HOMEWORK 3 - V0

CSC2532 WINTER 2024

University of Toronto

VERSION HISTORY: V0 → V1:

- **Deadline:** Apr 5, by 23:59.
- **Submission:** You need to submit your solutions through Crowdmark, including all your derivations, plots, and your code. You can produce the file however you like (e.g. LaTeX, Microsoft Word, etc), as long as it is readable. Points will be deducted if we have a hard time reading your solutions or understanding the structure of your code.

1. Stieltjes Transform and Double descent - 30 pts. In the lecture, as $d/n \to \gamma$, we proved that the risk of ridge regression can be written as

$$\text{Risk}(\lambda) = \text{V}(\lambda) + \text{B}(\lambda), \tag{1.1}$$

where the variance and the bias terms are given as

$$\begin{aligned} \text{V}(\lambda) \to& \sigma^2\gamma\{s(-\lambda) - \lambda s'(-\lambda)\} \\ \text{B}(\lambda) \to& \lambda^2 s'(-\lambda) \end{aligned}$$

with $s(z) = \int \frac{1}{x-z} d\mu(z)$ denoting the Stieltjes transform of the M-P law (explicit form given in lecture). Compute the risk of *ridgeless* regression as $\lambda \to 0_+$ by deriving expressions for V(0+) and B(0+). Plot the bias, variance and the risk as a function of γ (No need to submit code).

2. Implicit bias and Double descent- 70 pts. We have n data points $\{(\boldsymbol{x}_i, y_i)\}$, each of which is a pair of feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ and corresponding label y_i, and our goal is to find some parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ that accurately predicts a linear relation between the features and the label. We do so by minimizing the squared difference between the predictions of our linear model and the labels, summed over n data points, i.e., the least squares objective

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{R}(\boldsymbol{\theta}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$$

where $\boldsymbol{y} = (y_i) \in \mathbb{R}^n$ is the response, $\boldsymbol{X} = (\boldsymbol{x}_i) \in \mathbb{R}^{n\times d}$ is the feature matrix, and $\boldsymbol{\theta}$ is the least squares parameter. We assume that the data matrix is not degenerate, i.e., $\text{rank}(\boldsymbol{X}) = \min\{n, d\}$. This implies that when $n > d$, then $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible, and when $n < d$, $\boldsymbol{X}\boldsymbol{X}^\top$ is invertible.

Since we have n data points, and we aim to learn d parameters from data, we know that when $n < d$, the problem is underdetermined since we have more parameters than data points; we refer to this setting as the overparameterized regime; conversely, the underparameterized regime refers to the $n > d$ setting.

We solve this problem with gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}_t = -\nabla\hat{R}(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_0 = 0, \tag{2.1}$$

where $\nabla\hat{R}(\boldsymbol{\theta}) = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})$.

1. **Underparametrized regime:** Assume $n > d$ and set $\lambda_{\min/\max} = \lambda_{\min/\max}(\boldsymbol{X}^\top \boldsymbol{X}) > 0$ so the problem is strongly convex. Prove that

$$\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}\|^2 \le e^{-\mu t}\|\hat{\boldsymbol{\theta}}\|^2 \quad \text{with} \quad \hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

where $\mu = 2\lambda_{\min}$. Remark: A similar result also holds for the gradient descent.
2. **Overparametrized regime:** When $n < d$, we have that $\lambda_{\min} = 0$, thus the objective is no longer strongly convex (still convex). Since in this case, the equation $\boldsymbol{X}\theta = \boldsymbol{y}$ is underdetermined, there can be infinitely many solutions achieving zero loss: $\hat{R}(\boldsymbol{\theta}) = 0$. However, as it turns out, GF (starting from 0) has some implicit bias and does not return an arbitrary zero-loss solution.
Prove that (2.1) at $t = \infty$ returns the min-norm solution

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|^2 \quad \text{such that} \quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}.$$

In other words, in the overparameterized setting, GF finds the zero-loss solution with the smallest Euclidean norm. This phenomenon is called *implicit bias*. Hint: GF solution is always spanned by the rows of $\boldsymbol{X}$ for all t.
3. Conclude that GF finds the following solutions to the least squares objective

$$\boldsymbol{\theta}^\infty = \begin{cases} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}, & n > d \\ \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}, & n < d. \end{cases} \tag{2.2}$$

4. (Digression) Prove that the ridge regression solution $\boldsymbol{\theta}(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ in the overparametrized regime converges to the same minimum norm solution in the limit $\lambda \to 0_+$. This is what we analyzed in the lecture as well as Problem 1 above.
5. The above calculations do not rely on a particular statistical model. In what follows, we will assume that the data generating process satisfies

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}_* \rangle + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

where ϵ_i is independent of $\boldsymbol{x}_i$. If we assume that the features are Gaussian $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$, show that the population risk $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}[(y - \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)^2]$ of any (possibly random) $\hat{\boldsymbol{\theta}}$ is

$$\mathcal{R}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|^2] + \sigma^2.$$

Thus the excess risk is

$$\mathcal{ER}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|^2].$$

6. Using the explicit form of the GF solution (2.2), prove that

$$\mathcal{ER}(\boldsymbol{\theta}^\infty) = \mathbb{E}[\|\boldsymbol{\theta}^\infty - \boldsymbol{\theta}_*\|^2] = \begin{cases} \sigma^2 \mathbb{E}[\mathrm{Tr}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big)], & n > d + 1 \\ \frac{d-n}{d}\|\boldsymbol{\theta}_*\|^2 + \sigma^2 \mathbb{E}[\mathrm{Tr}\big((\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\big)], & n < d - 1 \end{cases}$$

Hint: In the case $d > n$, you will need to compute $\mathbb{E}[\boldsymbol{P}_R]$ where $\boldsymbol{P}_R = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$ is the projection matrix to the row space of the Gaussian matrix $\boldsymbol{X}$. Note that Gaussian matrices are rotationally invariant, i.e. $\boldsymbol{X} \overset{d}{=} \boldsymbol{X}\boldsymbol{Q}$ for any unitary matrix $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$. Due to this property, if we write the EVD of $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$, the (diagonal) matrix $\boldsymbol{D}$ containing the eigenvalues is independent of the matrix $\boldsymbol{V}$. This in hand, show that the projection matrix is given as $\boldsymbol{P}_R = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^T$ where $\boldsymbol{S}$ is a $d \times d$ diagonal matrix with entries either 0 or 1, with trace n. Argue that $\boldsymbol{S}$ and $\boldsymbol{V}$ are independent, and by symmetry, $\mathbb{E}[\boldsymbol{S}] = \frac{n}{d}\boldsymbol{I}_d$.

7. Using the properties of the inverse Wishart matrices[1], show

$$\mathcal{E}\mathcal{R}(\boldsymbol{\theta}^\infty) = \begin{cases} \sigma^2 \frac{d}{n-d-1}, & n > d+1 \\ \frac{d-n}{d}\|\boldsymbol{\theta}_*\|^2 + \sigma^2 \frac{n}{d-n-1}, & n < d-1. \end{cases}$$

8. Compare this non-asymptotic result to the asymptotic result obtained via M-P law. You may assume $\boldsymbol{\theta}_*$ is a multivariate Gaussian. Do you observe the same asymptotic behavior as $n, d \to \infty$ and $d/n \to \gamma$?

3. Course evaluations - 0 pts. Can you please fill out the course evaluations?

[1] https://en.wikipedia.org/wiki/Inverse-Wishart_distribution