Exercises 5: Learning in interpolation regimes

Exercise 1 (Interpolation of (S)GD for least squares)**.**

For a dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$, we consider linear regression by minimization of least squares criterion

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(X_i^\top \theta - Y_i \right)^2 = \frac{1}{n} \left\| \mathbb{X}\theta - Y \right\|_2^2$$

We assume that $d \gg n$, and that $\mathbb{X}\mathbb{X}^\top$ is invertible.

1. *Show that minimizers of the least square problem are solutions to the system $Y = \mathbb{X}\theta$.*

2. *Recall the expression of the interpolator $\theta^\star$ of minimal ℓ^2-norm.*

3. *Characterize the set of solution of the least squares problem, i.e., characterize how any solution of the system $Y = \mathbb{X}\theta$ can be written.*

   We want to run a gradient-based strategy to solve the least squares problem.

4. *Show that when an iterate θ_t of (S)GD strategies lives in $\operatorname{span}(X_1, \ldots, X_n) = \operatorname{Im}(\mathbb{X}^\top)$, then all the subsequent iterates stay in $\operatorname{span}(X_1, \ldots, X_n)$.*

5. *Call P the orthogonal projector on $\operatorname{span}(X_1, \ldots, X_n)$, and remark that the initial point θ_0 can be decomposed as follows*

   $$\theta_0 = P\theta_0 + (I - P)\theta_0.$$

   Call θ_∞ the limit iterate of (S)GD. Deduce that

   $$\theta_\infty = \theta^\star + (I - P)\theta_0.$$

6. *Show that*

   $$\theta_\infty = \operatorname{proj}_{\mathrm{sol}}(\theta_0) := \operatorname{argmin}_\theta \frac{1}{2} \left\| \theta - \theta_0 \right\|_2^2 \quad \textit{such that} \quad Y = \mathbb{X}\theta.$$

We have shown the following result.

Proposition 1. *In the case of overparameterized linear model, assuming that $\mathbb{X}\mathbb{X}^\top$ is invertible, (S)GD-based methods started from θ_0 converge to*

$$\theta_\infty = \operatorname{proj}_{\mathrm{sol}}(\theta_0) := \operatorname{argmin}_\theta \frac{1}{2} \left\| \theta - \theta_0 \right\|_2^2 \quad \textit{such that} \quad Y = \mathbb{X}\theta.$$

Exercise 2 (Separable data and exponential loss). *In a context of classification, we consider a set of training points $(X_i, Y_i)_{1\le i\le n}$ for $X_i \in \mathbb{R}^d$ and $Y_i \in \{-1; +1\}$. The training points are assumed to be linearly separable, i.e., there exists $\theta \in \mathbb{R}^d$ such that for all $1 \le i \le n$, $Y_i(X_i^\top \theta) > 0$ (we suppose that no bias is needed). In order to perform the classification, we want to train a linear classifier using an exponential-kind loss:*

$$\min_{\theta\in\mathbb{R}^d} \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{-Y_i X_i^\top \theta}\right)$$

or equivalently

$$\max_{\theta\in\mathbb{R}^d} -\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{-Y_i X_i^\top \theta}\right) =: F(\theta).$$

In the lecture, we study the implicit bias of the "gradient flow" algorithm, i.e.., we know that the gradient should diverge but the question is in which way. The goal of this exercise is to show an intermediate result.

Lemma 2. *The following holds:*

(i) $\min_i Y_i X_i^\top \theta \le F(\theta) \le \min_i Y_i X_i^\top \theta + \log(n)$ *for all* $\theta \in \mathbb{R}^d$*;*

(ii) $\|\nabla F(\theta)\|_2 \ge \gamma$ *for* γ *the* ℓ^2*-max margin, i.e.,*

$$\gamma := \max_{\|\theta\|_2\le 1} \min_{1\le i\le n} Y_i X_i^\top \theta.$$

1. *To demonstrate* (i)*, show that*

$$e^{-m_\theta} \ge \frac{1}{n}\sum_{i=1}^{n} e^{-Y_i X_i^\top \theta} \ge \frac{1}{n} e^{-m_\theta}$$

for $m_\theta := \min_i Y_i X_i^\top \theta$*, and conclude.*

2. *The rest of the exercise is dedicated to demonstrate* (ii)*. Call*

$$Z := \begin{pmatrix} Y_1 X_1^\top \\ \vdots \\ Y_n X_n^\top \end{pmatrix} \in \mathbb{R}^{n\times d}$$

and $\Delta_n = \left\{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\right\}$ *the simplex.*

   (a) ☕ *Check that*

$$\gamma = \max_{\|\theta\|_2\le 1} \min_{p\in\Delta_n} p^\top Z\theta = \min_{p\in\Delta_n} \max_{\|\theta\|_2\le 1} p^\top Z\theta.$$

   (The second equality can be justified by the minimax theorem for bilinear functions, but you can use your own way.)

   (b) *Deduce that*

$$\gamma = \min_{p\in\Delta_n} \left\|Z^\top p\right\|_2.$$

   (c) *Express the gradient* $\nabla F(\theta)$ *with the help of matrix* Z.

   (d) *Conclude.*