3.4 Penalized estimators

Adaptivity

Theorem 3.3.3 and 3.3.4 are very attractive since they provide fast decrease of the mean squared error in high dimensional settings. However, they require the knowledge of properties of the unknown θ^*. It is possible to produce adaptive estimators which do not require such knowledge.

Consider the sub-gaussian sequence model: $y = \theta^* + \xi \in \mathbb{R}^d$, where $\xi \sim \mathrm{subG}(\sigma^2/n)$. This allows to capture the intuition about penalization. Using Theorem 2.2.1 and Remark 2.2.1, we have for any $\delta > 0$, with probability at least $1-\delta$

$$\max_{1\le i\le d} |\xi_i| \le \sigma\sqrt{\frac{2\log(2d/\delta)}{n}} = \tau.$$

If $|y_j| \gg \tau$ for some j, then it must correspond to $\theta^*_j \neq 0$. On the other hand, if $|y_j| \le \tau$, then $|\theta^*_j| \le |y_j| + |\xi_j| \le 2\tau$ with high probability. This motivates the use of the following estimator, called the hard-thresholding estimator:

$$\hat\theta^{HT}_j = y_j \mathbb{I}(|y_j| \ge 2\tau), \quad j = 1, \dots, d.$$

Indeed, conditioning on the event:

$$\mathcal{A} = \left\{\max_i |\xi_i| \le \tau\right\},$$

we have for all j, $|y_j| \ge 2\tau \Rightarrow |\theta^*_j| \ge \tau$ and $|y_j| \le 2\tau \Rightarrow |\theta^*_j| \le 3\tau$ and

$$\begin{aligned}
\|\hat\theta^{RT} - \theta^*\|^2 &= \sum_{i=1}^d \left(|y_i - \theta^*_i|\mathbb{I}(|y_i| \ge 2\tau) + |\theta^*_i|\mathbb{I}(|y_i| < 2\tau)\right)^2 \\
&\le \sum_{i=1}^d \left(\tau\mathbb{I}(|\theta^*_i| \ge \tau) + (\theta^*_i)\mathbb{I}(|\theta^*_i| < 3\tau)\right)^2 \\
&\le \sum_{i=1}^d \left(4\min\left\{|\theta^*_j|^2, \tau\right\}\right)^2 \le 16\|\theta^*\|_0\tau^2 = \frac{32\|\theta\|_0\sigma^2\log(2d/\delta)}{n}.
\end{aligned}$$

Furthermore, if $\min_{j\in\mathrm{supp}(\theta^*)} |\theta^*_j| \ge 3\tau$, then $\mathrm{supp}(\hat\theta^{HT}) = \mathrm{supp}(\theta^*)$.

It turns out that $\hat\theta^{HT}$ is obtained by penalization using ℓ_0 pseudo norm ball:

$$\hat\theta^{HT} = \arg\min_{\theta\in\mathbb{R}^d} \|y-\theta\|^2 + 4\tau^2\|\theta\|_0.$$

This is easily seen as if $|y_i| < 2\tau$ for some j, then $4\tau^2\mathbb{I}(\theta_j \neq 0) > y_j^2$. This motivates the use of penalized estimators which are more adaptative to unknown properties of θ^*.

Under model (LM), we set, for any $\lambda \ge 0$,

$$\begin{aligned}
\hat\theta^{\ell_0} &\in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_0 \\
\hat\theta^{\ell_1} &\in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1
\end{aligned}$$

The second estimator is commonly called the Lasso estimator.

ℓ_0 penalized least squares

Theorem 3.4.1. *Assume that model* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$ *then choosing* $\lambda = 8\log(6)\sigma^2/n + 16\sigma^2\log(ed)/n$*, we have for any* $\delta > 0$ *with probability at least* $1-\delta$*,*

$$\mathrm{MSE}(\hat\theta^{\ell_0}) \le \frac{32\sigma^2\left(2\|\theta^*\|_0\left(\log(6)+\log(ed)\right)+\log(1/\delta)+\log(2)\right)}{n}$$

Proof. We have by definition

$$\frac{1}{2n}\|\mathbb{X}\hat\theta^{\ell_0} - Y\|^2 + \lambda\|\hat\theta^{\ell_0}\|_0 \le \frac{1}{2n}\|\mathbb{X}\theta^* - Y\|^2 + \lambda\|\theta^*\|_0.$$

Similarly as in Lemma 3.3.2, we have

$$\|\mathbb{X}\hat\theta^{\ell_0} - \mathbb{X}\theta^*\|^2 \le 2\epsilon^T\mathbb{X}\left(\hat\theta^{\ell_0}-\theta^*\right) + 2n\lambda(\|\theta^*\|_0 - \|\hat\theta^{\ell_0}\|_0).$$

For any $a, b \in \mathbb{R}^d$, we have

$$2a^Tb = 2a^T\frac{b}{\|b\|_2}\|b\|_2 \le 2\left(a^T\frac{b}{\|b\|_2}\right)^2 + \frac{1}{2}\|b\|_2^2,$$

and hence

$$\|\mathbb{X}\hat\theta^{\ell_0} - \mathbb{X}\theta^*\|^2 \le 4\left(\frac{\epsilon^T\mathbb{X}\left(\hat\theta^{\ell_0}-\theta^*\right)}{\left\|\mathbb{X}\left(\hat\theta^{\ell_0}-\theta^*\right)\right\|_2^2}\right)^2 + 4n\lambda(\|\theta^*\|_0 - \|\hat\theta^{\ell_0}\|_0). \qquad (3.3)$$

Setting $\mathcal{U}(\hat\theta^{\ell_0}-\theta^*) = \mathbb{X}\left(\hat\theta^{\ell_0}-\theta^*\right)/\left\|\mathbb{X}\left(\hat\theta^{\ell_0}-\theta^*\right)\right\|_2$, we have

$$\begin{aligned}\left(\epsilon^T\mathcal{U}(\hat\theta^{\ell_0}-\theta^*)\right)^2 - n\lambda\|\hat\theta^{\ell_0}\|_0 &\le \sup_{\theta\in\mathbb{R}^d}\left(\epsilon^T\mathcal{U}(\theta-\theta^*)\right)^2 - n\lambda\|\theta\|_0\\ &\le \max_{1\le k\le d}\max_{|S|=k}\sup_{\mathrm{supp}(\theta)=S}\left(\epsilon^T\mathcal{U}(\theta-\theta^*)\right)^2 - n\lambda k\\ &\le \max_{1\le k\le d}\max_{|S|=k}\sup_{u\in\mathbb{R}^{r_{S*}},\|u\|_2\le 1}\left(\epsilon^T\Phi_{S*}u\right)^2 - n\lambda k\end{aligned}$$

where $\Phi_{S*} \in \mathbb{R}^{n\times r_{S*}}$ denotes an orthonormal basis of the span of the columns of $\mathbb{X}$ indexed by $S \cup \mathrm{supp}(\theta^*)$, and $r_{S*} \le |S| + \|\theta^*\|_0$. For any $t > 0$, k and S with $|S| = k$, we have using Theorem 2.2.2.

$$\begin{aligned}\mathbb{P}\left[4\sup_{u\in\mathbb{R}^{r_{S*}},\|u\|_2\le 1}\left(\epsilon^T\Phi_{S*}u\right)^2 - 4n\lambda k > t\right] &= \mathbb{P}\left[\sup_{u\in\mathbb{R}^{r_{S*}},\|u\|_2\le 1}\left\|\epsilon^T\Phi_{S*}u\right\| > \sqrt{\frac{t}{4}+n\lambda k}\right]\\ &\le 2\cdot 6^{r_{S*}}\exp\left(-\frac{\frac{t}{4}+n\lambda k}{8\sigma^2}\right)\\ &\le 2\exp\left(-\frac{t}{32\sigma^2} - \frac{n\lambda k}{8\sigma^2} + (k+\|\theta^*\|_0)\log(6)\right). \qquad (3.4)\end{aligned}$$

Using the definition of λ, we have

$$\begin{aligned}-\frac{n\lambda k}{8\sigma^2} + (k+\|\theta^*\|_0)\log(6) &= -k\log(6) - 2k\log(ed) + (k+\|\theta^*\|_0)\log(6)\\ &= -2k\log(ed) + \|\theta^*\|_0\log(6).\end{aligned}$$

Using a union bound with (3.3) and (3.4), we obtain, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\|\mathbb{X}\hat{\theta}^{\ell_0} - \theta^*\|_2^2 \geq 4n\lambda\|\theta^*\|_0 + t\right] \\
&\leq \sum_{k=1}^{d}\sum_{|S|=k} 2\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(6)\right) \\
&\leq 2\sum_{k=1}^{d}\binom{d}{k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(6)\right) \\
&\leq 2\sum_{k=1}^{d}\exp\left(-\frac{t}{32\sigma^2} - k\log(ed) + \|\theta^*\|_0\log(6)\right) && \text{Lemma 3.3.2} \\
&\leq 2\sum_{k=1}^{d}(ed)^{-k}\exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right) \\
&\leq 2\exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right)
\end{aligned}$$

Choosing $t = 32\sigma^2\left(\log(1/\delta) + \|\theta^*\|_0\log(6) + \log(2)\right)$, the right hand side is equal to δ and we obtain that with probability $1-\delta$,

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\ell_0} - \mathbb{X}\theta^*\|_2^2 &\leq 4n\lambda\|\theta^*\|_0 + t \\
&= 32\sigma^2\left(\|\theta^*\|_0\left(\log(6) + 2\log(ed)\right) + \left(\log(1/\delta) + \|\theta^*\|_0\log(6) + \log(2)\right)\right) \\
&= 32\sigma^2\left(2\|\theta^*\|_0\left(\log(6) + \log(ed)\right) + \log(1/\delta) + \log(2)\right)
\end{aligned}$$

□

This is a very strong result as it provides an estimator which completely adapts to unknwon support, including its size.

ℓ_1 penalized least squares

Theorem 3.4.2. *Assume that model* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$. *Moreover assume that the columns of* $\mathbb{X}$ *have norm at most* $\sqrt{n}$. *Then, for any* $\delta > 0$, *choosing* $\lambda = 2\sigma/\sqrt{n}\left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right)$, *we have for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$\mathrm{MSE}(\hat{\theta}^{\ell_1}) \leq \frac{4\|\theta^*\|_1\sigma}{\sqrt{n}}\left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right).$$

Proof. We have by definition

$$\frac{1}{2n}\|\mathbb{X}\hat{\theta}^{\ell_1} - Y\|^2 + \lambda\|\hat{\theta}^{\ell_1}\|_1 \leq \frac{1}{2n}\|\mathbb{X}\theta^* - Y\|^2 + \lambda\|\theta^*\|_1.$$

Similarly as in Lemma 3.3.2, we have

$$\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|^2 \leq 2\epsilon^T\mathbb{X}\left(\hat{\theta}^{\ell_1} - \theta^*\right) + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1).$$

Hölder's inequality states that for any $a, b \in \mathbb{R}^d$, we have $a^Tb \leq \|a\|_\infty\|b\|_1$, and hence

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|_2^2 &\leq 2\|\epsilon^T\mathbb{X}\|_\infty\left(\|\hat{\theta}^{\ell_1}\|_1 + \|\theta^*\|_1\right) + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1) && (3.5) \\
&= \|\hat{\theta}^{\ell_1}\|_1(2\|\epsilon^T\mathbb{X}\|_\infty - \lambda n) + \|\theta^*\|_1(2\|\epsilon^T\mathbb{X}\|_\infty + \lambda n). && (3.6)
\end{aligned}$$

Now for any $t > 0$, and any column $\mathbb{X}_j$ of $\mathbb{X}$, we have that $\mathbb{X}_j^T \epsilon \sim \mathrm{subG}(\sigma^2 n)$ and from Theorem 3.4.2

$$\mathbb{P}\left[\|\mathbb{X}^T\epsilon\|_\infty > t\right] \leq 2de^{-\frac{t^2}{2n\sigma^2}}.$$

Taking $t = \sigma(\sqrt{2n\log(2d)} + \sqrt{2n\log(1/\delta)}) = n\lambda/2$, we obtain using (3.6), that with probability $1-\delta$,

$$\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|_2^2 \leq 2n\lambda\|\theta^*\|_1.$$

□

3.5 Incoherence and fast rates for Lasso

Incoherence, random matrices and cone condition

Definition 3.5.1. *A matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$ is said to have incoherence $k \in \mathbb{N}^*$, if*

$$\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right\|_\infty \leq \frac{1}{32k},$$

where $\|\cdot\|_\infty$ denotes the largest absolute value of a matrix.

For $k \to \infty$ this entails that $\mathbb{X}$ is orthonormal and prevents situations where $d > n$. However, finite values of k, amount to relax this constraint and allow for much larger d.

Proposition 3.5.1. *Let $\mathbb{A} \in \mathbb{R}^{n\times d}$ be a random matrix which entries are independent Rademacher variables (± 1 with probability $1/2$). Then, for any $\delta > 0$, if $n \geq 2^{11}k^2\log(1/\delta) + 2^{13}k^2\log(d)$, with probability $1-\delta$ over the random draw of its entries, $\mathbb{A}$ has incoherence k.*

Proof. The diagonal entries of $\mathbb{A}^T\mathbb{A}$ are equal to n and the off-diagonal elements are sum of n independant Rademacher random variables. From Hoeffding's lemma (2.1.1), Rademarcher random variables are sub gaussian with variance proxy 1 and using Theorem 2.1.2, their sum is $\mathrm{subG}(n)$. Using a union bound, we have, for any $t \geq 0$, using Theorem 2.1.1 and summing over the d^2 entries of $\mathbb{A}^T\mathbb{A}$,

$$\mathbb{P}\left[\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right\|_\infty > t\right] \leq 2d^2e^{\frac{-nt^2}{2}}.$$

Choosing $t = 1/(32k)$, we have

$$\mathbb{P}\left[\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right\|_\infty > \frac{1}{32k}\right] \leq e^{\log(2)+2\log(d)-\frac{n}{2^{11}k^2}} \leq \delta,$$

for the choice of n which has been made.

□

The k^2 term can actually be improved to k. For any $\theta \in \mathbb{R}^d$, $S \subset \{1,\dots,d\}$, we denote by θ_S, the vector which support is S and which entries agree with those of θ on S. We have $\|\theta\|_1 = \|\theta_S\|_1 + \|\theta_{S^c}\|_1$.

Lemma 3.5.1. *For any $k \leq d$ and $\mathbb{X}$ having incoherence k, any S with $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ satisfying the cone condition:*

$$\|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1,$$

we have $\|\theta\|_2^2 \leq 2\frac{\|\mathbb{X}\theta\|_2^2}{n}$.

Proof. We have $\theta = \theta_S + \theta_{S^c}$, and hence

$$\|\mathbb{X}\theta\|_2^2 = \|\mathbb{X}\theta_S\|_2^2 + \|\mathbb{X}\theta_{S^c}\|_2^2 + 2\theta_S\mathbb{X}^T\mathbb{X}\theta_{S^c}.$$

From the incoherence condition, since $\|\theta_S\|_0 \le k$, we have

$$\|\mathbb{X}\theta_S\|_2^2 = n\|\theta_S\|_2^2 + n\theta_S^T\left(\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right)\theta_S \ge n\|\theta_S\|_2^2 - n\frac{\|\theta_S\|_1^2}{32k}.$$

This also holds for θ_{S^c} and using the cone condition, we obtain

$$\|\mathbb{X}\theta_{S^c}\|_2^2 \ge n\|\theta_{S^c}\|_2^2 - n\frac{\|\theta_{S^c}\|_1^2}{32} \ge n\|\theta_{S^c}\|_2^2 - 9n\frac{\|\theta_S\|_1^2}{32k}.$$

Using the incoherence property again as well as Hölder's inequality, we obtain

$$2\left|\theta_S^T\mathbb{X}^T\mathbb{X}\theta_{S^c}\right| \le \frac{2}{32k}\|\theta_S\|_1\|\theta_{S^c}\|_1 \le \frac{6}{32k}\|\theta_S\|_1^2.$$

Finnally, from Cauchy-Schwartz inequality, one has $\|\theta_S\|_1^2 \le |S|\|\theta_S\|_2^2 \le k\|\theta_S\|_2^2$ and

$$\frac{\|\mathbb{X}\theta\|_2^2}{n} \ge \|\theta_S\|_2^2 + \|\theta_{S^c}\|_2^2 - \frac{16|S|\|\theta_S\|_2^2}{32k} \ge \frac{\|\theta_S\|_2^2}{2}.$$

□

Fast rate for the Lasso estimator

Theorem 3.5.1. *For $n \neq 2$, assume that model LM holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$. Assume that $\|\theta_0\|_0 \le k$ and that $\mathbb{X}$ has incoherence k. Then, for any $\delta > 0$, the Lasso estimator $\hat{\theta}^{\ell_1}$ with $\lambda = 8\sigma/n(\sqrt{\log(2d)} + \sqrt{\log(1/\delta)})$ satisfies with probability $1-\delta$*

$$\mathrm{MSE}(\hat{\theta}^{\ell_1}) \le (2^{12})\frac{k\sigma^2\log(2d/\delta)}{n}$$
$$\|\hat{\theta}^{\ell_1} - \theta^*\|_2^2 \le (2^{13})\frac{k\sigma^2\log(2d/\delta)}{n}$$

Proof. We have by definition

$$\frac{1}{2n}\|\mathbb{X}\hat{\theta}^{\ell_1} - Y\|^2 \le \frac{1}{2n}\|\mathbb{X}\theta^* - Y\|^2 + \lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1).$$

and similarly as in Lemma 3.3.1,

$$\|\mathbb{X}\hat{\theta}^{\ell_1} - Y\|^2 + n\lambda\|\hat{\theta}^{\ell_1} - \theta^*\|_1 \le 2\epsilon^T\mathbb{X}(\hat{\theta}^{\ell_1} - \theta^*) + n\lambda\|\hat{\theta}^{\ell_1} - \theta^*\|_1 + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1).$$

Similarly as in the proof of Theorem 3.4.2, $\mathbb{X}$ has columns satisfying $\|\mathbb{X}_j\|_2^2 \le n + \frac{1}{32k} \le 2n$ from the incoherence condition. Hence, for any $t > 0$,

$$\mathbb{P}\left[\|\mathbb{X}^T\epsilon\|_\infty > t\right] \le 2de^{-\frac{t^2}{4n\sigma^2}}.$$

Taking $t = 2\sigma(\sqrt{n\log(2d)} + \sqrt{n\log(1/\delta)}) = n\frac{\lambda}{4}$, the right hand side is smaller than δ, and we obtain that with probability $1-\delta$,

$$\begin{aligned}\epsilon^T\mathbb{X}(\hat{\theta}^{\ell_1} - \theta^*) &\le \|\mathbb{X}^T\epsilon\|_\infty\|\hat{\theta}^{\ell_1} - \theta^*\|_1 \\ &\le \frac{n\lambda}{4}\|\hat{\theta}^{\ell_1} - \theta^*\|_1.\end{aligned}$$

Setting S the support of θ^* and noting that $\|\hat\theta^{\ell_1}-\theta^*\|_1-\|\hat\theta^{\ell_1}\|_1=\|\hat\theta^{\ell_1}_S-\theta^*\|_1-\|\hat\theta^{\ell_1}_S\|_1$, we obtain, with probability $1-\delta$

$$\begin{aligned}\|\mathbb{X}\hat\theta^{\ell_1}-Y\|^2+n\lambda\|\hat\theta^{\ell_1}-\theta^*\|_1 &\le 2n\lambda\|\hat\theta^{\ell_1}-\theta^*\|_1+2n\lambda(\|\theta^*\|_1-\|\hat\theta^{\ell_1}\|_1) &(3.7)\\ &\le 2n\lambda\|\hat\theta^{\ell_1}_S-\theta^*\|_1+2n\lambda(\|\theta^*\|_1-\|\hat\theta^{\ell_1}_S\|_1) &(3.8)\\ &\le 4n\lambda\|\hat\theta^{\ell_1}_S-\theta^*\|_1. &(3.9)\end{aligned}$$

In particular, we have

$$\|\hat\theta^{\ell_1}_{S^c}-\theta^*_{S^c}\|_1\le 3\|\hat\theta^{\ell_1}_S-\theta^*\|_1$$

which is the cone condition of Lemma 3.5.1. Using this and Cauchy-Schwartz inequality, we obtain

$$\|\hat\theta^{\ell_1}_S-\theta^*\|_1\le\sqrt{|S|}\|\hat\theta^{\ell_1}_S-\theta^*\|_2\le\sqrt{|S|}\|\hat\theta^{\ell_1}-\theta^*\|_2\le\sqrt{\frac{2k}{n}}\left\|\mathbb{X}\left(\hat\theta^{\ell_1}-\theta^*\right)\right\|_2.$$

Combining with (3.9), we have

$$\left\|\mathbb{X}\left(\hat\theta^{\ell_1}-\theta^*\right)\right\|_2^2\le 32nk\lambda^2\le(2^{12})k\sigma^2\log(2d/\delta).$$

The secon inequality follows because from Lemma 3.5.1, we have $\|\hat\theta^{\ell_1}-\theta^*\|_2^2\le 2\mathrm{MSE}(\hat\theta^{\ell_1})$. □

For the proof, we only used Lemma 3.5.1 and more precisely

$$\inf_{|S|\le k}\inf_{\theta\in C_S}\frac{\|\mathbb{X}\theta\|_2^2}{n\|\theta\|_2^2}\ge\frac12,$$

where C_S is the cone defined by $\|\theta_{S^c}\|_1\le 3\|\theta_S\|_1$. This condition is called restricted eigenvalue condition. It can be seen as a lower bound on the eigenvalues of $\mathbb{X}$ when restricted to sparse eigen vectors. In particular it implies that the smallest singular value of $\mathbb{X}_S$ is at least $n/2$ for all $|S|\le k$. To summarize, Proposition 3.5.1 and Theorem 3.5.1 en sure that there exists design matrices $\mathbb{X}$ such that the Lasso estimators has a fast convergence rate in high dimensions.

3.6 Compressed sensing

High dimentional statistics have an important intersection with compressed sensing [5, 3] in signal processing. Traditional approaches separate signal aquisition and signal compression which is performed on a signal which is fully characterized in the memory of a device (or at least very accurately described). The field of compressed sensing emerge as different approach for this problem based on two observations.

- Natural signals such as speach, sounds, images, are not generic or completely random and they have a strong intrinsic strucutre.
- If this structure was known it should be possible to take advantage in a signal aquisition / compression scheme.

Compressed sensing emerged as a development of the preceeding observation based on two ideas.

- the undelying structure of natural signals is captured by sparsity patterns in a certain basis.
- if a signal is sparse in a given basis, one could probably mix the aquisition and compression phase by aquiring only a very limited number of measurements.

We describe a signal recovery result from random measurements relying on linear programming. Further readings on the topic include [2, 3, 4].

Signal recovery

Althouth the notations will be the same as in the high dimensional statistics context, the viewpoint is a bit different. The signal to be recovered is $\theta^* \in \mathbb{R}^d*$ which is unknown and assumed to be sparse, that is $\|\theta^*\|_0 = k < d$. The operator has the possibility to choose a sensing matrix $\mathbb{X} \in \mathbb{R}_{n\times d}$ which will result in the following measurements:

$$\mathbb{X}\theta^* = y \tag{3.10}$$

The goal of compressed sensing is to establish methods and conditions ensuring large classes of values of θ^* can be infered accuratly only from the knowledge of y and $\mathbb{X}$. Other questions of interest include robustness to noise and exact recovery of $\mathrm{supp}(\theta^*)$. For simplicity we will only touch the noiseless setting in (3.10). We will deduce compressed sensing type results from MSE estimates of the previous sections.

Exact recovery using ℓ_0 minimization

We introduce the estimator

$$\hat{\theta}^{\ell_0}_{CS} \in \min_{\theta\in\mathbb{R}_d} \quad \|\theta\|_0 \quad \text{s.t.} \quad \mathbb{X}\theta = y. \tag{3.11}$$

under mild assumption on the sensing matrix $\mathbb{X}$, this estimator deterministically recovers the unknown signal θ^*.

Proposition 3.6.1. *Given $k \in \mathbb{N}$, $k \le d$, and assume that for any S, $|S| \le 2k$, that $\mathbb{X}_S$ has full column rank. Then, the solution of* (3.11) *is unique and is equal to θ^*.*

Proof. Assume that $\hat{\theta}^{\ell_0}_{CS} \neq \theta^*$. We have $\|\hat{\theta}^{\ell_0}_{CS}\|_0 \le \|\theta^*\|_0 = k$. Set $S = \mathrm{supp}(\theta^* - \hat{\theta}^{\ell_0}_{CS})$. We have $|S| \le 2k$ and $\mathbb{X}(\theta^* - \hat{\theta}^{\ell_0}_{CS}) = 0$ and hence $\theta^* = \hat{\theta}^{\ell_0}_{CS}$. □

Exact recovery from random measurements with ℓ_1 minimization

Intuitively if one is interested in signal recovery over large classes of signals using ℓ_1 norm, the sensing matrix in (3.10) should not have structure fooling the ℓ_1 norm. This happens if $\mathbb{X}$ is generic in some sense. One way to achieve this is to use random measurements. This amounts to choose a random $\mathbb{X}$ in (3.10) such as the one described in Proposition 3.5.1 for example. Furthermore, since there is no noise, in the measurements, the least squares approach does not really make sense. We introduce an estimator.

$$\hat{\theta}^{\ell_1}_{CS} \in \min_{\theta\in\mathbb{R}_d} \quad \|\theta\|_1 \quad \text{s.t.} \quad \mathbb{X}\theta = y. \tag{3.12}$$

Corollary 3.6.1. *Given $k \in \mathbb{N}$, $k \le d$, and $\delta > 0$, assume that $\mathbb{X}$ is a Rademacher matrix with $n \ge 2^{11}k^2\log(1/\delta) + 2^{13}k^2\log(d)$. Assume furthermore that $\|\theta^*\|_0 \le k$ in* (3.10)*. Then with probability $1-\delta$ over the random draw of $\mathbb{X}$, the solution of* (3.12) *is unique and is equal to θ^*.*

Proof. Assume that $\hat{\theta}^{\ell_1}_{CS} \neq \theta^*$ and set $d = \hat{\theta}^{\ell_1}_{CS} - \theta^*$. We have $\|\hat{\theta}^{\ell_1}_{CS}\|_1 \le \|\theta^*\|_1$ and $\mathbb{X}d = 0$. Set $S = \mathrm{supp}(\theta^*)$, we have

$$\|\theta^*\|_1 \ge \|\hat{\theta}^{\ell_1}_{CS}\|_1 = \|d_{S^c}\|_1 + \|d_S + \theta^*\|_1 \ge \|d_{S^c}\|_1 + \|\theta^*\|_1 - \|d_S\|_1.$$

As a result, we have $\|d_S\|_1 \ge \|d_{S^c}\|_1$ and $\mathbb{X}d = 0$. Lemma 3.5.1 implies that $d = 0$. □

This result shows that it is possible to recover θ^* with high probability only from the order of $O(k^2\log(d))$ measurements provided that $\|\theta^*\|_0 \le k$. The k^2 term can be improved further. In the context of noisy measurements, conditioning both on the realization of $\mathbb{X}$ and the realization of the noise, one can obtain results similar to Theorem 3.5.1 for signal processing.

Exercises

Exercise 3.6.1. *Given $x \in \mathbb{R}^d$ and $\lambda > 0$, show that the solution to the problem*

$$\min_{y \in \mathbb{R}^p} \frac{1}{2}\|y - x\|_2^2 + \lambda\|y\|_1$$

is given by coordinatwise application of $p_\lambda \colon \mathbb{R} \mapsto \mathbb{R}$ to x, where, for any $s \in \mathbb{R}$

$$p_\lambda(s) = \begin{cases} s - \lambda, & \textit{if } s > \lambda \\ 0, & \textit{if } |s| \le \lambda \\ s + \lambda, & \textit{if } s < \lambda \end{cases} .$$

This is the soft-thresholding operation and the result is called the proximity operator of the function $\lambda\|\cdot\|_1$. Give a graphical representation of p_λ and compare it to the hard-thresholding operator given by $t \mapsto t\mathbb{I}(|t| \ge \lambda)$.

Exercise 3.6.2. *Let $X = (1, Z, \dots, Z^d)^T \in \mathbb{R}^{d+1}$ be a random vector where Z is a real random variable. Show that $\mathbb{E}\left[XX^T\right] \in \mathbb{R}^{d+1\times d+1}$ is positive definite when Z admits a density with respect to Lebesgue measure on $\mathbb{R}$. Provide a counter example for which $\mathbb{E}\left[XX^T\right]$ is singular.*

Exercise 3.6.3. *Under the linear model* (LM),

- *Assuming that $\mathbb{X}^T\mathbb{X}$ is invertible and $\mathbb{E}[\epsilon] = 0$, show that $\mathbb{E}[\theta_{LS}] = \theta^*$.*
- *Assuming in addition that $\epsilon \sim \mathrm{subG}(\sigma^2)$, show that $\theta_{LS} - \theta^* \sim \mathrm{subG}\left(\frac{\sigma^2}{\lambda_{\min}^2}\right)$ where $\lambda_{\min}$ denotes the smallest eigenvalue of $\mathbb{X}^T\mathbb{X}$. Propose a generalization of the result when the invertibility assumption is dropped.*
- *If $\mathbb{X}^T\mathbb{X}$ is not invertible, show that $\theta_{LS} = \arg\min_\theta \|\theta\|_2$, such that $\mathbb{X}^T\mathbb{X}\theta = \mathbb{X}^T Y$.*

Exercise 3.6.4. *We consider the model* (LM), *and define the ridge regression estimator, for any $\lambda > 0$*

$$\hat{\theta}^{\ell_2} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_2^2.$$

- *Show that $\hat{\theta}^{\ell_2}$ is indeed uniquely defined and propose a closed form expression for it.*
- *Compute the bias: $\mathbb{E}\left[\hat{\theta}^{\ell_2} - \theta^*\right]$ and show that it is bounded by $\|\theta^*\|_2^2$.*
- *Show that $\hat{\theta}^{\ell_2} - \mathbb{E}\left[\hat{\theta}^{\ell_2}\right] \sim \mathrm{subG}\left(\frac{\sigma^2}{\lambda^2}\right)$.*
- *Show the bias variance decomposition identity:*

$$\mathbb{E}\left[\|\hat{\theta}^{\ell_2} - \theta^*\|_2^2\right] = \mathbb{E}\left[\left\|\hat{\theta}^{\ell_2} - \mathbb{E}\left[\hat{\theta}^{\ell_2}\right]\right\|_2^2\right] + \left\|\mathbb{E}\left[\hat{\theta}^{\ell_2} - \theta^*\right]\right\|_2^2.$$

- *Using the previous exercise, suggest as situation for which*

$$\mathbb{E}\left[\|\hat{\theta}^{\ell_2} - \theta^*\|_2^2\right] < \mathbb{E}\left[\|\hat{\theta}^{LS} - \theta^*\|_2^2\right]$$