# Statistics 210B Lecture 21 Notes 

Daniel Raban
April 7, 2022

## 1 LASSO Prediction Error Bound and High-Dimensional Principal Component Analysis

### 1.1 Recap: overview of results for noisy, sparse linear regression

Let's finish up our analysis of noisy, sparse linear regression. Our model is $y=X \theta^{*}+w \in$ $\mathbb{R}^{n}$, where

$$
w \in \mathbb{R}^{n}, \quad X=\left[\begin{array}{c}
x_{1}^{\top} \\
\vdots \\
x_{n}^{\top}
\end{array}\right] \in \mathbb{R}^{n \times d}, \quad \theta^{*} \in \mathbb{R}^{d}, \quad\left|S\left(\theta^{*}\right)\right| \leq s
$$

We looked at the $\lambda$ formulation of the LASSO problem, where

$$
\widehat{\theta} \in \underset{\theta \in \mathbb{R}^{d}}{\arg \min } \frac{1}{2 n}\|y-X \theta\|_{2}^{2}+\lambda_{n}\|\theta\|_{1} .
$$

We also looked at the 1-norm constrained and error-constrained formulations of the problem. We defined the $\mathbb{C}_{\alpha}$ cone

$$
\mathbb{C}_{\alpha}(S)=\left\{\Delta \in \mathbb{R}^{d}:\left\|\Delta_{S^{c}}\right\|_{1} \leq \alpha\left\|\Delta_{S}\right\|_{1}\right\} .
$$

Using this cone, we defined the restricted eigenvalue condition for efficient bounds on estimation.

Definition 1.1. $X \sim \operatorname{RE}(S,(\kappa, \alpha))$ if

$$
\frac{1}{n}\|X \Delta\|_{2}^{2} \geq \kappa\|\Delta\|_{2}^{2} \quad \forall \Delta \in \mathbb{C}_{\alpha}(S) .
$$

We proved the following result, upper bounding the estimation error.
Theorem 1.1. Assume that $\operatorname{RE}(s,(\kappa, 3))$. With a proper choice of hyperparameter, we have

$$
\left\|\widehat{\theta}-\theta^{*}\right\|_{2} \lesssim \frac{1}{\kappa} \sqrt{s}\left\|\frac{X^{\top} w}{n}\right\|_{\infty} \lesssim \sigma \sqrt{\frac{s \log d}{n}} .
$$

We also showed that Gaussian random matrices satisfy this condition with high probability.

Theorem 1.2. Let $X \in \mathbb{R}^{n \times d}$ have iid $N(0,1)$ entries. If $n \gtrsim s \log d$, then with high probability, $X \sim \operatorname{RE}(S,(\kappa, 3))$ for all $|S| \leq s$.

### 1.2 LASSO prediction error bound

Instead of bounding $\left\|\widehat{\theta}-\theta^{*}\right\|_{2}$, we would like to bound the prediction errror (with fixed design):

$$
\frac{1}{n} \mathbb{E}_{\widetilde{w}}\left[\|\widetilde{y}-X \widehat{\theta}\|_{2}^{2}\right]=\frac{1}{n}\left\|X\left(\widehat{\theta}-\theta^{*}\right)\right\|_{2}^{2}+\sigma^{2}
$$

where $\widetilde{y}=X \theta^{*}+\widetilde{w}$ and $\approx N\left(0, \sigma^{2} I_{d}\right)$. We can upper bound $\frac{1}{n} \| X\left(\widehat{\theta}-\theta^{*}\left\|_{2}^{2} \leq\right\| \widehat{\theta}-\right.$ $\theta^{*}\left\|_{2}^{2}\right\| X^{\top} X / n \|_{\mathrm{op}}$; however, this is not always a good bound because $\left\|X^{\top} X / n\right\|_{\mathrm{op}}$, which has order $d / n$ (which blows up for $n \ll d$ ). Instead, we want to bound the prediction error directly

Theorem 1.3 (Prediction error bound). Let $\theta^{*}$ be s-sparse. Assume that the hyperparameter in the $\lambda$-formulation of the LASSO problem is $\lambda_{n} \geq 2\left\|\frac{X^{\top} w}{n}\right\|_{\infty}$. Then

1. Any optimal solution $\widehat{\theta}$ satisfies the bound

$$
\frac{1}{n}\left\|X\left(\widehat{\theta}-\theta^{*}\right)\right\|_{2}^{2} \leq 12\left\|\theta^{*}\right\|_{1} \lambda_{n}
$$

2. If $X$ satisfies $\operatorname{RE}(S,(\kappa, 3))$, then

$$
\frac{1}{n}\left\|X\left(\widehat{\theta}-\theta^{*}\right)\right\|_{2}^{2} \leq \frac{9}{\kappa} s \lambda_{n}^{2}
$$

Proof. As before, the proof is a basic inequality, plus some algebra.
Remark 1.1. The first bound is $\lesssim\left\|\theta^{*}\right\|_{1} \sqrt{\frac{\log d}{n}}$, so we get decay $O(1 / \sqrt{n})$. This is called the slow rate bound. The second bound is $\lesssim s\left(\sqrt{\frac{\log d}{n}}\right)^{2}$, so we get decay $O(1 / n)$. This is called the fast rate bound. Usually, without imposing any geometric assumptions, we get a slower rate bound than we get with such assumptions.

This phenomenon occurs in many settings such as in the empirical risk minimization
problem.


The setting is that we have data $\left(z_{i}\right)_{i \in[n]} \stackrel{\mathrm{iid}}{\sim} \mathbb{P}_{z}$ and a loss function $\ell: \Theta \times Z \rightarrow \mathbb{R}$. The empirical risk is

$$
\widehat{R}_{n}(\theta)=\frac{1}{n} \sum_{i=1}^{n} \ell\left(\theta ; X_{i}\right),
$$

and the population risk is

$$
R(\theta)=\mathbb{E}\left[\ell\left(\theta ; Z_{i}\right)\right] .
$$

If we take $\widehat{\theta}=\arg \min _{\theta} \widehat{R}_{n}(\theta)$, the minimizer of the empirical risk, then our generalization error is

$$
R(\widehat{\theta})-R\left(\theta^{*}\right)
$$

Without geometric assumptions, we can show a uniform convergence bound

$$
R(\widehat{\theta})-R\left(\theta^{*}\right) \leq 2 \sup _{\theta \in \Theta}\left|\widehat{R}_{n}(\theta)-R(\theta)\right| .
$$

Suppose $\Theta=B\left(0,10\left\|\theta^{*}\right\|\right)$. The upper bound of such an empirical process usually scales linearly in $\left\|\theta^{*}\right\|$, which does not give a very sharp prediction error bound.

Here is what we get with a geometric assumption. Assume that $\kappa\left\|\widehat{\theta}-\theta^{*}\right\|_{2}^{2} \leq(R(\widehat{\theta})-$ $\left.R\left(\theta^{*}\right)\right)$. Here, $\kappa$ is a strong convexity parameter. With this assumption, we can show an upper bound that is like

$$
R(\widehat{\theta})-R\left(\theta^{*}\right) \leq 2 \sup _{\theta \in B\left(\theta^{*},\left\|\widehat{\theta}-\theta^{*}\right\|_{2}\right)}\left|\widehat{R}_{n}(\theta)-R(\theta)\right| \lesssim\left\|\widehat{\theta}-\theta^{*}\right\|_{2} \sqrt{\frac{d \log d}{n}}
$$

This is nice because it scales linearly in the estimation error, which is usually smaller than $\left\|\theta^{*}\right\|$. We can bound $\left\|\widehat{\theta}-\theta^{*}\right\|_{2} \lesssim \sqrt{\frac{d \log d}{n}}$. Applying the geometric assumption gives the bound

$$
R(\widehat{\theta})-R\left(\theta^{*}\right) \leq \frac{d \log d}{n}
$$

### 1.3 Principal component analysis in high dimensions

Suppose we observe covariates $X_{1}, X_{2}, \ldots, X_{n} \stackrel{\text { iid }}{\sim} X \in \mathbb{R}^{d}$ with $\mathbb{E}[X]=0$ and $\operatorname{Cov}(X)=$ $\Sigma \in S_{+}^{d \times d}$. Let the eigenvalues of $\Sigma$ be $\lambda_{1}(\Sigma) \geq \lambda_{2}(\Sigma) \geq \cdots \geq \lambda_{d}(\Sigma) \geq 0$. We can find an orthonormal basis of eigenvectors $v_{1}(\Sigma), \ldots, v_{d}(\Sigma) \in \mathbb{R}^{d}$ such that $\Sigma v_{i}=\lambda_{i} v_{i}$ for all $i \in[d]$. If we let $\Lambda=\operatorname{diag}\left(\lambda_{1}, \ldots, \lambda_{d}\right) \in \mathbb{R}^{d \times d} \operatorname{nad} B=\left[v_{1}, \ldots, v_{d}\right] \in \mathbb{R}^{d \times d}$, then we can write $\Sigma=V \Lambda V^{\top}$.

The statistical interpretation of $v_{1}$ is that

$$
\begin{aligned}
v_{1} & \in \underset{\|v\|_{2}=1}{\arg \max } \operatorname{Var}(\langle x, v\rangle) \quad X \in \mathbb{R}^{d}, \mathbb{E}[X]=0 . \\
& =\underset{\|v\|_{2}=1}{\arg \max }\left\langle v, \mathbb{E}\left[X X^{\top}\right] v\right\rangle \\
& =\underset{\left\|v_{2}\right\|=1}{\arg \max }\langle v, \Sigma v\rangle .
\end{aligned}
$$

More generally, if we let $V_{k}=\left[v_{1}, \ldots, v_{k}\right] \in \mathbb{R}^{d \times k}$, then

$$
V_{k} \in \underset{\begin{array}{c}
U \in R^{d \times k} \\
\text { partial orth. }
\end{array} \underset{\sum_{i=1}^{k} \operatorname{Var}\left(\left\langle X, u_{i}\right\rangle\right)}{\arg \max }}{\mathbb{E}\left[\left\|U^{\top} X\right\|_{2}^{2}\right]} .
$$

Here is our statistical question: Given samples $\left\{X_{i}\right\}_{i \in[n]} \stackrel{\text { iid }}{\sim} X \in \mathbb{R}^{d}$, how can we estimate the principal components? Straightforwardly, we can use the eigenvectors of the sample covariance. If we define the sample covariance matrix

$$
\widehat{\Sigma}=\frac{1}{n} \sum_{i=1}^{n} X_{i} X_{i}^{\top}, \quad \mathbb{E}[\widehat{\Sigma}]=\Sigma
$$

then our estimator is

$$
\widehat{\theta}=\underset{\theta}{\arg \max }\langle\theta, \widehat{\Sigma} \theta\rangle .
$$

By comparison, the ground truth is

$$
\theta^{*}=\underset{\|\theta\|_{2}=1}{\arg \max }\langle\theta, \Sigma \theta\rangle
$$

How close is $\widehat{\theta}$ to $\theta^{*}$ ? We want to translate the closeless of $\Sigma$ and $\widehat{\Sigma}$ to closeness of $\theta$ and $\theta^{*}$. To quantify this, recall Weyl's eigenvalue perturbation inequality:

Lemma 1.1 (Weyl's inequality). For any matrices $\widehat{\Sigma}, \Sigma$,

$$
\left|\lambda(\widehat{\Sigma})-\lambda_{i}(\Sigma)\right| \leq\|\widehat{\Sigma}-\Sigma\|_{\mathrm{op}} .
$$

The proof of this fact comes from the variational characterization of the eigenvalues.
For a perturbation inequality for the eigenvectors, we also need the first eigen-gap to be large.

Definition 1.2. Let $\lambda_{1}(\Sigma) \geq \lambda_{2}(\Sigma) \geq \cdots \geq \lambda_{d}(\Sigma)$ be the eigenvalues of $\Sigma$. Then $k$-th eigen-gap is $\nu_{k}=\lambda_{k}-\lambda_{k+1}$.

We will write $\nu=\nu_{1}$ to refer to the first eigen-gap. You can think of having a large eigen-gap as similar to the restricted eigenvalue condition for LASSO. The parameter $\nu$ plays a similar role to $\kappa$ in LASSO, where $\operatorname{RE}(S,(\kappa, 3))$ means that $\Delta^{\top} \frac{X^{\top} X}{n} \Delta \geq \kappa\|\Delta\|_{2}^{2}$.

Example 1.1. Here is an example of instability of a matrix with a small eigengap. Suppose we have a diagonal matrix

$$
Q_{0}=\left[\begin{array}{cc}
1 & 0 \\
0 & 1.01
\end{array}\right]
$$

The eigenvalues are $\lambda_{1}\left(Q_{0}\right)=1.01$ and $\lambda_{2}\left(Q_{0}\right)=1$, so the eigengap is $\nu\left(Q_{0}\right)=0.01$. In this case, $\theta^{*}\left(Q_{0}\right)=\left[\begin{array}{l}0 \\ 1\end{array}\right]$. Now look at the perturbation

$$
Q_{\varepsilon}=Q_{0}+\varepsilon\left[\begin{array}{ll}
0 & 1 \\
1 & 0
\end{array}\right]=\left[\begin{array}{cc}
1 & \varepsilon \\
\varepsilon & 1.01
\end{array}\right]
$$

where $\varepsilon$ is small. If $\varepsilon=0.01$, then $\theta^{*}\left(Q_{\varepsilon}\right) \approx\left[\begin{array}{l}0.53 \\ 0.85\end{array}\right]$, which is far from $\left[\begin{array}{l}0 \\ 1\end{array}\right]$.

### 1.4 General perturbation bound for eigenvectors

Theorem 1.4. Let $\Sigma \in S_{+}^{d \times d}$, and let $\theta^{*} \in \mathbb{R}^{d}$ be an eigenvector for $\lambda_{1}(\Sigma)$. Let $\nu=$ $\lambda_{1}(\Sigma)-\lambda_{2}(\Sigma)>0$ be the first eigen-gap. Let the perturbation $P \in S^{d \times d}$ be such that $\|P\|_{\mathrm{op}}<\nu / 2$, and let $\widehat{\Sigma}=\Sigma+P$. If $\widehat{\theta} \in \mathbb{R}^{d}$ is an eigenvector for $\lambda_{1}(\widehat{\Sigma})$, then

$$
\left\|\widehat{\theta}-\theta^{*}\right\|_{2} \leq \frac{2\|\widetilde{P}\|_{2}}{\nu-2\|P\|_{\mathrm{op}}}
$$

Here

$$
\widetilde{P}=U^{\top} P U=\left[\begin{array}{cc}
\widetilde{P}_{1,1} & \widetilde{P}^{\top} \\
\widetilde{P} & \widetilde{P}_{2,2}
\end{array}\right] \in \mathbb{R}^{d \times d}
$$

where $U$ is the orthogonal matrix such that $\Sigma=U \Lambda U^{\top}$ and the blocks of $\widetilde{P}$ have sizes

$$
\left[\begin{array}{cc}
1 \times 1 & d \times(d-1) \\
(d-1) \times 1 & (d-1) \times(d-1)
\end{array}\right]
$$

If $\|P\|_{\mathrm{op}}$, then we get the bound

$$
\left\|\widehat{\theta}-\theta^{*}\right\|_{2} \leq \frac{4}{\nu}\|\widetilde{P}\|_{2} \leq \frac{4}{\nu}\|P\|_{\mathrm{op}}
$$

To prove this, first let $\widehat{\Delta}=\widehat{\theta}-\theta^{*}$, and define the quantity

$$
\begin{aligned}
\Psi(\widehat{\Delta} ; P) & =\langle\widehat{\theta}, P \widehat{\theta}\rangle-\left\langle\theta^{*}, P \theta^{*}\right\rangle \\
& =\langle\widehat{\Delta}, P \widehat{\Delta}\rangle+2\left\langle\widetilde{\Delta}, P \theta^{*}\right\rangle
\end{aligned}
$$

Here is the basic inequality of PCA:
Lemma 1.2 (PCA basic inequality).

$$
\nu \cdot\left(1-\left\langle\widehat{\theta}, \theta^{*}\right\rangle^{2}\right) \leq|\psi(\widehat{\Delta} ; P)| .
$$

The left hand side measures the distance between $\widehat{\theta}$ and $\theta^{*}$. We first prove this basic inequality:

Proof. The zero order optimality condition for $\widehat{\theta}$ says that $\widehat{\theta}=\arg \max _{\theta}\langle\theta, \widehat{\Sigma} \theta\rangle$. Then

$$
\langle\widehat{\theta}, \widehat{\Sigma} \widehat{\theta}\rangle \geq\left\langle\theta^{*}, \widehat{\Sigma} \theta^{*}\right\rangle
$$

Recall that $\widehat{\Sigma}=\Sigma+P$. We can express this inequality as

$$
\langle\widehat{\theta}, \Sigma \widehat{\theta}\rangle+\langle\widehat{\theta}, P \widehat{\theta}\rangle \geq\left\langle\theta^{*}, \Sigma \theta^{*}\right\rangle+\left\langle\theta^{*}, P \theta^{*}\right\rangle
$$

Putting the like terms on each side gives

$$
\left\langle\theta^{*}, \Sigma \theta^{*}\right\rangle-\langle\widehat{\theta}, \Sigma \widehat{\theta}\rangle \leq\langle\widehat{\theta}, P \widehat{\theta}\rangle-\left\langle\theta^{*}, P \theta^{*}\right\rangle
$$

The right hand side is $\psi(\widehat{\Delta} ; P)$.
To figure out the left hand side, write $\widehat{\theta}=\rho \theta^{*}+\sqrt{1-\rho^{2}} z$, where $\|z\|_{2}=1,\left\langle z, \theta^{*}\right\rangle=0$. Then $\rho=\left\langle\widehat{\theta}, \theta^{*}\right\rangle$. We can then expand

$$
\begin{aligned}
\langle\widehat{\theta}, \Sigma \widehat{\theta}\rangle & =\left\langle\rho \theta^{*}+\sqrt{1-\rho^{2}} z, \Sigma\left(\rho \theta^{*}+\sqrt{1-\rho^{2}} z\right)\right\rangle \\
& =\rho^{2} \underbrace{\left\langle\theta^{*}, \Sigma \theta^{*}\right\rangle}_{=\lambda_{1}}+2 \rho \sqrt{1-\rho^{2}} \underbrace{\left\langle\theta^{*}, \Sigma z\right\rangle}_{=0}+\left(1-\rho^{2}\right) \underbrace{\langle z, \Sigma z\rangle}_{\leq 2} .
\end{aligned}
$$

The bound on the last term is because $\langle z, \Sigma z\rangle \leq \sup _{\|z\|_{2}=1,\left\langle z, \theta^{*}\right\rangle=0}\langle z, \Sigma z\rangle=\lambda_{2}$.

$$
\leq \rho^{2} \lambda_{1}+\left(1-\rho^{2}\right) \lambda_{2}
$$

So the left hand side is

$$
\left\langle\theta^{*}, \Sigma \theta^{*}\right\rangle-\langle\widehat{\theta}, \Sigma \widehat{\theta}\rangle \geq \lambda_{1}-\left(\rho^{2} \lambda_{1}+\left(1-\rho^{2}\right) \lambda_{2}\right)
$$

$$
\begin{aligned}
& =\left(\lambda_{1}-\lambda_{2}\right)\left(1-\rho^{2}\right) \\
& =\nu\left(1-\rho^{2}\right) .
\end{aligned}
$$

So we get

$$
\nu\left(1-\left\langle\widehat{\theta}, \theta^{*}\right\rangle^{2}\right) \leq \Psi(\widehat{\Delta} ; P)
$$

Proof. Given the basic inequality, we now upper bound

$$
\Psi(\widehat{\Delta} ; P)=\langle\widehat{\theta}, P \widehat{\theta}\rangle-\left\langle\theta^{*}, P \theta^{*}\right\rangle
$$

Write $\Sigma=U \Lambda U^{\top}$ and $P=U \widetilde{P} U^{\top}$. We know that $U^{\top} \theta^{*}=e_{1}$, the first standard basis vector, so

$$
U^{\top} \widehat{\theta}=U^{\top}\left(\rho \theta^{*}+\sqrt{1-\rho^{2}} z\right)+\rho e_{1}+\sqrt{1-\rho^{2}} \underbrace{U^{\top} z}_{=:: \tilde{z}},
$$

where $\|\widetilde{z}\|_{2}=1$. Then

$$
\begin{aligned}
\Psi(\widehat{\Delta} ; P) & =\left\langle U^{\top} \widehat{\theta}, \widetilde{P} U^{\top} \widehat{\theta}\right\rangle-\left\langle U^{\top} \theta^{*}, \widetilde{P} U^{\top} \theta^{*}\right\rangle \\
& =\left\langle\rho e_{1}+\sqrt{1-\rho^{2}} \widetilde{z}, \widetilde{P}\left(\rho e_{1}+\sqrt{1-\rho^{2}} \widetilde{z}\right\rangle-\left\langle e_{1}, \widetilde{P} e_{1}\right\rangle\right. \\
& =\rho^{2}\left\langle e_{1}, \widetilde{P} e_{1}\right\rangle+2 \rho \sqrt{1-\rho^{2}}\left\langle\widetilde{z}, \widetilde{P} e_{1}\right\rangle+\left(1-\rho^{2}\right)\langle\widetilde{z}, \widetilde{P} \widetilde{z}\rangle-\left\langle e_{1}, \widetilde{P} e_{1}\right\rangle \\
& =\left(1-\rho^{2}\right) \underbrace{\left\langle e_{1}, \widetilde{P} e_{1}\right\rangle}_{\leq\|P\|_{\text {op }}}+\left(1-\rho^{2}\right)\langle\widetilde{z}, \widetilde{P} \widetilde{z}\rangle+2 \rho \sqrt{1-\rho^{2}} \underbrace{\left\langle\widetilde{z}, \widetilde{P} e_{1}\right\rangle}_{\leq\|P\|_{2}} .
\end{aligned}
$$

So, using the basic inequality, we get

$$
\nu\left(1-\rho^{2}\right) \leq 2\left(1-\rho^{2}\right)\|P\|_{\text {op }}+2 \rho \sqrt{1-\rho^{2}}\|\widetilde{P}\|_{2} .
$$

We can solve this to get

$$
\sqrt{1-\rho^{2}} \leq \frac{2 \rho\|\widetilde{P}\|_{2}}{\nu-2\|P\|_{\mathrm{op}}}
$$

So

$$
\begin{aligned}
\left\|\widehat{\theta}-\theta^{*}\right\|_{2} & =\sqrt{2(1-\rho)} \\
& \leq \frac{\sqrt{2} \rho}{\sqrt{1+\rho}} \frac{2\|\widetilde{P}\|_{2}}{\nu-2\|P\|_{\mathrm{op}}} \\
& \leq \frac{2\|\widetilde{P}\|_{2}}{\nu-2\|P\|_{\mathrm{op}}}
\end{aligned}
$$

