High-dimensional Estimation
10716, Spring 2020
Pradeep Ravikumar

1 Introduction

In classical sampling regimes, the number of parameters d is typically much smaller than n. Classical statistical analyses in fact assume that d is fixed, and analyze the behavior of the estimator risk as the number of samples n scales to infinity, treating d as just another constant. In high-dimensional estimation, we no longer make such an assumption, and even allow for the number of samples n to be smaller than the dimension d. This introduces a number of challenges, ranging over a lack of identifiability, and a low signal to noise ratio. This has thus led to devising various types of additional "side information" or constraints that ensures identifiability, and estimation with strong guarantees. Showing the latter has also necessitaed new tools for statistical analyses.

Some good recent on high-dimensional estimation are: Hastie, Tibshirani & Wainwright (2015), Buhlmann & van de Geer (2011), Wainwright (2017).

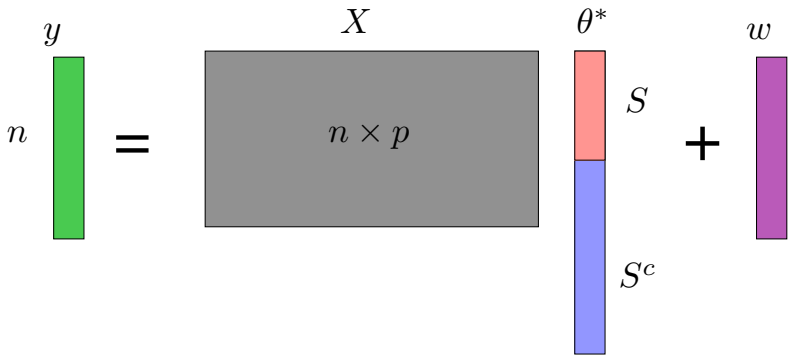
2 Preliminaries: Sparsity and the Lasso

Before we consider a more general account, let us briefly review the case of additional constraint of sparsity and the Lasso estimator. A natural additional constraint one could impose on the model parameter is that it be sparse, so that many of the values of the model parameter are identically zero. In other words,

$$\|\theta^*\|_0 = |\{j \in \{1, \ldots, p\} : \theta^*_j \neq 0\}|,$$

is small.

Consider the sparse linear regression model: $y_i = x_i^T \theta^* + w_i$, for $i \in [n]$, which can be collated in vector form as $y = X\theta^* + w$, and where θ^* is sparse.

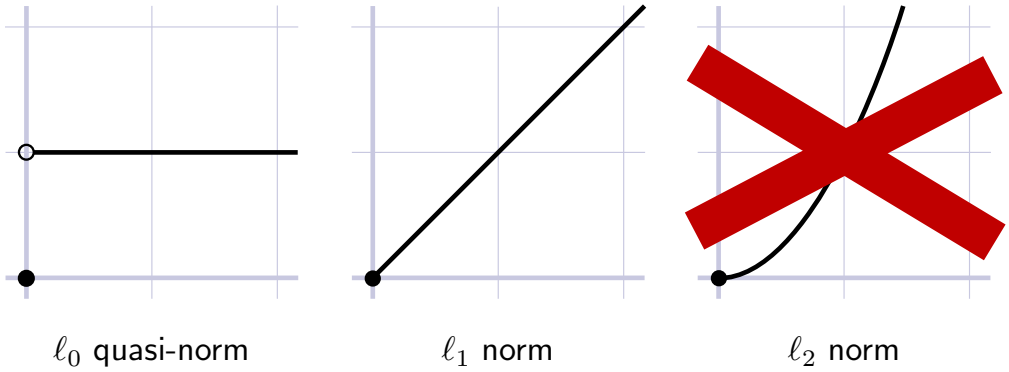


We can then estimate the parameters of this sparse linear model via ℓ_0 constrained linear regression:

$$\min_{\theta} \|y - X\theta\|_2^2$$
$$\text{s.t. } \|\theta\|_0 \leq k.$$

But the estimation problem is non-convex, and NP-Hard (Davis 1994, Natarajan 1995).

The ℓ_1 norm $\|\theta\|_1 = \sum_{j=1}^{p}$ is the closest "convex" norm to the ℓ_0 penalty, and can thus be thought of as a measure of sparsity. This can be readily seen even when $p = 1$:



(Image from Tropp 06)

As an example, consider the vectors $x = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ and $y = (1, 0, \ldots, 0)$ have the same L_2 norm. But $||y||_1 = 1 < ||x||_1 = \sqrt{d}$.

This thus motivates ℓ_1 regularized least squares estimator, also known as the Lasso.

Estimator: Lasso program

$$\widehat{\theta} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \lambda_n \sum_{j=1}^{p}|\theta_j|$$

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Haupt & Nowak, 2006; Zhao/Yu, 2006; Wainwright, 2006; Zou, 2006; Koltchinskii, 2007; Meinshausen/Yu, 2007; Tsybakov et al., 2008

In an equivalent constrained form, the estimator is given as:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2$$
$$\text{s.t. } \|\theta\|_1 \leq C.$$

These are convex problems which can thus be solved efficiently.

3 Theoretical analysis of the lasso

We now discuss some theoretical properties of the Lasso.

There has been an enormous amount theoretical work analyzing the performance of the lasso. Some references are Greenshtein & Ritov (2004), Fuchs (2005), Donoho (2006), Candes & Tao (2006), Meinshausen & Buhlmann (2006), Zhao & Yu (2006), Candes & Plan (2009), Wainwright (2009).

3.1 ℓ_2 error

A common assumption imposed is that X satisfies the **restricted eigenvalue condition** with constant $\phi_0 > 0$, i.e.,

$$\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2 \quad \text{for all subsets } J \subseteq \{1,\ldots,p\} \text{ such that } |J| = s_0$$
$$\text{and all } v \in \mathbb{R}^p \text{ such that } \|v_{J^c}\|_1 \leq 3\|v_J\|_1. \quad (1)$$

Think of v as an error direction $\widehat{\beta} - \beta_0$. The assumption ensures that error in parameter space is bounded by error in prediction space. This cannot hold for all directions since in high-

dimensions X is low-rank, and its minimum eigenvalue (the unconstrained minimum above) is zero. The condition thus only requires this hold for some directions: these essentially are directions we know (or can show) that the Lasso error $\widehat{\beta} - \beta_0$ will lie in.

It can be shown that under this assumption, the Lasso estimator satisfies:

$$\|\widehat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0 \log p}{n\phi_0^2} \tag{2}$$

with probability tending to 1. (This condition can be slightly weakened, but not much.)

The condition is unlikely to hold in any real problem. Nor is it checkable. The proof is in the appendix.

3.2 Support recovery

Here we discuss results on support recovery of the lasso estimator. There are a few versions of support recovery results and Buhlmann & van de Geer (2011) is a good place to look for a thorough coverage. Here we describe a result due to Wainwright (2009), who introduced a proof technique called the *primal-dual witness method*. The assumptions are even stronger (and less believable) than in the previous section. In addition to the previous assumptions we need:

Mutual incoherence: for some $\gamma > 0$, we have

$$\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq 1 - \gamma, \quad \text{for } j \notin S,$$

This condition requires the regression coefficients

$$\eta_j(S) = (X_S^T X_S)^{-1} X_S^T X_j,$$

given by regressing each X_j on the truly active variables X_S, to be small (in ℓ_1 norm) for all $j \notin S$. In other words, no truly inactive variables can be highly correlated (or well-explained, in a linear projection sense) by any of the truly active variables.

Minimum eigenvalue: for some $C > 0$, we have

$$\Lambda_{\min}\left(\frac{1}{n} X_S^T X_S\right) \geq C,$$

where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A. This is somewhat like the restricted eigenvalue condition on X. Here we are restricting to directions with support in S.

Minimum signal:

$$\beta_{0,\min} = \min_{j \in S} |\beta_{0,j}| \geq \lambda \|(X_S^T X_S)^{-1}\|_\infty + \frac{4\gamma\lambda}{\sqrt{C}},$$

where $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^q |A_{ij}|$ denotes the ℓ_∞ norm of an $m \times q$ matrix A. This minimum signal condition ensures that the nonzero entries of the true coefficient vector β_0 are big enough to detect.

Under these assumptions, once can show that, if λ is chosen just right, then

$$P(\text{support}(\widehat{\beta}) = \text{support}(\beta)) \to 1. \tag{3}$$

These conditions are quite restrictive, and are not needed for risk bounds, but are crucial to support recovery.

3.3 What happens if assumptions don't hold?

If these assumptions don't hold, one can still show consistency, but with what are known as *slow rates*. It is this inherent robustness that is likely responsible for the practical popularity of the Lasso.

Thes results to follow don't place any real assumptions on the predictor matrix X, but deliver slow(er) rates for the risk of the lasso estimator than what we would get under more assumptions, hence their name.

We will assume the standard linear model with X fixed, and $\epsilon \sim N(0, \sigma^2)$. We will also assume that $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$. That the errors are Gaussian can be easily relaxed to sub-Gaussianity.

The lasso estimator in bound form (**??**) is particularly easy to analyze. Suppose that we choose $t = \|\beta_0\|_1$ as the tuning parameter. Then, simply by virtue of optimality of the solution $\widehat{\beta}$ in (**??**), we find that

$$\|y - X\widehat{\beta}\|_2^2 \leq \|y - X\beta_0\|_2^2,$$

or, expanding and rearranging,

$$\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 2\langle \epsilon, X\widehat{\beta} - X\beta_0 \rangle.$$

Here we denote $\langle a, b \rangle = a^T b$. The above is sometimes called the *basic inequality* (for the lasso in bound form). Now, rearranging the inner product, using Holder's inequality, and recalling the choice of bound parameter:

$$\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 2\langle X^T\epsilon, \widehat{\beta} - \beta_0 \rangle \leq 4\|\beta_0\|_1 \|X^T\epsilon\|_\infty.$$

Notice that $\|X^T\epsilon\|_\infty = \max_{j=1,\ldots,p} |X_j^T\epsilon|$ is a maximum of p Gaussians, each with mean zero and variance upper bounded by $\sigma^2 n$. By a standard maximal inequality for Gaussians, for any $\delta > 0$,

$$\max_{j=1,\ldots,p} |X_j^T\epsilon| \leq \sigma\sqrt{2n\log(ep/\delta)},$$

with probability at least $1-\delta$. Plugging this to the second-to-last display and dividing by n, we get the finite-sample result for the lasso estimator

$$\frac{1}{n}\|X\widehat{\beta} - X\beta_0\|_2^2 \leq 4\sigma\|\beta_0\|_1\sqrt{\frac{2\log(ep/\delta)}{n}}, \tag{4}$$

with probability at least $1-\delta$.

The high-probability result (4) implies an in-sample risk bound of

$$\frac{1}{n}\mathbb{E}\|X\widehat{\beta} - X\beta_0\|_2^2 \lesssim \|\beta_0\|_1\sqrt{\frac{\log p}{n}}.$$

Compare to this with the risk bound of Lasso scaling as $s_0 \log p/n$ when β_0 has s_0 nonzero components. If each of the nonzero components here has constant magnitude, then above risk bound for the lasso estimator is on the order of $s_0\sqrt{\log p/n}$, which is much slower.

Predictive risk. Instead of in-sample risk, we might also be interested in out-of-sample risk, as after all that reflects actual (out-of-sample) predictions. In least squares, recall, we saw that out-of-sample risk was generally higher than in-sample risk. The same is true for the lasso Chatterjee (2013) gives a nice, simple analysis of out-of-sample risk for the lasso. He assumes that x_0, x_i, $i = 1,\ldots,n$ are i.i.d. from an arbitrary distribution supported on a compact set in $\mathbb{R}^p$, and shows that the lasso estimator in bound form (**??**) with $t = \|\beta_0\|_1$ has out-of-sample risk satisfying

$$\mathbb{E}(x_0^T\widehat{\beta} - x_0^T\beta)^2 \lesssim \|\beta_0\|_1^2\sqrt{\frac{\log p}{n}}.$$

The proof is not much more complicated than the above, for the in-sample risk, and reduces to a clever application of Hoeffding's inequality, though we omit it for brevity. Note here the dependence on $\|\beta_0\|_1^2$, rather than $\|\beta_0\|_1$ as in the in-sample risk.

Oracle inequality. If we don't want to assume linearity of the mean then we can still derive an *oracle inequality* that characterizes the risk of the lasso estimator in excess of the risk of the best linear predictor. For this part only, assume the more general model

$$y = \mu(X) + \epsilon,$$

with an arbitrary mean function $\mu(X)$, and normal errors $\epsilon \sim N(0,\sigma^2)$. We will analyze the bound form lasso estimator (**??**) for simplicity. By optimality of $\widehat{\beta}$, for any other $\widetilde{\beta}$ feasible

for the lasso problem in (**??**), it holds that[1]

$$\langle X^T(y - X\widehat{\beta}), \widetilde{\beta} - \widehat{\beta}\rangle \leq 0. \tag{5}$$

Rearranging gives

$$\langle \mu(X) - X\widehat{\beta}, X\widetilde{\beta} - X\widehat{\beta}\rangle \leq \langle X^T\epsilon, \widehat{\beta} - \widetilde{\beta}\rangle. \tag{6}$$

Now using the polarization identity $\|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2 = 2\langle a, b\rangle$,

$$\|X\widehat{\beta} - \mu(X)\|_2^2 + \|X\widehat{\beta} - X\widetilde{\beta}\|_2^2 \leq \|X\widetilde{\beta} - \mu(X)\|_2^2 + 2\langle X^T\epsilon, \widehat{\beta} - \widetilde{\beta}\rangle,$$

and from the exact same arguments as before, it holds that

$$\frac{1}{n}\|X\widehat{\beta} - \mu(X)\|_2^2 + \frac{1}{n}\|X\widehat{\beta} - X\widetilde{\beta}\|_2^2 \leq \frac{1}{n}\|X\widetilde{\beta} - \mu(X)\|_2^2 + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1-\delta$. This holds simultaneously over all $\widetilde{\beta}$ with $\|\widetilde{\beta}\|_1 \leq t$. Thus, we may write, with probability $1-\delta$,

$$\frac{1}{n}\|X\widehat{\beta} - \mu(X)\|_2^2 \leq \left\{\inf_{\|\widetilde{\beta}\|_1 \leq t} \frac{1}{n}\|X\widetilde{\beta} - \mu(X)\|_2^2\right\} + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}}.$$

Also if we write $X\widetilde{\beta}^{\text{best}}$ as the best linear that predictor of ℓ_1 at most t, achieving the infimum on the right-hand side (which we know exists, as we are minimizing a continuous function over a compact set), then

$$\frac{1}{n}\|X\widehat{\beta} - X\widetilde{\beta}^{\text{best}}\|_2^2 \leq 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1-\delta$.

3.4 Problems With Sparsity

Sparse estimators are convenient and popular but they can some problems. Say that $\widehat{\beta}$ is **weakly sparsistent** if, for every β,

$$P_\beta\big(I(\widehat{\beta}_j = 1) \leq I(\beta_j = 1) \ \text{ for all } j\big) \to 1 \tag{7}$$

as $n \to \infty$. In particular, if $\widehat{\beta}_n$ is sparsistent, then it is weakly sparsistent. Suppose that d is fixed. Then the least squares estimator $\widehat{\beta}_n$ is minimax and satisfies

$$\sup_\beta E_\beta(n\|\widehat{\beta}_n - \beta\|^2) = O(1). \tag{8}$$

But sparsistent estimators have much larger risk:

[1] To see this, consider minimizing a convex function $f(x)$ over a convex set C. Let $\widehat{x}$ be a minimizer. Let $z \in C$ be any other point in C. If we move away from the solution $\widehat{x}$ we can only increase $f(\widehat{x})$. In other words, $\langle \nabla f(\widehat{x}), z - \widehat{z}\rangle \geq 0$.

Theorem 1 (Leeb and Pötscher (2007)) *Suppose that the following condiitons hold:*

1. *d is fixed.*
2. *$n^{-1}\mathbb{X}^T\mathbb{X} \to Q$ for some positive definite matrix Q.*
3. *The errors ϵ_i are independent with mean 0, finite variance σ^2 and have a density f satisfying*
$$0 < \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx < \infty.$$

If $\widehat{\beta}$ is weakly sparsistent then

$$\sup_\beta E_\beta(n\|\widehat{\beta}_n - \beta\|^2) \to \infty. \tag{9}$$

More generally, if ℓ is any nonnegative loss function then

$$\sup_\beta E_\beta(\ell(n^{1/2}(\widehat{\beta}_n - \beta))) \to \sup_s \ell(s). \tag{10}$$

Proof. Choose any $s \in \mathbb{R}^d$ and let $\beta_n = -s/\sqrt{n}$. Then,

$$\begin{aligned}\sup_\beta E_\beta(\ell(n^{1/2}(\widehat{\beta} - \beta)) &\geq E_{\beta_n}(\ell(n^{1/2}(\widehat{\beta} - \beta)) \geq E_{\beta_n}(\ell(n^{1/2}(\widehat{\beta} - \beta))I(\widehat{\beta} = 0)) \\ &= \ell(-\sqrt{n}\beta_n)P_{\beta_n}(\widehat{\beta} = 0) = \ell(s)P_{\beta_n}(\widehat{\beta} = 0).\end{aligned}$$

Now, $P_0(\widehat{\beta} = 0) \to 1$ by assumption. It can be shown that we also have $P_{\beta_n}(\widehat{\beta} = 0) \to 1$ (due to a property called contiguity.) Hence, with probability tending to 1,

$$\sup_\beta E_\beta(\ell(n^{1/2}(\widehat{\beta} - \beta)) \geq \ell(s).$$

Since s was arbitrary the result follows. □

It follows that, if R_n denotes the minimax risk then

$$\sup_\beta \frac{R(\widehat{\beta}_n)}{R_n} \to \infty.$$

The implication is that when d is much smaller than n, sparse estimators have poor behavior. However, when d_n is increasing and $d_n > n$, the least squares estimator no longer satisfies (8). Thus we can no longer say that some other estimator outperforms the sparse estimator. In summary, sparse estimators are well-suited for high-dimensional problems but not for low dimensional problems.

4 Beyond Sparsity

Let us now consider some other popular classes of structural constraints on model parameters that go beyond simple sparsity.

Group-Sparsity. Suppose the model parameters can be grouped into natural groups, so that $\theta = \big(\underbrace{\theta_1, \ldots, \theta_{|G_1|}}_{\theta_{G_1}}, \ldots, \underbrace{\theta_{p-|G_m|+1}, \ldots, \theta_p}_{\theta_{G_m}}\big)$

A **group** analog of sparsity is where many of the groups are identically zero. For instance $\theta^* = \big(\underbrace{*, \ldots, *}_{\theta_{G_1}}, 0, \ldots, 0, \ldots\big)$ has a single non-zero group. Thus, if we construct a vector of ℓ_q norms (for $q \geq 2$) of the group sub-vectors $\{\|\theta_{G_j}\|_q\}_{j=1}^m$, then group sparsity entails that this vector is sparse, so that $\{j \in [m] : \|\theta_{G_j}\|_q \neq 0\}$ is small. This thus motivates the group-analogue of the ℓ_1 norm, also called the group Lasso norm: $\sum_{j=1}^m \|\theta_{G_j}\|_q$.

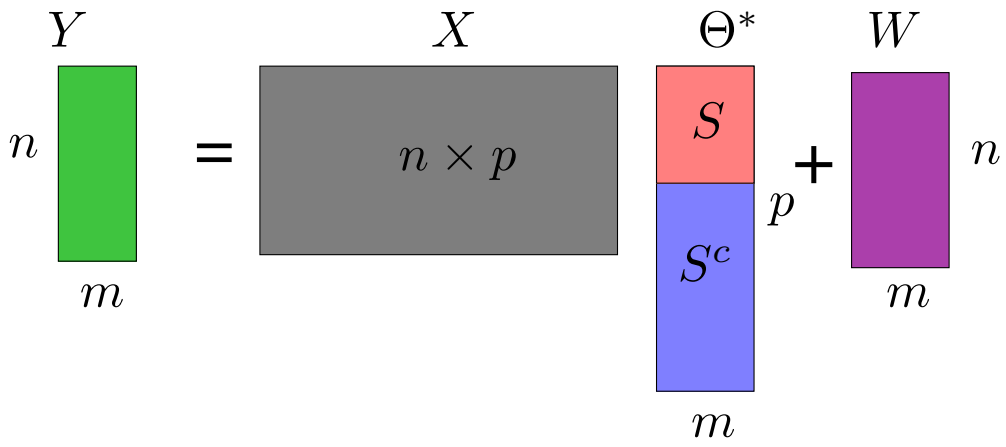
Block Sparsity. A natural variant of such group sparsity is what is known as block-sparsity. Consider the task of multiple linear regression. Here, we have m response variables and m corresponding linear regression models:

$$Y_i^{(l)} = X_i^T \Theta^{(l)} + w_i^{(l)},\ i = 1, \ldots, n.$$

We can collate these into matrices $Y \in \mathbb{R}^{n \times m}$, $X \in \mathbb{R}^{n \times p}$ and $\Theta \in \mathbb{R}^{m \times p}$ so that we get the following compact representation of the multiple linear regression model:

$$Y = X\Theta + W.$$

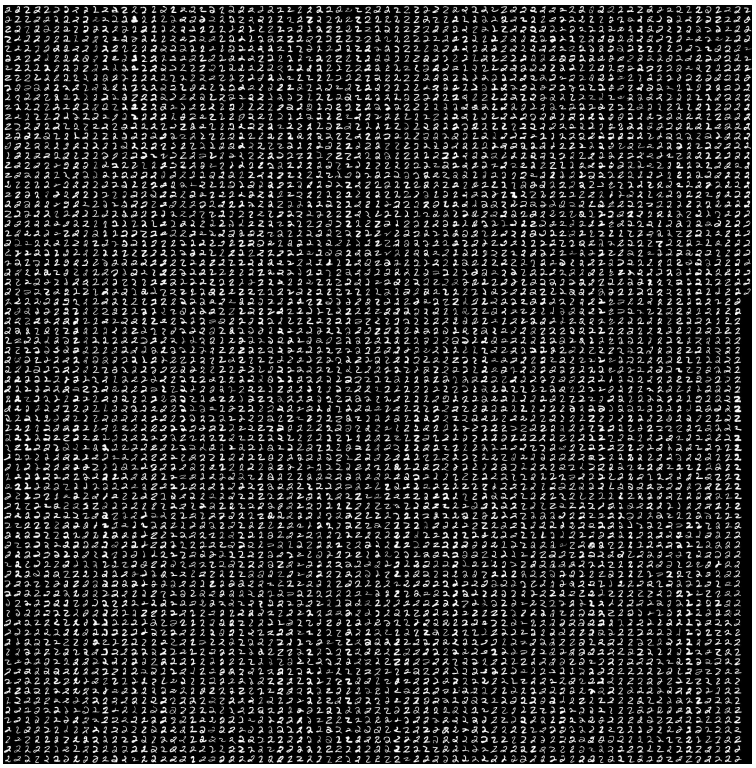
A natural set of groups given the parameter matrix Θ are the rows of the matrix, so that group sparsity here would correspond to most of the rows being entirely zero, other than a few rows with non-zero entries. Such block sparsity here has a very natural interpretation: the set of relevant (specifically, the set of irrelevant) features are shared across the multiple linear regression tasks, even while the actual values of the relevant feature weights can be different.



Thus block sparsity of Θ would entail that: $|\{j \in \{1, \ldots, p\} : \Theta_{j\cdot} \neq 0\}|$ is small. While the block Lasso norm is given as: $\sum_{j=1}^{m} \|\Theta_{j\cdot}\|_q$. The block-sparse estimator of multiple linear regression is then given as:

$$\min_{\Theta} \left\{ \sum_{l=1}^{m} \sum_{i=1}^{n} (Y_i^{(l)} - X_i^T \Theta_{\cdot l})^2 + \lambda \sum_{j=1}^{p} \|\Theta_{j\cdot}\|_q \right\}.$$

Example: Handwriting Recognition.



Consider data consisting of written images of digits from multiple writers. The task is that of classification: recognize the digit given a new image. One could model digit recognition for each writer separately, or mix all digits for training. An alternative in the middle is to use block sparsity. We could model digit recognition for each writer, but make the models share relevant features (where an image is represented as a vector of features).

Low Rank Structure. Suppose we have matrix-structured observations: $X \in \mathbb{R}^{k \times m}$, $Y \in \mathbb{R}$. The corresponding parameters are matrices as well: $\Theta \in \mathbb{R}^{k \times m}$.

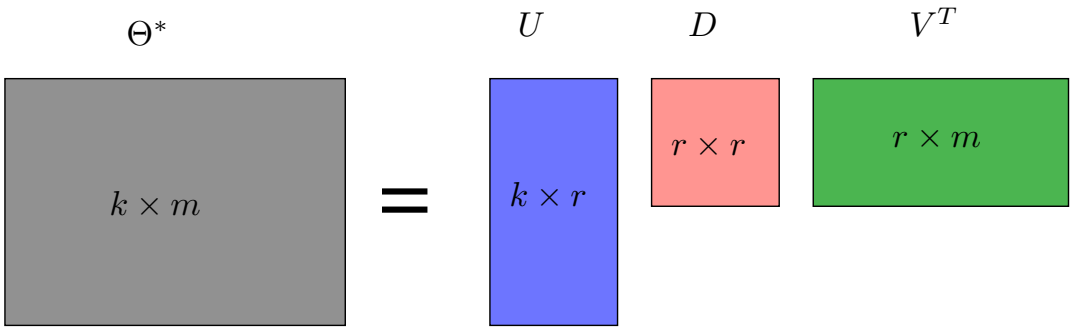
A linear model in this setting would take the form:

$$Y_i = \mathrm{tr}(X_i \Theta) + W_i, \ i = 1, \ldots, n.$$

Such matrix-structured observations are ubiquitous in fMRI image data, EEG data decoding, neural response modeling, as well as financial data. They also arise in collaborative filtering: predicting user preferences for items (such as movies) based on their and other users' ratings of related items.

Recall that the singular values of a matrix $A \in \mathbb{R}^{k \times m}$ are the square-roots of non-zero eigenvalues of $A^T A$. The rank of a matrix A can then be expressed as $\mathrm{rank}(A) = |\{i \in \{1, \ldots, \min(k, m)\} : \sigma_i \neq 0\}|$. Computing an ℓ_1 norm of the singular values rather than the

ℓ_0 quasi norm yields what is known as the nuclear norm: $\|A\|_* = \sum_{i=1}^{\min\{k,m\}} \sigma_i$. Just as the ℓ_1 norm regularization encourages vectors to be sparse, nuclear norm regularization encourages matrices to be low-rank.



Set-up: Matrix $\Theta^* \in \mathbb{R}^{k \times m}$ with rank $r \ll \min\{k, m\}$.

Estimator:

$$\widehat{\Theta} \in \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\langle X_i, \Theta \rangle\rangle)^2 + \lambda_n \sum_{j=1}^{\min\{k,m\}} \sigma_j(\Theta)$$

Some past work: Frieze et al., 1998; Achilioptas & McSherry, 2001; Srebro et al., 2004; Drineas et al., 2005; Rudelson & Vershynin, 2006; Recht et al., 2007; Bach, 2008; Meka et al., 2009; Candes & Tao, 2009; Keshavan et al., 2009

5 Unified Frameworks for Analyzing High-dimensional Models

We will now briefly review the unified framework from (Negahban, Ravikumar, Wainwright, Yu 12) for statistical analyses of general high-dimensional models, with general structural constraints.

In the previous section, we saw varied structural constraints ranging over sparse, group-sparse, block-sparse, low-rank, among many others (e.g. graph based) that all allow us to estimate high-dimensional models with guarantees. Is there an underlying commonality to them? It can be shown that all these structural constraints essentially impose the condition that the model parameters lie in some low-dimensional subspace from a specific collection of low-dimensional subspaces.

Example: Sparse Vectors. Consider the set of s-sparse vectors in p dimensions. For any particular subset $S \subseteq \{1, \ldots, p\}$, with cardinality s, define the subspace:

$$A(S) = \{\theta \in \mathbb{R}^p \, : \, \theta_j = 0, \quad \forall j \notin S\}.$$

It can then be seen an s-sparse vector lies in one of the collection of low-dimensional subspaces $\{A(S)\}_{S \subseteq [p]}$.

Example: Group-Sparse Vectors. Suppose that $\{1, \ldots, p\}$ can be partitioned into a set of T disjoint groups $\mathcal{G} = \{G_1, \ldots, G_T\}$. Given any subset $S_{\mathcal{G}} \subset [T]$ of the group indices, we can define the subspace:

$$A(S_{\mathcal{G}}) = \{\theta \in \mathbb{R}^p \, : \, \theta_{G_j} = 0, \quad \forall j \notin S_{\mathcal{G}}\}.$$

It can be again be seen that a group-sparse vector lies in a collection of these low-dimensional subspaces.

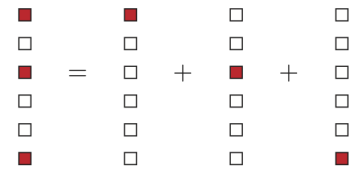
Low-Rank Matrices. For any matrix $\Theta \in \mathbb{R}^{p_1 \times p_2}$, let $\text{row}(\Theta) \in \mathbb{R}^{p_1}$ denote its row space, and $\text{col}(\Theta) \in \mathbb{R}^{p_2}$ its column space. For a given pair (U, V) or r-dimensional subspaces $U \subseteq \mathbb{R}^{p_1}$ and $V \subseteq \mathbb{R}^{p_2}$, we can define the subspaces:

$$A(U, V) = \{\Theta \in \mathbb{R}^{p_1 \times p_2} \, : \, \text{row}(\Theta) \subseteq U, \, \text{col}(\Theta) \subseteq V\}.$$

It can then be seen that any low-rank matrix $\Theta \in \mathbb{R}^{p_1 \times p_2}$ of rank $r \leq \min(p1, p_2)$ lies in a collection of the low-dimensional subspaces above.

(Chandrasekharan, Recht, Parillo, Willsky, 10) further refine this notion of structure as follows: If the "structured" parameter belongs to a low-dimensional subspace, then it can be written down as the sum of a small number of dictionary vectors of that low-dimensional subspace. One could call the dictionary elements of the low-dimensional subspaces as "atoms". Thus, a collection of low-dimensional subspaces is just a collection of simple "atoms" — which provides an alternative notion of structure. We briefly list some examples of such structured atoms.

Sparsity



Atoms are $\pm \mathbf{e}_j$

Low Rank

Atoms are rank one matrices $\mathbf{u}\mathbf{v}^\top$
(with $\mathbf{u}, \mathbf{v}$ having unit ℓ_2 norm)

Group Sparsity

Atoms are matrices with single non-zero row
(with unit ℓ_2 norm)

Permutations

Atoms are permutation matrices

5.1 Unified Analysis

Consider the following general statistical estimation problem: we have a set of probability distributions $\{P_\theta : \theta \in \Theta\}$, for some parameter space Θ. Suppose we are given n observations $\{Z_i\}_{i=1}^n \sim P_\theta^*(\cdot)$, and wish to estimate θ^* given these samples.

The general class of M-estimators solve an optimization problem of the form:

$$\widehat{\theta} \in \arg\min_\theta \left\{ \mathcal{L}(\theta; \{Z_i\}_{i=1}^n) + \lambda R(\theta) \right\},$$

where $\mathcal{L}(\theta, D)$ is some loss function measuring the goodness of fit of the candidate parameter θ to the data $D = \{Z_i\}_{i=1}^n$, and $R(\theta)$ is some regularization function that encodes the prior information about the true parameter, and encourages the candidate parameter θ to have specific structure. But would the regularization function be suited to the specific structural constraint we wish to impose? What would be the performance of the M-estimator above for general losses, and general regularization functions?

Surprisingly, one can analyze these general class of M-estimators with respect to general structured parameters via a very simple theorem. Before discussing this theorem, let us first introduce some terminology.

5.1.1 Decomposability

We will first introduce the property of decomposability, which captures whether a regularization function $R(\cdot)$ is "suited" to a class of structural constraints, as encoded by a collection of subspaces. We will say that a regularization function $R(\cdot)$ is **decomposable** with respect to a collection of subspaces $\mathcal{A}$ if for any $V \in \mathcal{A}$, any $u \in V$, and $u^\perp \in V^\perp$, it holds that:

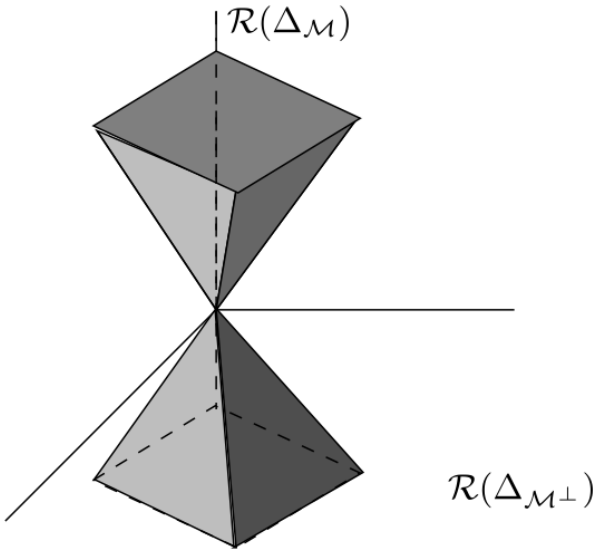
$$R(u + u^\perp) = R(u) + R(u^\perp).$$

Note that, assuming $R(\cdot)$ is some norm, by the triangle inequality, the RHS is always larger than the LHS. By requiring an equality, the decomposability property states that $R(\cdot)$ penalizes "unstructured" deviations $u^\perp$ (since they belong to $V^\perp$ for $V \in \mathcal{A}$) by the maximum possible.

Example: Sparse Vectors. The ℓ_1 norm is decomposable with respect to the sparse vector subspaces introduced earlier. To see this, note that any $\theta \in A(S)$ can be written as $\theta = (\theta_S, 0_{S^c})$, and any $\theta^\perp \in A(S)^\perp$ can be written as $\theta^\perp = (0_S, \theta^\perp_{S^c})$, so that:

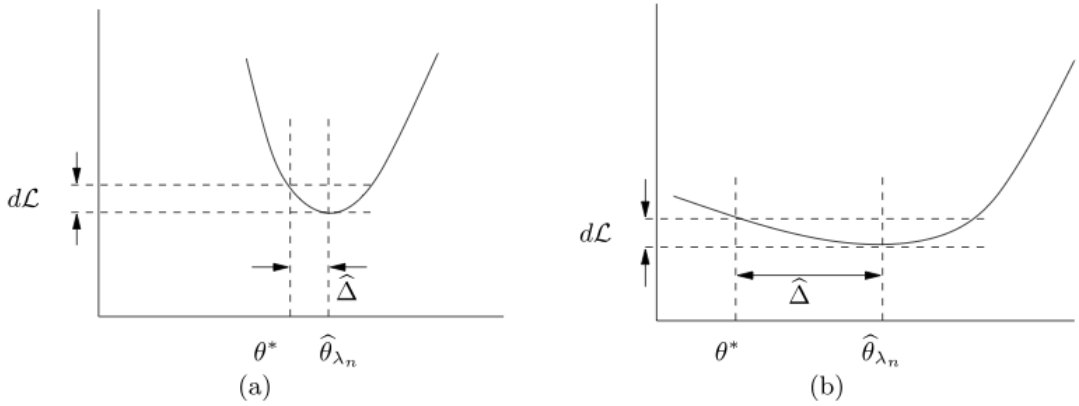
$$\|\theta + \theta^\perp\|_1 = \|(\theta_S, 0_{S^c}) + (0_S, \theta^\perp_{S^c})\|_1 = \|\theta\|_1 + \|\theta^\perp\|_1.$$

The key advantage of a decomposable regularizer when used within an M-estimator, is that one can then show that the error of the M-estimator will always lie in a cone, as shown in the following figure.

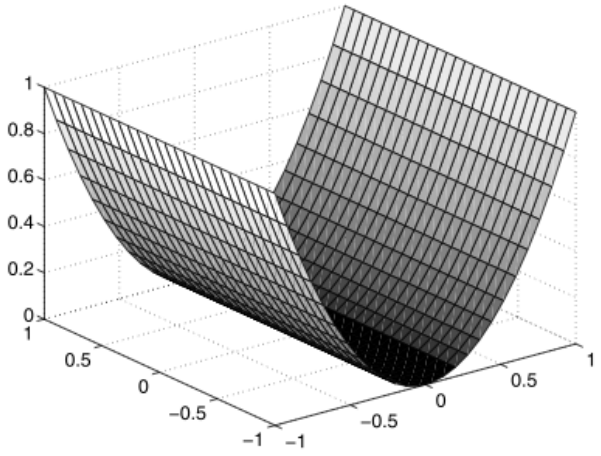


5.1.2 Restricted Strong Convexity

When we solve an M-estimation problem, we are minimizing in the range space of the loss functions. But suppose we require guarantees in our estimate of the parameter. Then, one should be able to translate small values of excess loss (from optimum) to small values of parameter error. But this requires curvature of the loss function, as can be seen by the figure below. For a loss with large curvature around the optimum, a small excess loss translates to small parameter error. But with flatter losses, a small excess loss could still entail a large parameter error.



But this is a problem in high-dimensions, because the empirical risk will always have flat directions, since the Hessian will be low-rank. For instance, with the squared loss in linear regression, the Hessian is $\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$, which is a $p \times p$ matrix with rank at most n, which in turn is smaller than p under high-dimensional sampling regimes. Thus, the Hessian is rank deficient, which means that it lack curvature at least along some directions. But on the positive side, it will have curvature at least along some directions, as in the following figure.



And we do not need curvature along directions in the first place: we only need curvature along directions the parameter estimate deviations $\widehat{\theta} - \theta^*$ can take, since that is all we need to translate excess loss to error in parameters. We just saw previously that the error of the M-estimator lies in a cone: so all we need is the guarantee that the loss has curvature along any direction in the cone. We call this property restricted strong convexity:

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \nabla\langle \mathcal{L}_n(\theta^*), \Delta\rangle \geq \gamma(\mathcal{L}_n)\, d^2(\Delta),$$

for all directions $\Delta \in \mathcal{C}$ in some restricted set $\mathcal{C}$, and where $d(\cdot)$ is the error norm with respect to which we desire to obtain error bounds.

5.1.3 Subspace Sizes.

Finally, we will measure the size of the subspace A in which the true parameter θ^* lies via the compability constant between the regularization function $R(\cdot)$ and the error norm $d(\cdot)$:

$$\Psi_{R,d}(A) = \min_{\theta^* \in A; d(\theta^*) \neq 0} \frac{R(\theta^*)}{d(\theta^*)}.$$

Example: Sparse Vectors. Suppose we use ℓ_1 regularization and wish to measure ℓ_2 error, so that $R(\theta) = \|\theta\|_1$, and $d(\Delta) = \|\Delta\|_2$. Then, for sparse vector subspaces $A(S) = \{\theta : \theta_{S^c} = 0\}$, we have:

$$\Psi_{\ell_1,\ell_2}(A(S)) = \sqrt{|S|}.$$

5.1.4 Noise Level.

The final ingredient we will need is a measure of the "noise" in the estimation problem, since that should naturally affect the parameter estimation error rates. It can be seen that the score function $\mathbb{E}(\nabla \mathcal{L}_n(\theta^*)) = 0$, when θ^* is the optimal parameter with respect to the expected loss $\mathbb{E}(\mathcal{L}_n(\theta))$. Thus, quantifying how far the sample estimate of the score function is from zero seems a reasonable approach to quantify the "noise level" in the estimation problem. We will specifically use:

$$R^*(\nabla \mathcal{L}_n(\theta^*)),$$

where $R^*(\cdot)$ is the dual norm of the regularization function $R(\cdot)$.

Example: Lasso. Consider the linear regression model: $Y = X\theta^* + W$, with $Y \in \mathbb{R}^n$, and $X \in \mathbb{R}^{n \times p}$. With Lasso, the loss function is the squared loss $\mathcal{L}_n(\theta) = \frac{1}{2n}\|Y - X\theta\|_2^2$, and the regularization function is the ℓ_1 norm $R(\theta) = \|\theta\|_1$, so that its dual norm is the ℓ_∞ norm: $R^*(\theta) = \|\theta\|_\infty$. We can then measure the noise level via:

$$\|\nabla \mathcal{L}_n(\theta^*)\|_\infty = \|\frac{1}{n} X^T W\|_\infty,$$

which when $W \sim N(0, 1)$, and with bounded columns $\|X_i\|_2^2 \leq C$, can be bounded as:

$$\|\nabla \mathcal{L}_n(\theta^*)\|_\infty \preceq \sqrt{\frac{\log p}{n}}.$$

5.1.5 Error Bounds

Suppose we wish to bound the error of the M-estimator $\widehat{\theta}$ from the true parameter θ^* with respect to some error norm $d(\cdot)$.

Theorem 2 *(Negahban, Ravikumar, Wainwright, Yu 2012) Suppose the regularization parameter is set so that $\lambda_n \geq 2R^*(\nabla \mathcal{L}_n(\theta^*))$, the regularization function satisfies the decomposability condition with respect a collection of subspaces $\mathcal{A}$, the true parameter $\theta^* \in A$ for some $A \in \mathcal{A}$, and the loss satisfies the restricted strong convexity condition with parameter $\gamma(\mathcal{L})$. Then:*

$$d(\widehat{\theta} - \theta^*) \preceq \frac{1}{\gamma(\mathcal{L})} \Psi(A) \lambda_n.$$

Example: Sparse Linear Regression Model. Suppose $Y = X\theta^* + W$, with $W \sim N(0, \sigma^2 I)$, and $\|\theta^*\|_0 = k$. Consider the Lasso estimator:

$$\widehat{\theta} \in \arg\min_\theta \left\{ \frac{1}{2n}\|Y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}.$$

The restricted strong convexity condition reduces to a restricted eigenvalue condition on $\frac{1}{n}X^TX$. The noise level can be bound as

$$\|\frac{1}{n}X^TW\|_\infty \preceq \sqrt{\frac{\sigma^2 \log p}{n}},$$

with high probability, so that by setting $\lambda_n \asymp \sqrt{\frac{2\sigma^2 \log p}{n}}$, we have that with high probability:

$$\|\widehat{\theta} - \theta^*\|_2 \preceq \frac{\sigma}{\gamma(\mathcal{L})}\sqrt{\frac{k \log p}{n}}.$$

Example: Low Rank Matrices. Consider a matrix-structured linear regression model: $Y_i = \langle X_i, \Theta^* \rangle + W_i$, with $W_i \sim N(0, \sigma^2)$, for $i \in [n]$. Suppose that the true parameter $\Theta^* \in \mathbb{R}^{k \times m}$ has rank $r \ll \min(k, m)$, so that we solve for the nuclear norm regularized least squares estimator:

$$\widehat{\Theta} \in \arg\min_{\Theta} \left\{ \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \langle X_i, \Theta \rangle^2 + \lambda_n |||\Theta|||_* \right\}.$$

Then, the noise level can be bounded as

$$|||\frac{1}{n}X_iW_i|||_2 \preceq \sigma\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{m}{n}}\right),$$

with high probability. Thus, assuming the design matrix X satisfies the restricted strong convexity condition, and by setting the regularization parameter as $\lambda_n \asymp \sigma\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{m}{n}}\right)$, we have with high probability:

$$|||\widehat{\Theta} - \Theta^*||| \preceq \frac{\sigma}{\gamma(\mathcal{L})}\left(\sqrt{\frac{rk}{n}} + \sqrt{\frac{rm}{n}}\right).$$

References

Buhlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer.

Candes, E. J. & Plan, Y. (2009), 'Near ideal model selection by ℓ_1 minimization', *Annals of Statistics* **37**(5), 2145–2177.

Candes, E. J. & Tao, T. (2006), 'Near optimal signal recovery from random projections: Universal encoding strategies?', *IEEE Transactions on Information Theory* **52**(12), 5406–5425.

Chatterjee, S. (2013), Assumptionless consistency of the lasso. arXiv: 1303.5817.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Transactions on Information Theory* **52**(12), 1289–1306.

Fuchs, J. J. (2005), 'Recovery of exact sparse representations in the presense of bounded noise', *IEEE Transactions on Information Theory* **51**(10), 3601–3608.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**(6), 971–988.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, Chapman & Hall.

Meinshausen, N. & Buhlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* **34**(3), 1436–1462.

Wainwright, M. (2017), *High-Dimensional Statistics: A Non-Asymptotic View*, Cambridge University Press. To appear.

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2564.