Stat 928: Statistical Learning Theory **Lecture: 3**

Ridge Regression; Dimensionality Reduction; and Feature Selection

Instructor: Sham Kakade

1 Ridge Regression

1.1 Bias Variance in the Fixed Design Setting

Lemma 1.1. *(bias-variance for risk) We can decompose the expected risk as:*

$$R(\hat{\beta}) = \mathbb{E}_Y \|\hat{\beta} - \mathbb{E}[\hat{\beta}]\|_\Sigma^2 + \|\mathbb{E}[\hat{\beta}] - \beta\|_\Sigma^2$$
$$= \frac{1}{n}\mathbb{E}_Y \|\mathbb{E}[\hat{Y}] - \hat{Y}\|^2 + \frac{1}{n}\|Y^* - \mathbb{E}[\hat{Y}]\|^2$$

where we have that:

$$\textit{(average) variance} = \frac{1}{n}\mathbb{E}_Y \|X\hat{\beta} - X\mathbb{E}[\hat{\beta}]\|^2 = \frac{1}{n}\mathbb{E}_Y \|\mathbb{E}[\hat{Y}] - \hat{Y}\|^2$$

and

$$\textit{prediction bias vector} = X\beta - X\mathbb{E}[\hat{\beta}] = Y^* - \mathbb{E}[\hat{Y}]$$

1.2 Ridge Regression and the Bias-Variance Tradeoff

The ridge regression estimator using an outcome Y is just:

$$\hat{\beta}_\lambda = \arg\min_w \frac{1}{n}\|Y - Xw\|^2 + \lambda\|w\|^2$$

The estimator is then:

$$\hat{\beta}_\lambda = (\Sigma + \lambda I)^{-1}(\frac{1}{n}X^\top Y) = (\Sigma + \lambda I)^{-1}(\frac{1}{n}\sum Y_i X_i^\top)$$

For simplicity, let us rotate X such that:

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

(note this rotation does not alter the predictions of rotationally invariant algorithms). With this choice, we have that:

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

It is straightforward to see that:

$$\beta = \mathbb{E}[\hat{\beta}_0]$$

and it follows that:

$$[\mathbb{E}[\hat{\beta}]_\lambda]_j := \mathbb{E}[\hat{\beta}_\lambda]_j = \frac{\lambda_j}{\lambda_j + \lambda}\beta_j$$

by just taking expectations.

Lemma 1.2. *(Risk Bound) If* $\mathrm{Var}(Y_i) = \sigma^2$*, we have that:*

$$R(\hat{\beta}_\lambda) = \frac{\sigma^2}{n} \sum_j (\frac{\lambda_j}{\lambda_j + \lambda})^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2}$$

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

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

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

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

and the result follows from algebraic manipulations. □

1.3 Margin Based Bound

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

Theorem 1.3. *Assume* $\mathrm{Var}(Y_i) \leq 1$

- *(Finite Dims) For* $\lambda = 0$*,*

$$R(\hat{\beta}_\lambda) \leq \frac{d}{n}$$

  And if $Var(Y_i) = 1$*, then* $R(\hat{\beta}_\lambda) = \frac{d}{n}$.

- *(Infinite Dims) For* $\lambda = \frac{\sqrt{\|\Sigma\|_{trace}}}{\|\beta\|\sqrt{n}}$*, then:*

$$R(\hat{\beta}_\lambda) \leq \frac{\|\beta\|\sqrt{\|\Sigma\|_{trace}}}{2\sqrt{n}} \leq \frac{\|\beta\|\|\mathcal{X}\|}{2\sqrt{n}}$$

  where the trace norm is the sum of the singular values and $\|\mathcal{X}\| = \max_i ||X_i||$*. Furthermore, for all* n *there exists a distribution* $\Pr[Y]$ *and an* X *such that the* $\inf_\lambda R(\hat{\beta}_\lambda)$ *is* $\Omega^*(\frac{\|\beta\|\|\mathcal{X}\|}{\sqrt{n}})$ *(so the above bound is tight up to log factors in* n*).*

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 $(a + b)^2 \geq 2ab$, we can bound the variance term for general λ as follows:

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

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

$$\sum_j \beta_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2} \leq \sum_j \beta_j^2 \frac{\lambda_j}{2\lambda_j/\lambda} = \frac{\lambda}{2}||\beta||^2$$

So we have that:

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

and using the choice of λ completes the proof.

To see the above bound is tight, consider the following problem. Let $X_i = \sqrt{\frac{n}{i}}$ and $\beta_i = \sqrt{\frac{1}{i}}$ and let $Y = X\beta + \eta$ where η is unit variance. Here, we have that $\lambda_i = \frac{1}{i}$ so $\sum_j \lambda_j \leq \log n$ and $\|\beta\|^2 \leq \log n$, so the upper is $\frac{\log n}{\sqrt{n}}$. Now one can write the risk as:

$$\begin{aligned}
R(\hat{\beta}_\lambda) &= \frac{1}{n}\sum_j (\frac{\frac{1}{i}}{\frac{1}{i} + \lambda})^2 + \sum_j \frac{\frac{1}{i^2}}{(1 + \frac{1}{i\lambda})^2} && (1)\\
&= \sum_j \frac{\frac{1}{n} + \lambda^2}{(1 + i\lambda)^2} && (2)\\
&\geq \int_1^n \frac{\frac{1}{n} + \lambda^2}{(1 + x\lambda)^2} dx && (3)\\
&= (\frac{1}{n} + \lambda^2)(\frac{1}{\lambda(1 + \lambda)} - \frac{1}{\lambda(1 + n\lambda)}) && (4)\\
&= (\frac{1}{n\lambda} + \lambda)(\frac{1}{1 + \lambda} - \frac{1}{1 + n\lambda}) && (5)\\
& && (6)
\end{aligned}$$

and this is $\Omega(\sqrt{n})$, for all λ. □

2 PCA Projections and MLEs

Fix some λ. Consider the following 'keep or kill' estimator, which uses the MLE estimate if $\lambda_i \geq \lambda$ and 0 otherwise:

$$[\hat{\beta}_{PCA,\lambda}]_j = \begin{cases} [\hat{\beta}_0]_j & \text{if } \lambda_i \geq \lambda \\ 0 & \text{else} \end{cases}$$

where $\hat{\beta}_0$ is the MLE estimator. This estimator is 0 for the small values of λ_i (those in which we are effectively regularizing more anyways).

Theorem 2.1. *(Risk Inflation of $\hat{\beta}_{PCA,\lambda}$)*

Assume $\mathrm{Var}(Y_i) = 1$*, then*

$$\mathbb{E}_Y[R(\hat{\beta}_{PCA,\lambda})] \leq 4\mathbb{E}_Y[R(\hat{\beta}_\lambda)]$$

Note that the the actual risk (not just an upper bound) of the simple PCA estimate is within a factor of 4 of the ridge regression risk on a wide class of problems.

Proof. Recall that:

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] = \frac{1}{n}\sum_j \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1 + \lambda_j/\lambda)^2}$$

Since we can write the risk as:

$$\mathbb{E}_Y[R(\hat{\beta})] = \mathbb{E}_Y\|\hat{\beta} - \overline{\beta}\|_\Sigma^2 + \|\overline{\beta} - \beta\|_\Sigma^2$$

we have that:

$$\mathbb{E}_Y[R(\hat{\beta}_{PCA,\lambda})] = \frac{1}{n}\sum_j \mathbb{I}(\lambda_j > \lambda) + \sum_{j:\lambda_j < \lambda} \lambda_j \beta_j^2$$

where $\mathbb{I}$ is the indicator function.

We now show that each term in the risk of $\hat{\beta}_{PCA,\lambda}$ is within a factor of 4 for each term in $\hat{\beta}_\lambda$. If $\lambda_j > \lambda$, then the ratio of the $j - th$ terms is:

$$\begin{aligned}
\frac{\frac{1}{n}}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2 + \beta_j^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2}} &\leq \frac{\frac{1}{n}}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2} \\
&= \frac{(\lambda_j + \lambda)^2}{\lambda_j^2} \\
&\leq \left(1 + \frac{\lambda}{\lambda_j}\right)^2 \\
&\leq 4
\end{aligned}$$

Similarly, if $\lambda_j \leq \lambda$, then the ratio of the j-th terms is:

$$\begin{aligned}
\frac{\lambda_j \beta_j^2}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2 + \frac{\lambda_j \beta_j^2}{(1+\lambda_j/\lambda)^2}} &\leq \frac{\lambda_j \beta_j^2}{\frac{\lambda_j \beta_j^2}{(1+\lambda_j/\lambda)^2}} \\
&= (1 + \lambda_j/\lambda)^2 \\
&\leq 4
\end{aligned}$$

Since each term is within a factor of 4, the proof is completed. □