norm) cannot be concentrated on a support of small size $S$. Thus, the discrete time-frequency uncertainty principle [11] ensures that the nullspace of $X$ does not meet the cone $C'(S)$.

The bounds produced by time frequency uncertainty principle are quite weak. Stronger results are possible for partial fourier matrices with random frequencies, using a different argument [15, 4].

### 2.7 Testing the Nullspace Property

We need only show that $X$ when restricted to the cone of points where $\|d_{S^c}\|_1 < \|d_S\|_1$ has eigenvalue bounded away from zero. Note that this involves minimizing a ratio of norms on a cone. For a fixed design, it is possible to test the nullspace property using SDP relaxations [13, 7] or using second order conic programs [16].

### 2.8 Restricted Eigenvalue Condition

One way of showing that the nullspace property (*i.e.,* condition for theorem 1) holds is by proving that the minimum eigenvalue $1/K_0$ of $X$ restricted to $C_0'$ is bounded away from zero. Define

$$\frac{1}{K_0} = \min_{z\neq 0}\left\{ \frac{\|Xz\|_2}{\|z\|_2} \;\middle|\; z \in C_0'(S) \right\}.$$

The condition $K_0 > 0$ is called the restricted eigenvalue condition [1]. For random design matrices, this is in turn proved by showing that $C_0$ is approximately like the set of sparse vectors and that random designs behave like approximate isometries for sparse vectors. Candes, Romberg and Tao [3], showed that the nullspace property holds for various classes of random matrices with high probability.

## 3 Noisy Case

To handle noise, note that we can no longer have $X\beta = y$ as a constraint. If we knew that $\|X\beta - y\|_2 \le \delta$, inspired by (2), we can propose the optimization problem:

$$\begin{aligned} &\underset{\beta}{\text{minimize}} && \|\beta\|_1 \\ &\text{subject to} && \|X\beta - y\|_2 \le \delta. \end{aligned}$$

by simply replacing the equality constraint with a bound on the residue. There are two other alternative formulations which are equivalent in that there is a choice of $(\delta, \lambda, t)$ such that all the problems yield the same solution:

[2] Nullspace property is often stated in a slightly different form from the one we have chosen. For a direct proof of the usual form, see appendix A.1.

1. **Basis Pursuit Denoising (BPDN)** [5]

$$\text{minimize}_{\beta} \, \frac{1}{2} \left\| X\beta - y \right\|_2^2 + \lambda \left\| \beta \right\|_1 \tag{5}$$

   which is the regularized version and

2. **Least absolute shrinkage and selection operator (LASSO)** [17]

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \left\| X\beta - y \right\|_2^2 \\ \text{subject to} \quad & \left\| \beta \right\|_1 \leq \tau. \end{aligned} \tag{6}$$

   which is the $\ell^1$ constrained version.

Note that all of these have the same Lagrangian and hence similar dual problems. We will pick the regularized version for our analysis. We will study how the *estimation error* $\left\| \beta - \hat{\beta} \right\|_2$ and the *prediction error* $\left\| X\beta - X\hat{\beta} \right\|_2$ scale with the number of measurements $n$, sparsity $s$ and the number of columns $p$. As a first step, we will characterize the optimality conditions of (5).

## 3.1 Optimality Conditions

Denote the objective function for BPDN by $f$, so that

$$f(\beta) = \frac{1}{2} \left\| X\hat{\beta} - y \right\|_2^2 + \lambda \left\| \hat{\beta} \right\|_1 .$$

Notice that $f$ is not smooth because $\left\| \cdot \right\|_1$ is not differentiable. We can characterize the minima using a very elementary argument as in the following lemma.

**Lemma 2. (Optimality Conditions)** *$f\left(\hat{\beta}\right) \leq f(\beta)$ for all $\beta$, if and only if*

*1.* $\left\| X^T \left( X\hat{\beta} - y \right) \right\|_\infty \quad \leq \quad \lambda,$

*2.* $\left\langle \hat{\beta}, X^T \left( X\hat{\beta} - y \right) \right\rangle \quad = \quad \lambda \left\| \hat{\beta} \right\|_1 .$

An elementary proof of this is provided in Appendix A.2.

*Remark* 3. Readers familiar with KKT optimality conditions can observe that the first condition is dual feasibility. This can be verified by deriving the dual. The second condition can be used to derive complementary slackness when we rewrite $\left\langle \hat{\beta}, X^T \left( X\hat{\beta} - y \right) \right\rangle$ as $\left\langle \left| \hat{\beta} \right|, \operatorname{sgn}\left( \hat{\beta} \right) X^T \left( X\hat{\beta} - y \right) \right\rangle$ where I follow the convention that the absolute value and the sgn functions can operate on vectors.

## 3.2 Shrinkage Interpretation

For any $\beta$ with $\left\| X^T (X\beta - y) \right\|_\infty \leq \lambda$, we have

$$\begin{aligned} \left\| X\beta \right\|_2^2 &= \left\| X\hat{\beta} \right\|_2^2 + \left\| X\beta - X\hat{\beta} \right\|_2^2 + \left\langle X\hat{\beta}, X\beta - X\hat{\beta} \right\rangle \\ &\geq \left\| X\hat{\beta} \right\|_2^2 + \left\langle \hat{\beta}, X^T (y - X\beta) \right\rangle + \left\langle \hat{\beta}, X^T \left( y - X\hat{\beta} \right) \right\rangle \\ &\geq \left\| X\hat{\beta} \right\|_2^2 + \lambda \|\hat{\beta}\|_1 - \left\| X^T (X\beta - y) \right\|_\infty \left\| \hat{\beta} \right\|_1 \\ &\geq \left\| X\hat{\beta} \right\|_2^2 . \end{aligned}$$

where we have used Holder's inequality and the optimality condition $(ii)$ from Lemma 2 in the second inequality. Thus, $\hat{\beta}$ is a solution to the problem

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \|X\beta\|_2 \\ \text{subject to} \quad & \left\|X^T(X\beta - y)\right\|_\infty \leq \lambda. \end{aligned} \tag{7}$$

Conversely, if $\hat{\beta}$ solves (7), we have

$$\begin{aligned} \|X\beta - y\|_2^2 + \lambda\|\beta\|_1 &= \left\|X\beta - X\hat{\beta}\right\|_2 + \left\|X\hat{\beta} - y\right\|_2^2 + \left\langle X\beta - X\hat{\beta}, X\beta - y\right\rangle + \lambda\|\beta\|_1 \\ &\geq \left\|X\hat{\beta} - y\right\|_2^2 + \left\langle \beta - \hat{\beta}, X^T(X\beta - y)\right\rangle - +\lambda\|\beta\|_1 \\ &\geq \left\|X\hat{\beta} - y\right\|_2^2 + \lambda\left\|\hat{\beta} - \beta\right\|_1 + \lambda\|\beta\|_1 \\ &\geq \left\|X\hat{\beta} - y\right\|_2^2 + \lambda\left\|\hat{\beta}\right\|_1 \end{aligned}$$

Thus, (7) is merely another way of describing the solution of (5). This gives the interpretation that BPDN shrinks $X\beta$ towards origin while requiring that $\left\|X^T(X\beta - y)\right\|_\infty \leq \lambda$. So, the regularization parameter $\lambda$ controls the amount of shrinking. If $X$ is the identity matrix, this corresponds to soft thresholding $\beta$ by $\lambda$. [8]

*Remark* 4. This equivalent formulation can also be obtained by deriving the dual problem, as shown in Appendix 3.

## 3.3 Choice of Regularization Parameter

We can suggest a choice of regularization parameter. Since we expect $\hat{\beta} \cong \beta^0$, this suggests we pick $\lambda$ bigger than $\mathbb{E}\left\|X^T\left(X\beta^0 - y\right)\right\|_\infty = \mathbb{E}\left\|X^T\varepsilon\right\|_\infty$. Assuming that the columns are normalized so that $\left\|X^Te_i\right\|_2 = 1$ for $i = 1, \ldots, p$, we have

$$\mathbb{E}\left\|X^T\varepsilon\right\|_\infty \leq \sqrt{2\sigma^2\log(n)}$$

where we have used the well-known bound for maximum of Gaussian random variables. By picking $C_0 > 1$, we can also show that overwhelming probability $\left\|X^T\varepsilon\right\|_\infty < C_0\sqrt{2\sigma^2\log(n)}$. For the following, set

$$\lambda = C_0\sqrt{2\sigma^2\log(n)}. \tag{8}$$

## 3.4 Approximate Descent Directions and the strong Restricted Eigenvalue Condition

If $\hat{\beta}$ is the solution of (5), we will see that it is approximately a descent direction for the $\ell^1$ norm. This will allow us to draw an analogue of nullspace property for robust recovery. Let $\gamma = \frac{\left\|X^T\varepsilon\right\|}{\lambda}$. If $\lambda$ is chosen according to (8), we would have $\gamma < 1$ with high probability. Since $f(\hat{\beta}) \leq f(\beta^0)$, which gives

$$\frac{1}{2}\left\|(X\hat{\beta} - X\beta^0) + X\beta^0 - y\right\|_2^2 + \lambda\left\|\hat{\beta}\right\|_1 \leq \frac{1}{2}\left\|X\beta^0 - y\right\|_2^2 + \lambda\left\|\beta^0\right\|_1$$

Rearranging, we get

$$\begin{aligned} \left\|\hat{\beta}\right\|_1 &\leq \left\|\beta^0\right\|_1 + \lambda^{-1}\left\langle X\beta^0 - X\hat{\beta}, X\beta^0 - y\right\rangle \\ &\leq \left\|\beta^0\right\|_1 + \lambda^{-1}\left\langle \beta^0 - \hat{\beta}, X^T\varepsilon\right\rangle \\ &\leq \left\|\beta^0\right\|_1 + \gamma\left\|\beta^0 - \hat{\beta}\right\|_1. \end{aligned} \tag{9}$$

Thus, we have shown that the direction $d = \hat{\beta} - \beta^0$ is in the cone of approximate descent directions, given by

$$C_\gamma\left(\beta^0\right) = \text{cone}\{d \mid \left\|\beta^0 + d\right\|_1 \leq \left\|\beta^0\right\|_1 + \gamma \left\|d\right\|_1\}.$$

Note that $C_0\left(\beta^0\right)$ coincides with the descent cone described in (4) whereas $C_1\left(\beta^0\right)$ includes all the points (by triangle inequality). So, we can think of $C_\gamma$ as the cone of *approximate* descent directions.

Like the descent cone, points in $C_\gamma$ have $\ell^1$ concentration on a small support. In fact, if $d \in C_\gamma$,

$$\left\|\beta^0 + \alpha d\right\|_1 \leq \left\|\beta^0\right\|_1 + \alpha\gamma \left\|d\right\|_1.$$

Writing $d = d_S + d_{S^c}$ where $S = \text{supp}\,\beta^0$, we have

$$\left\|\beta^0 + \alpha d_S\right\|_1 + \alpha \left\|d_{S^c}\right\|_1 \leq \left\|\beta^0\right\|_1 + \alpha\gamma \left\|d\right\|_1.$$

By using the triangle inequality, and rearranging, we get,

$$\left\|d_{S^c}\right\|_1 \leq \frac{1+\gamma}{1-\gamma} \left\|d_S\right\|_1. \tag{10}$$

or equivalently

$$\left\|d\right\|_1 \leq \frac{2}{1-\gamma} \left\|d_S\right\|_1. \tag{11}$$

For exact recovery in the noiseless case, we needed the restricted eigenvalue condition ($K_0 > 0$.) For the noisy case, achieving minimax stability rates requires only a slightly stronger condition. Define,

$$\frac{1}{K_\gamma} = \min_{z \neq 0} \left\{ \frac{\left\|Xz\right\|_2}{\left\|z\right\|_2} \;\middle|\; \left\|z_{S^c}\right\|_1 < \frac{1+\gamma}{1-\gamma} \left\|z_S\right\|_1 \right\}.$$

The condition $K_\gamma > 0$ is the strong restricted eigenvalue condition.

## 3.5 Statistical Performance

Two metrics for studying the quality of the reconstruction are the *prediction error* $\left\|X\beta^0 - X\hat{\beta}\right\|_2^2$ and the *estimation error* $\left\|\beta^0 - \hat{\beta}\right\|_2^2$. We will see that we can obtain good error bounds for both if the matrix $X$ satisfies restricted eigenvalue condition.

### 3.5.1 Slow Prediction Rates

Assuming no conditions on the design matrix, we can guarantee a statistical error rate that depends on $\beta^0$, instead of a global error bound that works for all $s$-sparse vectors. Using the optimality conditions in Lemma 2, we get

$$\begin{aligned}
\left\|X\hat{\beta} - X\beta^0\right\|_2^2 &= \left\langle X\hat{\beta} - X\beta^0, (y - X\beta^0) - (y - X\hat{\beta})\right\rangle \\
&\leq \left\langle \hat{\beta}, X^T(y - X\beta^0)\right\rangle - \left\langle \hat{\beta}, X^T(y - X\hat{\beta})\right\rangle - \left\langle \beta^0, X^T\varepsilon\right\rangle + \left\langle \beta^0, X^T(y - X\hat{\beta})\right\rangle \\
&\leq \left\|X^T\varepsilon\right\|_\infty \left\|\hat{\beta}\right\|_1 - \lambda \left\|\hat{\beta}\right\|_1 + \left\|X^T\varepsilon\right\|_\infty \left\|\beta^0\right\|_1 + \left\|\beta^0\right\|_1 \left\|X^T(y - X\hat{\beta})\right\|_\infty \\
&\leq 2\lambda \left\|\beta^0\right\|_1 \\
&= 2C_0\sqrt{\sigma^2 \log(n)} \left\|\beta^0\right\|_1.
\end{aligned}$$

with high probability.

#### 3.5.2 Faster Prediction Rates

If we assume that the design matrix satisfies the restricted eigenvalue condition, we can then get faster rates and a uniform upper bound on the error rate. See [18] for a discussion of various conditions under which we can get fast rates.

$$\begin{aligned}
\left\|X\hat{\beta} - X\beta^0\right\|_2^2 &= \left\langle X\hat{\beta} - X\beta^0, (y - X\beta^0) - (y - X\hat{\beta})\right\rangle \\
&= \left\langle \hat{\beta} - X\beta^0, X^T(y - X\beta^0) - X^T(y - X\hat{\beta})\right\rangle \\
&\leq \left\|\hat{\beta} - \beta^0\right\|_1 \left(\|X^T\left(y - X\beta^0\right)\|_\infty + \left\|X^T\left(y - X\hat{\beta}\right)\right\|_\infty\right) \\
&\leq 2\lambda \left\|\hat{\beta} - \beta^0\right\|_1 \\
&\leq \frac{4\lambda}{1-\gamma}\left\|\hat{\beta}_S - \beta^0\right\|_1 && \text{(since } \hat{\beta} - \beta^0 \in C_\gamma(\beta^0).\text{)} \\
&\leq \frac{4\sqrt{s}\lambda}{1-\gamma}\left\|\hat{\beta} - \beta^0\right\|_2 && \text{(by Cauchy-Schwarz inequality.)} \\
&\leq \frac{4\sqrt{s}\lambda}{1-\gamma} K_\gamma \left\|X\hat{\beta} - X\beta^0\right\|_2. && \text{(by Restricted Eigenvalue Condition.)}
\end{aligned}$$

Now, substituting the value of $\lambda$ from (8), we get

$$\frac{1}{n}\left\|X\hat{\beta} - X\beta^0\right\|_2^2 = O\left(\frac{\sigma^2 s\log(n)}{n}\right).$$

Using the restricted minimum eigenvalue condition and again using the fact that $\hat{\beta} - \beta^0$ is in $C'_\gamma(S)$, we get

$$\frac{1}{n}\left\|\hat{\beta} - \beta^0\right\|_2^2 = O\left(\frac{\sigma^2 s\log(n)}{n}\right).$$

It can be shown that these rates are actually minimax optimal [2].

# 4 Summary

We first looked at the noiseless case and saw that the $\ell^1$ heuristic can recover every $s$-sparse vector if and only if the design matrix $X$ satisfies the nullspace property. We looked at a couple of ways of verifying that nullspace property holds for a family of design matrices.

In the noisy case, we looked at natural regularized variant of $\ell^1$ minimization. A sufficient condition for achieving minimax estimation rate in the presence of noise is only slightly stronger than the condition for the usual condition (restricted eigenvalue condition) for exact recovery.

# A Proofs

## A.1 The Nullspace Property

If $D\left(\beta^0\right)$ never intersects $\ker(X)$ *for any choice of* $\beta^0$, the $\ell_1$ minimization heuristic will always work. Introducing the notation $z_S = z \circ 1_S$, this is equivalent to the statement that $\|d_S\|_1 \leq \|d_{S^c}\|_1$ for all $d \in \ker(X)$, and when this holds, we say $X$ satisfies **nullspace property**. If $X$ satisfies nullspace property, for any $d \in \ker(X)$,

$$\left\|\beta^0 + d\right\|_1 = \left\|\beta^0 + d_S\right\|_1 + \|d_{S^c}\|_1 \geq \left\|\beta^0 + d_S\right\|_1 + \|d_S\|_1 \geq \left\|\beta^0\right\|_1.$$

Conversely, suppose $\|d_S\|_1 > \|d_{S^c}\|_1$ for some $d \in \ker(X)$, then set $y = Xd_S = -Xd_{S^c}$. Now, $\ell_1$ minimization will not yield the sparse $d_S$ as the solution.

## A.2 Proof of Optimality Conditions

*Proof.* If $\hat{\beta}$ solves (5), if and only if, for every $\beta$, and every $\alpha > 0$, the direction $\alpha(\beta-\hat{\beta})$ reduces the objective:

$$\frac{1}{2}\left\|X\left(\hat{\beta}+\alpha\left(\beta-\hat{\beta}\right)\right)-y\right\|_2^2+\lambda\left\|\hat{\beta}+\alpha\left(\beta-\hat{\beta}\right)\right\|_1 \geq \frac{1}{2}\left\|X\hat{\beta}-y\right\|_2^2+\lambda\left\|\hat{\beta}\right\|_1$$

Rearranging some terms, we get,

$$\frac{1}{2}\alpha\left\|\beta-\hat{\beta}\right\|_2^2+\left\langle\beta-\hat{\beta},X^T\left(X\hat{\beta}-y\right)\right\rangle \geq \lambda\frac{\left(\left\|\hat{\beta}\right\|_1-\left\|\hat{\beta}+\alpha\left(\beta-\hat{\beta}\right)\right\|_1\right)}{\alpha} \tag{12}$$

By convexity of $\|\cdot\|_1$, we have

$$\left\|\beta+\alpha\left(\beta-\hat{\beta}\right)\right\|_1 \leq (1-\alpha)\left\|\hat{\beta}\right\|_1+\alpha\left\|\beta\right\|_1$$

Using this in (12), and letting $\alpha \to 0$, we conclude that $\hat{\beta}$ is a solution of (5) only if, for every $\beta$,

$$\left\langle\beta-\hat{\beta},\lambda^{-1}X^T\left(X\hat{\beta}-y\right)\right\rangle \geq \left\|\hat{\beta}\right\|_1-\left\|\beta\right\|_1. \tag{13}$$

*Remark* 5. Readers familiar with subgradients may note that the previous inequality says that the negative gradient of the quadratic term is a subgradient of the $\ell^1$ term. So, it can be obtained from the subgradient condition $0 \in \partial f\left(\hat{\beta}\right)$, which is a characterization of minima of a nonsmooth function like $f$.

Conversely, observe that the previous inequality implies $f(\hat{\beta}) \leq f(\beta)$ for every $\beta$. We can eliminate the universal quantifier by rearranging the equation and introducing an infimum:

$$\begin{aligned}\left\|\hat{\beta}\right\|_1-\left\langle\hat{\beta},\lambda^{-1}X^T\left(X\hat{\beta}-y\right)\right\rangle &= \inf_{\beta}\left\{\left\|\beta\right\|_1-\left\langle\beta,\lambda^{-1}X^T\left(X\hat{\beta}-y\right)\right\rangle\right\}\\ &= \begin{cases}0, & \left\|\lambda^{-1}X^T\left(X\hat{\beta}-y\right)\right\|_\infty \leq 1\\ -\infty, & \text{otherwise}.\end{cases}\end{aligned}$$

where the last inequality is Holder's. So, we have

$$\left\|X^T\left(X\hat{\beta}-y\right)\right\|_\infty \leq \lambda$$

and

$$\lambda\left\|\hat{\beta}\right\|_1 \leq \left\langle\hat{\beta},X^T\left(X\hat{\beta}-y\right)\right\rangle$$

By another application of Holder's inequality, it is clear that the second inequality must be an equality. This completes the proof. □

## A.3 Dual Problem

**Lemma 3** (Dual Problem)**.** *The dual problem of* (2) *is given by*

$$\begin{aligned}&\underset{z}{\mathit{maximize}} && \frac{1}{2}\left(\|y\|_2^2-\|y-z\|_2^2\right)\\ &\mathit{subject\ to} && \left\|X^Tz\right\|_\infty \leq \tau.\end{aligned} \tag{14}$$

*The dual problem admits a unique solution $\hat{z}$. The primal solution $\hat{x}$ and the dual solution $\hat{z}$ are specified by the optimality conditions and there is no duality gap.*

*1.* $y = \hat{X\hat{\beta}} + \hat{z}$,

*2.* $\left\|X^T\hat{z}\right\|_\infty \leq \lambda$,

*3.* $\left\langle \hat{z}, X\hat{\beta} \right\rangle = \lambda \left\|\hat{\beta}\right\|_1$.

*Proof.* We can rewrite the primal problem (5) as a constrained optimization problem:

$$\begin{array}{ll} \underset{x,u}{\text{minimize}} & \frac{1}{2}\left\|y-u\right\|_2^2 + \lambda\left\|\beta\right\|_1 \\ \text{subject to} & u = X\beta. \end{array}$$

Now, we can introduce the dual variable $z$ and thus write down Lagrangian:

$$L(\beta, u, z) = \frac{1}{2}\left\|y-u\right\|_2^2 + \lambda\left\|\beta\right\|_1 + \left\langle z, u - X\beta\right\rangle .$$

so that the dual function is given by

$$\begin{aligned} g(z) &= \inf_{\beta,u} L(\beta,u,z) \\ &= \inf_x \left(\frac{1}{2}\left\|y-u\right\|_2^2 + \left\langle z,u\right\rangle\right) + \inf_u\left(\lambda\left\|\beta\right\|_1 - \left\langle X^Tz,\beta\right\rangle\right) \\ &= \inf_u \frac{1}{2}\left(\left\|u-y+z\right\|_2^2 + \left\|y\right\|_2^2 - \left\|z-y\right\|_2^2\right) + \inf_u\left(\lambda\left\|\beta\right\|_1 - \left\langle X^Tz,\beta\right\rangle\right) \\ &= \begin{cases} \frac{1}{2}\left(\left\|y\right\|_2^2 - \left\|y-z\right\|_2^2\right), & \text{if } \left\|X^Tz\right\|_\infty \leq \lambda \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

where the second equality follows by completing the squares and the last equality uses Holder's inequality. Thus the dual problem of maximizing $g(z)$ can be written as in (14).

The solution to the dual problem is the unique projection $\hat{z}$ of $y$ on to the closed convex set $C = \{z : \left\|X^Tz\right\|_\infty \leq \lambda\}$. By projection theorem for closed convex sets, $\hat{z}$ is a projection of $y$ onto $C$ if and only if $\hat{z} \in C$ and $\left\langle z - \hat{z}, y - \hat{z}\right\rangle \leq 0$ for all $z \in C$, or equivalently if $\left\langle \hat{z}, y-\hat{z}\right\rangle \geq \sup_{z\in C}\left\langle z, y-\hat{z}\right\rangle$. This condition is satisfied for $\hat{z} = y - X\hat{\beta}$ where $\hat{\beta}$ minimizes $f(\beta)$ by Lemma 2. The absence of duality gap can be derived thus:

$$\begin{aligned} f(\hat{\beta}) &= \frac{1}{2}\left\|y - X\hat{\beta}\right\|_2^2 + \lambda\left\|\beta\right\|_1 \\ &= \frac{1}{2}\left\|y - X\hat{\beta}\right\|_2^2 + \left\langle \beta, X^T(y - X\hat{\beta})\right\rangle \\ &= \frac{1}{2}\left\|-X\hat{\beta}\right\|_2^2 + \left\langle \hat{z}, X\hat{\beta}\right\rangle \\ &= \frac{1}{2}\left\|\hat{z}\right\|_2^2 + \left\langle \hat{z}, y - \hat{z}\right\rangle \\ &= g(\hat{z}). \end{aligned}$$

□

## References

[1] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[2] E.J. Candes and M.A. Davenport. How well can we estimate a sparse vector? *Arxiv preprint arXiv:1104.5246*, 2011.

[3] E.J. Candes, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[4] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

[5] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, pages 129–159, 2001.

[6] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc*, 22(1):211–231, 2009.

[7] A. d'Aspremont and L. El Ghaoui. Testing the nullspace property using semidefinite programming. *Mathematical programming*, 127(1):123–144, 2011.

[8] D.L. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

[9] D.L. Donoho. Neighborly polytopes and sparse solution of underdetermined linear equations. 2004.

[10] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197, 2003.

[11] D.L. Donoho and P.B. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, pages 906–931, 1989.

[12] Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in n. *Geometric Aspects of Functional Analysis*, pages 84–106, 1988.

[13] A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via 1 minimization. *Mathematical programming*, 127(1):57–88, 2011.

[14] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[15] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[16] G. Tang and A. Nehorai. Performance analysis of sparse recovery based on constrained minimal singular values. *Arxiv preprint arXiv:1004.4222*, 2010.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[18] S.A. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.