CSE 546: Machine Learning **Lecture 3**

Risk of Ridge Regression

Instructor: Sham Kakade

0.1 Analysis

Let us rotate each X_i by $V^\top$, i.e.

$$X_i \leftarrow V^\top X_i$$

where V is the right matrix of the SVD of the $n \times d$ matrix $\mathbf{X}$ (note this rotation does not alter the predictions of rotationally invariant algorithms).

In this rotated, coordinate system, we have that:

$$\mathbf{\Sigma} := \frac{1}{n}\mathbf{X}^\top \mathbf{X} = diag(\lambda_1, \lambda_2, \ldots \lambda_d)$$

and that:

$$[\hat{w}_\lambda]_j = \frac{\frac{1}{n}\sum_{i=1}^n Y_i [X_i]_j}{\lambda_j + \lambda}$$

It is straightforward to see that:

$$w_* = \mathbb{E}[\hat{w}_0]$$

(where w_* is the minimizer defined in the previous lecture). It follows that:

$$[\mathbb{E}[\hat{w}]_\lambda]_j := \mathbb{E}[\hat{w}_\lambda]_j = \frac{\lambda_j}{\lambda_j + \lambda}(w_*)_j$$

by just taking expectations.

Lemma 0.1. *(Risk Bound) If* $\mathrm{Var}(Y_i) = \sigma^2$*, we have that:*

$$R(\hat{w}_\lambda) = \frac{\sigma^2}{n}\sum_j \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2 + \sum_j (w_*)_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2}$$

The above is an equality if $\mathrm{Var}(Y_i) \leq \sigma^2$.

Proof. Note that in our coordinate system we have $X = UD^\top$ (from the thin SVD), since $X^\top X$ is diagonal. Here, the diagonal entries are $\sqrt{n\lambda_j}$. Letting η be the noise:

$$Y = \mathbb{E}[Y] + \eta$$

and

$$\Sigma_\lambda = \Sigma + \lambda \mathrm{I}\,,$$

so that $\hat{w}_\lambda = \frac{1}{n}\Sigma_\lambda X^\top Y$. We have that:

$$\begin{aligned}
\mathbb{E}_Y \|\hat{w}_\lambda - \mathbb{E}[\hat{w}]_\lambda\|_\Sigma^2 &= \frac{1}{n^2}\mathbb{E}_\eta[\eta^\top X\Sigma_\lambda \Sigma \Sigma_\lambda X \eta] \\
&= \frac{1}{n^2}\mathbb{E}_\eta[\eta^\top U Diag(\ldots, \frac{n\lambda_j^2}{(\lambda_j+\lambda)^2}, \ldots) U^\top \eta] \\
&= \frac{1}{n}\sum_j \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}\mathbb{E}_\eta[U^\top \eta]_j^2 \\
&= \frac{\sigma^2}{n}\sum_j \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}
\end{aligned}$$

This holds with equality if $\mathrm{Var}(Y_i) = 1$. For the bias term,

$$\begin{aligned}
\|\overline{w}_\lambda - w_*\|_\Sigma^2 &= \sum_j \lambda_j([\overline{w}_\lambda]_j - [w_*]_j)^2 \\
&= \sum_j (w_*)_j^2 \lambda_j (\frac{\lambda_j}{\lambda_j+\lambda} - 1)^2 \\
&= \sum_j (w_*)_j^2 \lambda_j (\frac{\lambda}{\lambda_j+\lambda})^2
\end{aligned}$$

and the result follows from algebraic manipulations. □

0.2 A (dimension-free) margin bound

There following bound characterizes the risk for two natural settings for λ.

Theorem 0.2. *Assume the linear model is correct: Define $\overline{d}$ as:*

$$\overline{d} = \frac{1}{n}\sum_i \|X_i\|^2$$

For $\lambda = \frac{\sqrt{\overline{d}}}{\|w_*\|\sqrt{n}}$, then:

$$R(\hat{w}_\lambda) \le \frac{\|w_*\|\sqrt{\overline{d}}}{\sqrt{n}} \le \frac{\|w_*\| X_+}{\sqrt{n}}$$

where X_+ is a bound on the norm of $\|X\|_i$.

Conceptually, the second bound is 'dimension free', i.e. it does not depend explicitly on d, which could be infinite. And we are effectively doing regression in a large (potentially) infinite dimensional space.

Proof. The $\lambda = 0$ case follows directly from the previous lemma. Using that $(a+b)^2 \ge 2ab$, we can bound the variance term for general λ as follows:

$$\frac{1}{n}\sum_j (\frac{\lambda_j}{\lambda_j+\lambda})^2 \le \frac{1}{n}\sum_j \frac{\lambda_j^2}{2\lambda_j \lambda} = \frac{\sum_j \lambda_j}{2n\lambda}$$

Again, using that $(a+b)^2 \ge 2ab$, the bias term is bounded as:

$$\sum_j (w_*)_j^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2} \le \sum_j (w_*)_j^2 \frac{\lambda_j}{2\lambda_j/\lambda} = \frac{\lambda}{2}||w_*||^2$$

So we have that:

$$R(\hat{w}_\lambda) \leq \frac{\|\Sigma\|_{\text{trace}}}{2n\lambda} + \frac{\lambda}{2}||w_*||^2$$

and using the choice of λ completes the proof.

$\square$