Lecture 20

Last time

- Ordinary least squares
- Excess risk

Today

- Ridge Regression
- Random design
- Beyond linear features

Ridge Regression

The excess rate that we proved assumed that $d \leq n$ and X is full rank. But in high-dimensional settings $d \gg n$ and our guarantees do not hold.

Even when $\varepsilon = 0$ there are many solutions, which one to choose?

One way to do this is by enforcing certain structure with a regularizer, i.e., an additional term in the objective (common choices are the ℓ_1 norm and the ℓ_2-norm squared).

Def: For any $\lambda > 0$, define ridge least-squares estimator as

Prefers small norm

$$\theta^\lambda = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \underbrace{\frac{1}{n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2}_{\ell_\lambda(\theta)}.$$

Lemma: The ridge least-squares estimator exist and it is unique. Further,

$$\theta^\lambda = \frac{1}{n}\left(\hat{\Sigma} + \lambda I\right)^{-1} X^T y$$

This matrix doesn't have to be invertible.

Proof: Just as before, the lost is convex and

$$0 = \nabla \ell_\lambda(\theta) = \frac{2}{n} X^T(X\theta - y) + 2\lambda\theta$$
$$\Rightarrow \left(\hat{\Sigma} + \lambda I\right)\theta = \frac{1}{n} y. \qquad \square$$

Note that when $\hat{\Sigma}$ is invertible and $\lambda = 0$, this recovers θ^{OLS}. Indeed, as $\lambda \to 0$ $\theta^\lambda \to (X^TX)^+ X^T y$ where $(.)^+$ denotes the pseudo-inverse.

Let's understand what happens to the bias and variance.

Lemma: We have that

$$\text{Bias}_{\hat{\Sigma}}(\theta^\lambda) = \lambda^2 \theta^{*T}\left(\hat{\Sigma} + \lambda I\right)^{-2}\hat{\Sigma}\theta^*$$

$$\text{Var}_{\hat{\Sigma}}(\theta^\lambda) = \frac{\sigma^2}{n} \text{tr}\left(\hat{\Sigma}^2\left(\hat{\Sigma} + \lambda I\right)^{-2}\right)$$

Proof: We will use the following fact.

Fact: For any symmetric matrix A,

$$A(A+\lambda I)^{-1} = (A+\lambda I)^{-1}A.$$

This follows since A and $(A+\lambda I)^{-1}$ share the same eigenvectors (Why?).

With this,

$$\begin{aligned}
\mathbb{E}\,\theta^\lambda &= \frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1}X^T\mathbb{E}y \\
&= \frac{1}{n}(\hat{\Sigma}+\lambda I)^{-1}X^TX\theta^* \\
&= (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\theta^* \\
&= \theta^* + ((\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma} - I)\theta^* \\
&= \theta^* + ((\hat{\Sigma}+\lambda I)^{-1}(\hat{\Sigma}+\lambda I-\lambda I) - I)\theta^* \\
&= \theta^* + \lambda(\hat{\Sigma}+\lambda I)^{-1}\theta^*.
\end{aligned}$$

Add and subtract →

Thus,

$$\begin{aligned}
\mathrm{Bias}_{\hat{\Sigma}}(\theta^\lambda) &= \lambda^2\|(\hat{\Sigma}+\lambda I)^{-1}\theta^*\|^2_{\hat{\Sigma}} \\
&= \lambda^2\theta^{*T}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\theta^*.
\end{aligned}$$

The fact above →

Expanding for Var gives

$$\mathrm{Var}_{\hat{\Sigma}}(\theta^\lambda) = \mathbb{E}\|\theta^\lambda - \mathbb{E}\theta^\lambda\|^2_{\hat{\Sigma}}$$

$$= \mathbb{E} \left\| \frac{1}{n} (\hat{\Sigma} + \lambda I)^{-1} X^T y - \frac{1}{n} (\hat{\Sigma} + \lambda I)^{-1} X^T X \theta^* \right\|_{\hat{\Sigma}}^2$$

$$= \mathbb{E} \left\| \frac{1}{n} (\hat{\Sigma} + \lambda I)^{-1} X^T \varepsilon \right\|_{\hat{\Sigma}}^2$$

$$= \frac{1}{n} \mathbb{E} \, tr\left(\varepsilon^T X (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I) X^T \varepsilon \right)$$

Cyclic invariance of tr →

$$= \frac{\sigma^2}{n} tr\left(\hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I) \right)$$

Fact →

$$= \frac{\sigma^2}{n} tr\left(\hat{\Sigma}^2 (\Sigma + \lambda I)^{-2} \right).$$

□

Notice that

$$tr\left(\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2} \right) = \sum_{i=1}^{d} \frac{\lambda_i^2}{(\lambda_i + \lambda)^2} < d$$

(eigenvalues of $\hat{\Sigma}$)

quantifies the intrinsic dimension or degrees of freedom of the problem. It provides a soft count on how many eigenvalues $\lambda_i \gg \lambda$.

Altogether we have proven:

Corollary:

$$\mathbb{E} \, R(\hat{\theta}_\lambda) - R^* = \lambda^2 \theta^* (\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma} \theta^*$$

$$+ \frac{\sigma^2}{n} \operatorname{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right).$$

Remarks:

- No dimension dependence! This is key for learning in infinite dim.
- This estimator is not unbiased any more (unless $\lambda = 0$).
- We can optimize over λ!

Theorem: Take $\lambda = \frac{\sigma}{\|\theta^*\|_2}\sqrt{\frac{\operatorname{tr}(\hat{\Sigma})}{n}}$. Then,

$$\mathbb{E}R(\theta^\lambda) - R^* \leq \sigma\|\theta^*\|_2\sqrt{\frac{\operatorname{tr}(\hat{\Sigma})}{n}}.$$

Proof: Note that we can upper bound

$$\lambda^2\theta^{*\top}(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\theta^* \leq \lambda \underbrace{\|(\hat{\Sigma}+\lambda I)^{-2}\lambda\hat{\Sigma}\|_{op}}\|\theta^*\|_2^2$$

$$(\lambda_i+\lambda)^{-2}\lambda\lambda_i \leq \frac{1}{2} \iff 2\lambda\lambda_i \leq (\lambda_i+\lambda)^2 \quad \forall i.$$

$$\leq \frac{\lambda}{2}\|\theta^*\|_2^2.$$

Further,

$$\frac{\sigma^2}{n}\operatorname{tr}\left(\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\right) \leq \frac{\sigma^2}{\lambda n}\operatorname{tr}\left(\hat{\Sigma}\,\lambda\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-2}\right)$$

Show that $\mathrm{tr}(AM) \le \mathrm{tr}(A)\lambda_{\max}(M)$

$$\le \frac{\sigma^2}{2\lambda n}\mathrm{tr}(\hat{\Sigma}).$$

Altogether

$$\mathbb{E}R(\hat{\theta}^\lambda) - R^* \le \frac{\lambda}{2}\|\theta^*\|_2^2 + \frac{\sigma^2\,\mathrm{tr}(\hat{\Sigma})}{2\lambda n}.$$

Notice that this has the form

$$a\lambda + \frac{b}{\lambda} \quad \text{with} \quad a, b > 0,$$

which is always minimized at $\lambda = \sqrt{b/a}$. Check!

□

Remarks:

- The dependency on n is worst than for OLS. But there is no d and dependency on σ is better.
- It's hard to find λ (cross-validation in practice).
- This is extremely relevant for learning with kernels (see Project topics).

Random Design

So far we assumed X was fixed, next we suppose (x_i, y_i) are drawn iid with

$$y_i = \theta^{*T} x_i + \varepsilon_i$$

(x_i: ind; ε_i: iid $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = 1$.)

Lemma: Under the random design model.

$$\forall \theta,\ R(\theta) - R^* = \|\theta - \theta^*\|_\Sigma^2 \quad \text{with } \Sigma = \mathbb{E} x x^T$$

Proof: Just as before

$$R(\theta) = \mathbb{E}(y - x^T\theta)^2 \quad (y \in \mathbb{R},\ x^T\theta \in \mathbb{R}^d)$$

$$= \mathbb{E}(x^T(\theta^* - \theta) + \varepsilon)^2$$

$$= (\theta^* - \theta)^T \mathbb{E} x x^T (\theta^* - \theta)^T + \sigma^2 \quad \square$$

Lemma: Under the random design model and assuming $\hat{\Sigma}$ is invertible,

$$\mathbb{E} R(\theta^{OLS}) - R^* = \frac{\sigma^2}{n} \mathbb{E}\, tr(\Sigma \hat{\Sigma}^{-1})$$

(Population: Σ; Empirical: $\hat{\Sigma}^{-1}$)

Proof: Recall $\theta^{OLS} = \theta^* + \frac{1}{n}\hat{\Sigma}^{-1} X^T \varepsilon$. Thus,

$$\mathbb{E} R(\theta^{OLS}) - R^* = \mathbb{E}\|\theta^{OLS} - \theta^*\|_\Sigma^2$$

$$= \frac{1}{n^2} \mathbb{E}\, \varepsilon^T X \hat{\Sigma}^{-1} \Sigma^{-1} \hat{\Sigma}^{-1} X^T \varepsilon$$

$$= \frac{1}{n} \mathbb{E}\, \mathrm{tr}(\varepsilon\varepsilon^T \hat{\Sigma} \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1})$$

$$= \frac{1}{n} \mathop{\mathbb{E}}_{X} \mathrm{tr}(\mathop{\mathbb{E}}_{\varepsilon} \varepsilon\varepsilon^T \Sigma \hat{\Sigma}^{-1})$$

$$= \frac{\sigma^2}{n} \mathbb{E}\, \mathrm{tr}(\Sigma \hat{\Sigma}^{-1}).$$

□

Note that unlike before $\Sigma \hat{\Sigma}^{-1} \neq I$ and the bound depends on how fast does $\mathbb{E}\, \mathrm{tr}(\Sigma \hat{\Sigma}^{-1}) \to d$.

For instance if $z \sim N(0, \Sigma)$ one can show

$$\mathbb{E}\, \mathrm{tr}(\Sigma \hat{\Sigma}^{-1}) = \frac{d}{1 - \frac{d+1}{n}}.$$

Try to prove it.
Hint: Inverse Washart dist.

Beyond linear features.

Everything we developed holds

if we imagine that

$$y_i = \theta^{*T} \varphi(x_i) + \varepsilon_i$$

where $\varphi : \mathbb{R}^k \to \mathbb{R}^d$. Then we substitute the matrix X by

$$\Phi = \begin{pmatrix} - \varphi(x_1) - \\ \vdots \\ - \varphi(x_n) - \end{pmatrix},$$

and $\hat{\Sigma}$ by $\frac{1}{n} \Phi^T \Phi$ and everything follows just as before.