Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 2-3 — Week 11

Instructor: S. Minsker *Scribe: S. Minsker*

1.1 Linear Regression.

(Also see the supplemental file "regression+sparsity.pdf" that has been uploaded to Brightspace)

Next, we will start the chapter devoted to some aspects of high-dimensional statistics. In particular, we will focus on the topic of linear regression. We start by recalling the basic facts about the least squares estimator. Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \hline \vdots \\ \hline X_n \end{pmatrix}$$

be the so-called "design matrix." The j-th row $\mathbf{X}$, vector $X_j \in \mathbb{R}^p$, is called the *design vector*. Without loss of generality, we will assume that $\|X_j\|_2 = 1$. Suppose that we observe n noisy linear measurements of an unknown vector $\lambda_* \in \mathbb{R}^p$,

$$Y = \mathbf{X}\lambda_* + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \mathbf{I}_n)$ is the Gaussian noise (or, more generally, a vector of independent centered random variables with variance σ^2). In other words, we have n independent observations of the form

$$Y_j = \langle \lambda_*, X_j \rangle + \varepsilon_j$$

We will start by assuming that $n \geq p$ and that the columns of matrix $\mathbf{X}$ are linearly independent. Consider the least squares estimator

$$\widehat{\lambda} = \operatorname*{argmin}_{v \in \mathbb{R}^p} \|Y - \mathbf{X}v\|_2^2 .$$

The solution is given by (check!) $\widehat{\lambda} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top Y$. We can verify that

$$\frac{1}{n}\left\|\mathbf{X}\left(\widehat{\lambda} - \lambda_*\right)\right\|_2^2 = \frac{1}{n}\left\|\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{X}\lambda_* + \varepsilon) - \mathbf{X}\lambda_*\right\|_2^2 = \frac{1}{n}\left\|\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\varepsilon\right\|_2^2 .$$

Hence the "in-sample" risk $\frac{1}{n}\mathbb{E}\left(\sum_{j=1}^n \langle \widehat{\lambda} - \lambda_*, X_j \rangle^2\right) = \frac{1}{n}\mathbb{E}\left\|\mathbf{X}(\widehat{\lambda} - \lambda_*)\right\|_2^2$ satisfies

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left\|\mathbf{X}(\widehat{\lambda} - \lambda_*)\right\|_2^2 &= \frac{1}{n}\mathbb{E}\left\|\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\varepsilon\right\|_2^2 \\
&= \frac{\sigma^2}{n}\operatorname{tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right) = \sigma^2\frac{p}{n},
\end{aligned}$$

where we used the fact that $\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is the matrix of the orthogonal projection onto the range of columns of X, implying that its trace equals p, the dimension of this subspace. Note that when p/n is large (e.g., $\frac{1}{2}$) and $\sigma^2 = O(1)$, the expected error is separated away from 0.

Exercise (Solved in class, see the supplemental file)**.** Let X' be a new measurement vector independent from $X_1, \ldots, X_n$. Show that the "out-of-sample" error $\mathbb{E}\left((\widehat{\lambda} - \lambda_*)^T X'\right)^2$ satisfies

$$\mathbb{E}\left((\widehat{\lambda} - \lambda_*)^T X'\right)^2 > \sigma^2 \frac{p}{n}.$$

What if we have additional information about λ_*? For example, what if we know that all but $s \ll p$ of coordinates of λ_* are 0? Can we use this information to obtain a better estimator? We address this question next.

1.1.1 Estimation with prior information

We will express "prior information" about λ_* as $\lambda_* \in \mathcal{K}$ where $\mathcal{K}$ is a known set, for instance the set of all s-sparse vectors, meaning that

$$\mathcal{K} = \{\lambda \in \mathbb{R}^p : |\lambda|_0 \le s\}$$

where $|\lambda|_0 = \text{card}\{j : \lambda_j \neq 0\}$, cardinality of the support of λ).

The problem that we will consider is the following: let $\lambda_* \in \mathcal{K} \subseteq \mathbb{R}^p$ be an unknown element of a known set $\mathcal{K}$, and $Y = \mathbf{X}\lambda_*$ where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a design matrix, and p can be larger than n. We will first consider the scenario when the design vectors $X_1, \ldots, X_n$, the rows of $\mathbf{X}$ are random. More specifically, let's assume that X_j, $j = 1, \ldots, n$ are i.i.d $N(0, \mathbf{I}_p)$. We will also start with a *noiseless case* (as opposed to the noisy case $Y = \mathbf{X}\lambda + \varepsilon$).

All we know is that (a) $\lambda_* \in \mathcal{K}$ and (b) $\lambda_* \in E'$ – an affine subspace of dimension $p - n$ defined by $E' = \{\lambda' \in \mathbb{R}^p \,|\, Y = \mathbf{X}\lambda'\}$. Hence, any $\widehat{\lambda} \in \mathcal{K} \cap E'$ can be viewed as an estimate of λ_*. Assume in addition that the set $\mathcal{K}$ is bounded (many situations can be reduced to the case of bounded $\mathcal{K}$, as we will see in examples later). It is clear that $\widehat{\lambda} - \lambda_* \in \mathcal{K} - \mathcal{K} := \{u - v,\ u, v \in \mathcal{K}\}$ and $\widehat{\lambda} - \lambda_* \in E := \{\lambda' \in \mathbb{R}^p \,|\, \mathbf{X}\lambda' = 0\}$, the kernel of $\mathbf{X}$, hence

$$\left\|\widehat{\lambda} - \lambda_*\right\|_2 \le \text{diam}\left((\mathcal{K} - \mathcal{K}) \cap E\right).$$

To bound the latter quantity, we will need to introduce several definitions.

Definition 1. Let $\eta \in \mathbb{R}^p$ be a unit vector. The <u>width of $\mathcal{K}$ in direction η</u> is defined as

$$w_\eta(\mathcal{K}) = \sup_{u,v\in\mathcal{K}} \langle \eta, u - v\rangle.$$

Definition 2 (Spherical mean width)**.** The <u>spherical mean width</u> of $\mathcal{K}$ is is defined as

$$\widetilde{w}(\mathcal{K}) = \mathbb{E}w_\eta(\mathcal{K}),$$

where $\eta \sim \text{Unif}(\mathcal{S}^{p-1})$ (in other words, η is uniformly distributed over the unit sphere).

Definition 3 (Gaussian mean width)**.** The Gaussian mean width of $\mathcal{K}$ is defined as

$$w(\mathcal{K}) = \mathbb{E} w_g(\mathcal{K}),$$

where $g \sim \mathcal{N}(0, \mathbf{I}_p)$.

The relationship between the Gaussian mean width and the spherical mean width is given by

$$\begin{aligned} w(\mathcal{K}) &= \mathbb{E} \sup_{u,v \in k} \langle g, u - v \rangle \\ &= \mathbb{E} \|g\|_2 \cdot \sup_{u,v} \left\langle \frac{g}{\|g\|_2}, u - v \right\rangle = \widetilde{w}(\mathcal{K}) \mathbb{E} \|g\|_2 , \end{aligned}$$

where we used the fact that $\|g\|_2$ and $\frac{g}{\|g\|_2}$ are independent and that $\frac{g}{\|g\|_2}$ has $\mathrm{Unif}(\mathcal{S}^{p-1})$ (check these claims!). It is also well known that $\mathbb{E} \|g\|_2 = \sqrt{2} \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}$ and that $\frac{p}{\sqrt{p+1}} \leq \mathbb{E} \|g\|_2 \leq \sqrt{p}$, hence the Gaussian and spherical mean widths are equivalent; however, it is easier to estimate the Gaussian mean width since the coordinates of g are independent, while the coordinates of $\eta \sim \mathrm{Unif}(\mathcal{S}^{p-1})$ are not. Moreover, the process $\mathcal{K} \ni z \mapsto \langle g, z \rangle$ is a Gaussian (hence, also sub-Gaussian) process indexed by the set $\mathcal{K}$, hence the Gaussian mean width can be bounded by the Dudley's entropy integral which depends only on the metric properties of $\mathcal{K}$.