Stat 928: Statistical Learning Theory **Lecture: 2**

Fixed Design and Least Squares

Instructor: Sham Kakade

1 Intro

We now attempt to understand the least squares algorithm in the fixed design setting. We start with the $p < n$ case, and then move to the $p > n$ case.

2 Review: The SVD; the "Thin" SVD; and the pseudo-inverse

Theorem 2.1. *(SVD) For any matrix $X \in \mathbb{R}^{n \times p}$, there exists $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ orthogonal matrices (e.g. matrices with ortogonal rows and columns, so that $UU^\top = \mathrm{I}$ and $VV^top = \mathrm{I}$) such that:*

$$X = \sum_i \lambda_i u_i v_i^\top = U diag(\lambda_1, \ldots \lambda_{\min\{n,d\}}) V^\top$$

where $diag(\cdot)$ is diagonal $\mathbb{R}^{n \times p}$ matrix and the λ_i's are referred to as the the singular values.

For $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^n$, suppose that the equation:

$$X\beta = Y$$

has a unique solution and that X is invertible, then:

$$\beta = X^{-1}Y$$

In regression, there is typically noise, and we find a β which minimizes:

$$\|X\beta - Y\|^2$$

Clearly, if there is no noise, then a solution is given by $\beta = X^{-1}Y$, assuming no degeneracies. In general though, the least squares solution is given by:

$$\beta = (X^\top X)^{-1} XY \tag{1}$$

which one can argue is a less intuitive (and elegant) expression than when there is an exact solution. Furthermore, Equation 1 above only holds if X is of rank d (else $(X^\top X)^{-1}$ would not be invertible).

Now let us define the Moore-Penrose pseudo-inverse. While there are a variety of more elegant definitions of the pseudo-inverse, in terms of certain desirable properties, we take the more brute force definition.

First, let us define the 'thin' SVD.

Definition 2.2. *We say $X = UDV^\top$ is the "thin" SVD of $X \in \mathbb{R}^{n \times p}$ if: $U^{n \times r}$ and $V^{p \times r}$ have orthonormal columns (e.g. where r is the number of columns) and $D \in \mathbb{R}^{r \times r}$ is diagonal, with all it's diagonal entries being non-zero.*

Now we define the pseudo-inverse as follows:

Definition 2.3. *Let $X = UDV^\top$ be the thin SVD of X. The Moore-Penrose pseudo-inverse of X, denoted by X^+, is defined as:*

$$X^+ = VD^{-1}U^\top$$

Let us make some observations:

1. First, if X is invertible (so X is square) then $X^+ = X^{-1}$.
2. Suppose that X isn't square and that $Xw = Y$ has a (unique) solution, then $w = X^+Y$.
3. Now suppose that $Xw = Y$ has (at least one) solution. Then one solution is given by $w = X^+Y$. This solution is the minimum norm solution w.
4. (geometric interpretation) The matrix X^+ maps any point in the range of X to the minimum norm point in the domain.

With the pseudo-inverse, we have the much more elegant least squares estimator:

Lemma 2.4. *The least squares estimator is:*

$$\beta = X^+Y$$

(Note that the above is alway a minimizer, while the solution provided in Equation 1 only holds if $X^\top X$ is invertible, in which case the minimizer is unique).

3 Risk and Fixed Design Regression

Let us now consider the 'normal means' problem, sometimes referred to as the fixed design setting (also sometimes referred to as the problem of signal reconstruction). Here, we have a set of n points $\mathcal{X} = \{X_i\} \subset \mathbb{R}^p$, and let X denote the $\mathbb{R}^{n\times p}$ matrix where the i row of X is X_i. We also observe a output vector $Y \in \mathbb{R}^n$. We desire to learn $\mathbb{E}[Y]$. In particular, we seek to predict $\mathbb{E}[Y]$ as $X\hat{\beta}$.

The square loss of an estimator w is:

$$L(w) = \frac{1}{n}\mathbb{E}_Y\|Y - Xw\|_2^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i - X_iw)^2$$

where the expectation is with respect to Y. Let β be the optimal predictor:

$$\beta = \arg\min_w L(w)$$

Let $\hat{\beta}_Y$ be an estimator constructed with the outcome Y — we drop the explicit Y dependence as this is clear from context. The (fixed design) risk of an estimator $\hat{\beta}$ is defined as:

$$R(\hat{\beta}) = E_Y[L(\hat{\beta}_Y) - L(\beta)] = E_Y\frac{1}{n}\|X\hat{\beta}_Y - X\beta\|^2 := E_Y\frac{1}{n}\|\hat{Y} - Y^*\|^2$$

where $Y^* = X\beta$ and $\hat{Y} = X\hat{\beta}$. Denoting,

$$\Sigma := \frac{1}{n}X^\top X$$

we can write the risk as:

$$R(\hat{\beta}) = E_Y(\hat{\beta} - \beta)^\top\Sigma(\hat{\beta} - \beta) := E_Y\|\hat{\beta} - \beta\|_\Sigma^2$$

Another interpretation of the risk is how well we accurately learn the parameters of the model.

3.1 Risk Bounds for Least Squares

The least squares estimator using an outcome Y is just:

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

The first derivative condition is that:

$$X^\top (Y - X\hat{\beta}) = 0$$

which is sometimes referred to as the *normal equations*.

The least squares estimator is then:

$$\hat{\beta} = \frac{1}{n}\Sigma^{-1}X^\top Y$$

Equivalently,

$$\hat{\beta} = X^\dagger Y$$

where $X^\dagger$ is the pseudo-inverse.

Also note that that:

$$\hat{Y} = \Pi Y$$

where Π is the orthogonal projection operator $\frac{1}{n}X\Sigma^{-1}X^\top$. Note that:

$$\Pi = UU^\top$$

where $U \in \mathbb{R}^{d\times p}$ is the left matrix from the SVD of X (with orthogonal columns).

It is straightforward to see that:

$$R(\hat{\beta}) = E_Y \frac{1}{n}\|\Pi E[Y] - \Pi Y\|^2$$

and

$$\beta = X^\dagger \mathbb{E}[Y] = \mathbb{E}[\hat{\beta}],\ Y^* = \Pi E[Y] = \mathbb{E}[\hat{Y}]$$

Lemma 3.1. *(Risk Bound) If* $\mathrm{Var}(Y_i) = \sigma^2$ *(i.e. the noise is homoskedastic), we have that:*

$$R(\hat{\beta}_\lambda) = \frac{d}{n}\sigma^2$$

If $\mathrm{Var}(Y_i) \le \sigma^2$*, then:*

$$R(\hat{\beta}_\lambda) \le \frac{d}{n}\sigma^2$$

Proof. Note that we can write:

$$Y = E[Y] + \eta$$

where $\mathbb{E}[\eta] = \vec{0}$. Since $Y^* = \Pi\mathbb{E}[Y]$

$$\|Y^* - \Pi Y\|^2 = \|\Pi\mathbb{E}[Y] - (\Pi\mathbb{E}[Y] + \Pi\eta)\|^2 = \|\Pi\eta\|^2$$

Hence the risk is:

$$\frac{1}{n}\mathbb{E}\|\Pi\eta\|^2 = \frac{1}{n}\mathbb{E}\eta UU^\top\eta = \frac{d}{n}\sigma^2$$

The second claim follows from using an inequality in the second to last step. □

3.2 What about high probability bounds?

Note that we have shown that:

Lemma 3.2. *We have that:*

$$L(\hat{\beta}) - L(\beta) = \frac{1}{n}\|Y^* - \Pi Y\|^2 = \frac{1}{n}\|\Pi\eta\|^2$$

where η is the noise vector on Y.

So if were interested in:

$$\Pr(L(\hat{\beta}) - L(\beta) \geq \epsilon)$$

The this is equivalent to understanding the following tail bound:

$$\Pr(\frac{1}{n}\|\Pi\eta\|^2 \geq \epsilon)$$

which will examine later.

4 What about if $d > n$?

If $d > n$, the risk of the least squares estimator is not useful (as $\hat{Y} = Y$). There are two common approaches we seek to understand in detail:

- Regularization. The idea is to "shrink" β in a certain manner to reduce variance (and increase bias).
- Feature Selection. The idea is to fit β only in certain directions (and exclude other irrelevant directions).

5 Ridge Regression

5.1 Bias Variance in the Fixed Design Setting

Lemma 5.1. *(bias-variance for risk) We can decompose the expected risk as:*

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

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

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

5.2 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 5.2. *(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. in next class □