MML lecture extra notes, week Nov 1 - 5, 2021

Bias-variance trade-off in linear/ridge regression

Below we show that, when assuming no model mismatch and in-distribution test settings, there exist choices of $\lambda > 0$ such that ridge regression returns smaller expected test error when compared with linear regression. For the assumption of no model mismatch, this means the training dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ is generated using a (noisy) underlying function that has the same form as the model:

$$y_n = f(\boldsymbol{x}_n; \boldsymbol{\theta}_0) + \epsilon_n, \quad f(\boldsymbol{x}, \boldsymbol{\theta}_0) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\theta}_0, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

In other words, there is a "ground truth" parameter $\boldsymbol{\theta}_0$ governing the generation of training data. This $\boldsymbol{\theta}_0$ parameter is unknown to us. We also denotes the above process as $\mathcal{D} \sim p_{data}^N$. The "in-distribution test" setting means the test data $(\boldsymbol{x}_{test}, y_{test})$ is also generated from the same process, i.e. $(\boldsymbol{x}_{test}, y_{test}) \sim p_{data}$.

Now we wish to learn the parameters $\boldsymbol{\theta}$ for our model $f(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\theta}$ using training data $\mathcal{D} \sim p_{data}^N$. For both linear regression and ridge regression, the minimiser of the loss function depends on $\mathcal{D}$. Therefore in the analysis we will write an estimator as $\boldsymbol{\theta}^*(\mathcal{D})$ to emphasise the dependency on training data.

We will look into the expected test error to understand how well the model performs; here the expectations are taken on both the training data $\mathcal{D} \sim p_{data}^N$ and the test data $(\boldsymbol{x}_{test}, y_{test}) \sim p_{data}$. Derivations show that for an estimator $\boldsymbol{\theta}^*$ which might not necessarily equal to the ground truth $\boldsymbol{\theta}_0$, the expected test error is related to the parameter estimation error:

$$\begin{aligned} error_{pred}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[\mathbb{E}_{(\boldsymbol{x}_{test}, y_{test}) \sim p_{data}}[||y_{test} - f(\boldsymbol{x}_{test}; \boldsymbol{\theta}^*(\mathcal{D}))||_2^2]] \\ &= \mathbb{E}_{\boldsymbol{x}_{test}}[\phi(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}^*) \phi(\boldsymbol{x}_{test})] + \sigma^2, \end{aligned} \tag{2}$$

$$\begin{aligned} Error(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[(\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0)(\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0)^\top] \\ &:= \mathbf{b}(\boldsymbol{\theta}^*)\mathbf{b}(\boldsymbol{\theta}^*)^\top + \mathbf{V}(\boldsymbol{\theta}^*), \end{aligned} \tag{3}$$

$$\begin{aligned} \text{bias:} \quad \mathbf{b}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})] - \boldsymbol{\theta}_0 \\ \text{variance:} \quad \mathbf{V}(\boldsymbol{\theta}^*) &= \mathbb{V}_{\mathcal{D} \sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})]. \end{aligned} \tag{4}$$

We can show that smaller parameter estimation error leads to smaller expected prediction error: for two estimators $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, using properties of positive semi-definite matrices, we have:

$$Error(\boldsymbol{\theta}_1) \preceq Error(\boldsymbol{\theta}_2) \quad \Rightarrow \quad error_{pred}(\boldsymbol{\theta}_1) \leq error_{pred}(\boldsymbol{\theta}_2).$$

So it remains to find settings of $\lambda > 0$ for ridge regression such that it achieves a smaller parameter estimation error when compared with linear regression. Note that when $\lambda = 0$ it corresponds to linear regression. The bias and variance of the ridge regression estimator are:

$$\mathbf{b}(\boldsymbol{\theta}_R^*) = \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] - \boldsymbol{\theta}_0 = -\sigma^2 \lambda (\sigma^2 \lambda \mathbf{I} + \Phi^\top \Phi)^{-1} \boldsymbol{\theta}_0 := \mathbf{b}(\lambda), \tag{5}$$

$$\mathbf{V}(\boldsymbol{\theta}^*) = \sigma^2 (\sigma^2 \lambda \mathbf{I} + \Phi^\top \Phi)^{-1} \Phi^\top \Phi (\sigma^2 \lambda \mathbf{I} + \Phi^\top \Phi)^{-1} := \mathbf{V}(\lambda).$$

The expressions indicate that linear regression returns an *unbiased estimator* of $\boldsymbol{\theta}_0$ as $\mathbf{b}(\lambda) = 0$ when $\lambda = 0$. By contrast, ridge regression ($\lambda > 0$) returns a biased estimator. Therefore the search for $\lambda > 0$ such that $Error(\boldsymbol{\theta}_R^*) \preceq Error(\boldsymbol{\theta}_L^*)$ is equivalent to searching for λ such that $\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0)$. After some linear algebra, we have:

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) - \mathbf{V}(0) = -\sigma^2 \lambda (\Phi^\top \Phi + \sigma^2 \lambda \mathbf{I})^{-1} \underbrace{(\sigma^2 [2\mathbf{I} + \sigma^2 \lambda (\Phi^\top \Phi)^{-1}] - \sigma^2 \lambda \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top)}_{:= \mathbf{E}} (\Phi^\top \Phi + \sigma^2 \lambda \mathbf{I})^{-1}. \tag{6}$$

Furthermore, one can show that

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0) \quad \Leftrightarrow \quad \mathbf{E} \text{ is positive semi-definite}, \tag{7}$$

which can be achieved by e.g. setting $0 \leq \lambda \leq \frac{2}{||\boldsymbol{\theta}_0||_2^2}$. To see this, first notice that in eq. (6) $\mathbf{E}$ is left- and right-multiplied by the same matrix, which supports the claim in eq. (7). Then a close inspection of $\mathbf{E}$ shows that if we make $2\mathbf{I} - \lambda \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top$ positive semi-definite then $\mathbf{E}$ will also be positive semi-definite. As $\boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top$ is a rank-1 matrix, the only non-zero eigenvalue of $\boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top$ is $||\boldsymbol{\theta}_0||_2^2$. Using the discussed indications of eigen-decomposition, we can show that $2\mathbf{I} - \lambda \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top$ is positive semi-definite when $0 \leq \lambda \leq \frac{2}{||\boldsymbol{\theta}_0||_2^2}$.

One can also show that $\mathbf{V}(\lambda) \preceq \mathbf{V}(0)$ for $\lambda > 0$:

$$\mathbf{V}(\lambda) - \mathbf{V}(0) = -\sigma^2 \lambda (\Phi^\top \Phi + \sigma^2 \lambda \mathbf{I})^{-1} \underbrace{(\sigma^2 [2\mathbf{I} + \sigma^2 \lambda (\Phi^\top \Phi)^{-1}])}_{:= \tilde{\mathbf{E}}} (\Phi^\top \Phi + \sigma^2 \lambda \mathbf{I})^{-1} \preceq 0,$$

because $\tilde{\mathbf{E}}$ is positive semi-definite. Combining both results, we see that ridge regression is useful in reducing the variance of parameter estimation, but this is in the price of increased bias. Therefore λ needs to be selected carefully (e.g. $0 \leq \lambda \leq \frac{2}{||\boldsymbol{\theta}_0||_2^2}$) such that the bias is not too large, and at the same time the variance of parameter estimation is reduced.

Solving PCA optimisation problems

Minimum reconstruction error perspective We have shown in the lecture that the PCA algorithm aims to find an orthonormal basis $\mathbf{B}_{full}$ which minimises the reconstruction error

$$L = \sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2, \quad \tilde{\boldsymbol{x}}_n = \sum_{j=1}^{M} z_{nj} \mathbf{b}_j, \quad M < D. \tag{8}$$

A few derivations show that

$$L = \sum_{j=M+1}^{D} \mathbf{b}_j^\top \underbrace{\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top}_{\mathbf{S}} \mathbf{b}_j, \tag{9}$$

where by plugging-in the eigen-decomposition of $\mathbf{S} = \mathbf{Q} \Lambda \mathbf{Q}^\top$ we have the optimisation problem as

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \boldsymbol{\beta}_j^\top \Lambda \boldsymbol{\beta}_j, \quad \boldsymbol{\beta}_j = \mathbf{Q}^\top \mathbf{b}_j, \quad \text{subject to } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j. \tag{10}$$

Now notice that $||\boldsymbol{\beta}_j||_2^2 = \mathbf{b}_j^\top \mathbf{Q} \mathbf{Q}^\top \mathbf{b}_j = \mathbf{b}_j^\top \mathbf{b}_j = 1$ since $\mathbf{Q} = [\boldsymbol{q}_1, ..., \boldsymbol{q}_D]$ represents an orthonormal basis. This means $\boldsymbol{\beta}_j^\top \Lambda \boldsymbol{\beta}_j, \quad \boldsymbol{\beta}_j = \sum_{d=1}^{D} \beta_{jd}^2 \lambda_d$ is a weighted sum of the eigenvalues $\{\lambda_1 \geq ... \geq \lambda_D\}$ and the weights $\{\beta_{jd}^2\}$ sum to 1. Therefore, we can conduct the following reasoning to iteratively solve the optimisation problem, using *proof by induction*:

1. For $j = D$, we can show that $\boldsymbol{\beta}_D^\top \Lambda \boldsymbol{\beta}_D$ is minimised by choosing $\mathbf{b}_D = \boldsymbol{q}_D$. This is done by choosing $\boldsymbol{\beta}_D = [0, ..., 0, 1]^\top$ which minimises the quadratic loss.

2. For each $j = D - 1, ..., M + 1$:

   - Assume we have obtained solutions $\mathbf{b}_i = \boldsymbol{q}_i$ for $i > j$;
   - As we can write $\mathbf{b}_j = \sum_{d=1}^{D} \beta_{jd} \boldsymbol{q}_d$, to make sure that $\mathbf{b}_j \perp \mathbf{b}_i$ for $i > j$, this means $\mathbf{b}_j^\top \mathbf{b}_i = \beta_{ji} = 0$;
   - So we seek for the other β_{jd} values ($d \leq j$) such that $\sum_{d=1}^{j} \beta_{jd}^2 \lambda_d$ is minimised. Notice that the weights for $\{\lambda_d\}$ sum to one. This leads to $\mathbf{b}_j = \boldsymbol{q}_j$, i.e. $\beta_{jj} = 1$ and $\beta_{jd} = 0$ for $d \neq j$.

3. Using proof by induction, we can show that the optimal solution is $\mathbf{b}_j = \boldsymbol{q}_j$ for $j = M + 1, ..., D$.

Maximum variance perspective The PCA algorithm can also be viewed as solving a sequence of optimisation problems to find the projection directions that maintain maximum variance. In detail, for each $m = 1, ..., M$, we have shown in the lecture that the corresponding constrained optimisation problem is

$$\max_{\mathbf{b}_m} \mathbb{V}[\mathbf{b}_m^\top \hat{\boldsymbol{x}}_n], \quad \hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \sum_{j=1}^{m-1} (\mathbf{b}_j^\top \boldsymbol{x}_n)\mathbf{b}_j, \quad \text{subject to } ||\mathbf{b}_m||_2^2 = 1, \mathbf{b}_m \perp \mathbf{b}_j, j < m. \tag{11}$$

In other words, PCA iteratively finds the "maximum variance directions" in the remainder information. Note that we have shown in the lecture that $\mathbb{V}[\mathbf{b}_m^\top \hat{\boldsymbol{x}}_n] = \mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n]$ which is due to the constraint of orthonormal basis. Also notice that $\mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n] = \mathbf{b}_m^\top \mathbf{S} \mathbf{b}_m = \sum_{d=1}^D \beta_{md}^2 \lambda_d$. So we apply the *proof by induction* technique again and solve the optimisation tasks as follows:

1. For $m = 1$, we can show that $\mathbf{b}_1^\top \mathbf{S} \mathbf{b}_1$ is minimised by choosing $\mathbf{b}_1 = \boldsymbol{q}_1$. The argument here is similar to that of step 1 in solving the reconstruction error minimisation problem.

2. For each $m = 2, ..., M$:

   - Assume we have obtained solutions $\mathbf{b}_i = \boldsymbol{q}_i$ for $i < m$;
   - As we can write $\mathbf{b}_m = \sum_{d=1}^D \beta_{md} \boldsymbol{q}_d$, to make sure $\mathbf{b}_m \perp \mathbf{b}_i$ for $i < m$, this means $\mathbf{b}_m^\top \mathbf{b}_i = \beta_{mi} = 0$;
   - So we seek for the other β_{md} values ($d \geq m$) such that $\sum_{d=m}^D \beta_{md}^2 \lambda_d$ is maximised. Notice that the weights for $\{\lambda_d\}$ sum to one. This leads to $\mathbf{b}_m = \boldsymbol{q}_m$, i.e. $\beta_{mm} = 1$ and $\beta_{md} = 0$ for $d \neq m$.

3. Using proof by induction, we can show that the optimal solution is $\mathbf{b}_m = \boldsymbol{q}_m$ for $m = 1, ..., M$.

Remark The above derivations for both perspectives solve a constrained optimisation problem in $\mathbf{b}$ space, by rewriting the problem as (a sequence of) constrained optimisation problem in β_{jd} space. The constrain in such case is simple (β_{jd}^2 sum to one) so solutions can be obtained fairly easily. In future lectures we will discuss constrained optimisation techniques and revisit the PCA optimisation example; using such techniques we can solve the PCA optimisation task jointly for all the principle components.

Remark The two perspectives of PCA, although resulting in the same projections, do not necessarily need the usage of the same $\mathbf{B}_{full}$ at optimum. From the minimum reconstruction error perspective, one just need to make sure that $\boldsymbol{x}_n$ is projected to the orthogonal complement space $span(\{\boldsymbol{q}_j\}_{j=M+1}^D)^\perp$, and for such space, $\{\boldsymbol{q}_m\}_{m=1}^M$ is not the only orthonormal basis. In other words, the minimum reconstruction error perspective just requires the optimal $\mathbf{B}_{full} = \{\mathbf{b}_1, ..., \mathbf{b}_M, \boldsymbol{q}_{M+1}, ..., q_D\}$ with $span(\{\mathbf{b}_1, ..., \mathbf{b}_M\}) = span(\{\boldsymbol{q}_j\}_{j=M+1}^D)^\perp$. Similarly, one can show that for the maximum variance perspective, the optimal $\mathbf{B}_{full} = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_M, \mathbf{b}_{M+1}, ..., \mathbf{b}_D\}$ with $span(\{\mathbf{b}_{M+1}, ..., \mathbf{b}_D\}) = span(\{\boldsymbol{q}_m\}_{m=1}^M)^\perp$. In practice we will use $\mathbf{B}_{full} = \mathbf{Q}$ though as a convention.

An extra exercise

Q1: Convergence analysis of constant step-size gradient descent (GD) for ridge regression:

1. Show that if GD converges, it would converge to $\boldsymbol{\theta}_R^*$.

2. Derive the "safe threshold" for the constant step size γ.

Solution of Q1:

The iterative update of GD for ridge regression is:

$$\boldsymbol{\theta}_{t+1} = ((1 - \gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top \mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top \mathbf{y}. \tag{12}$$

Solving the corresponding geometric sequence returns

$$\boldsymbol{\theta}_t = ((1 - \gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top \mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_R^*) + \boldsymbol{\theta}_R^*, \tag{13}$$

where $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I}+\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the minimiser of the loss function. Therefore it means GD, if converges, converges to the right solution. And GD converges if $((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_R^*) \to \mathbf{0}$.

Applying the analysis techniques of GD for linear regression, we see that it reduces to investigate the eigenvalues of matrix $((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^2$. Therefore we would like to make sure that

$$\lambda_{max} := \lambda_{max}(((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^2) = \max_{\lambda_x}(1-\gamma\lambda - \frac{\gamma}{\sigma^2}\lambda_x)^2 < 1, \tag{14}$$

where λ_x denotes possible eigenvalue of $\mathbf{X}^\top\mathbf{X}$. Therefore the “safe threshold” for step size selection is

$$\gamma < 2(\lambda + \lambda_{max}(\mathbf{X}^\top\mathbf{X})/\sigma^2)^{-1}. \tag{15}$$