Math of ML : Exercises 6 *

October 27, 2025

Exercise 1 (Conjugate kernels for step and ReLU activations)**.** *Consider the two-layer neural network*

$$f(x; w, \eta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x),$$

where (w, η) are the weights of the network and σ is the activation function.

Suppose that the weights $w = (w_1, \dots, w_j)$ are sampled independently from the multivariate standard normal distribution $w_j \sim \mathcal{N}(0, I_d)$, where I_d is the $d \times d$ identity matrix.

When the weights w are held fixed and only the weights η are trained, training the neural network $f(x; w\eta)$ corresponds to fitting a kernel method with the feature map

$$\phi(x)_j = \frac{1}{\sqrt{m}} \sigma(w_j^\top x)$$

with the corresponding kernel $\widehat{k}_m(x, x') = \frac{1}{m} \sum_{j=1}^{m} \sigma(w_j^\top x)\sigma(w_j^\top x')$. By the law of large numbers, we have

$$\lim_{m \to \infty} \widehat{k}_m(x, x') = k(x, x') = \mathbf{E}_{w \sim \mathcal{N}(0, I_d)} \left[\sigma(w^\top x)\sigma(w^\top x'). \right]$$

The kernel k is called the conjugate kernel. *Fix any x, x' and let $\theta = \arccos(x^\top x'/(\|x\|_2 \|x'\|_2))$. Prove the following closed-form expressions for k.*

1. *For the step activation function $\sigma(x) = 1$ if $x \geq 0$ and $\sigma(x) = 0$ if $x < 0$ we have*

$$k(x, x') = \frac{1}{2\pi}(\pi - \theta).$$

2. ($\star$) *For the ReLU activation function $\sigma(x) = \max\{0, x\}$ we have*

$$k(x, x') = \frac{\|x\|_2 \|x'\|_2}{2\pi} \left(\sin\theta + (\pi - \theta)\cos\theta \right). \tag{1}$$

Hint: *use the fact that for any orthogonal matrix P we have $k(x, x') = k(Px, Px')$.*

Exercise 2 (Two-layer neural networks as kernel methods)**.** *Consider the setting of Exercise 1. Consider sampling input vectors $(x, 1) \in \mathbb{R}^2$, where $x \sim \text{Uniform}([0, 1])$. Conditionally on the input vector $(x, 1)$, generate a response variable $y = f^*(x) + \mathcal{N}(0, 0.1)$ for the choices $f^*(x) = x, f^*(x) = 1$ and $f^*(x) = \sin(2\pi x)$. For each choice of f^*, generate a dataset $(x_i, y_i)_{i=1}^n$ of size $n = 100$. Fit the kernel ridge regression estimator $\widehat{f}_\infty$ for the conjugate kernel of ReLU network (1). For $m = 5, 10, 100, 1000$, fit the kernel ridge regression estimator $\widehat{f}_m$ using the kernel $\widehat{k}_m$ (cf. Exercise 1) with ReLU activation. Plot the learned functions $\widehat{f}_m$ (for $m = 5, 10, 100, 1000, \infty$) and the function f^* used to generate the data.*

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

The next exercise is not about kernels nor neural networks. It introduces the notion of duality, which plays a central role in convex optimization, and will be useful later in the course.

Exercise 3 (KKT Conditions)**.** *Consider the convex optimization problem:*

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} f(x) \\ &\text{subject to} \quad f_i(x) \leq 0 \text{ for } i = 1, \dots, m, \end{aligned} \tag{2}$$

where the functions $f, f_1, \dots, f_m : \mathbb{R}^d \to \mathbb{R}$ *are convex and differentiable.*

1. *The Lagrangian* $L : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ *is defined by* $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i f_i(x)$*. Prove that the problem* (2) *can be reformulated as*

$$\inf_{x\in\mathbb{R}^d} \sup_{\lambda\geq 0} L(x, \lambda), \tag{3}$$

   where the inequality $\lambda \geq 0$ *is to be interpreted coordinate-wise. We will refer to the problem* (3) *as the primal problem.*

2. *Denote the optimal (minimum) value attained by the primal minimization problem* (3) *by* $p^* \in \mathbb{R} \cup \{\pm\infty\}$*. We now introduce the dual problem, the maximization problem in* $\lambda \geq 0$ *defined by*

$$\sup_{\lambda\geq 0} \inf_{x\in\mathbb{R}^d} L(x, \lambda). \tag{4}$$

   Denote the optimal (maximum) value attained by the dual maximization problem (4) *by* $d^* \in \mathbb{R} \cup \{\pm\infty\}$*. Prove that* $d^* \leq p^*$ *(weak duality;* $p^* - d^*$ *is called the duality gap).*

3. *Prove that* $x^* \in \mathbb{R}^d$ *is optimal for* (3)*,* $\lambda^* \in \mathbb{R}^m$ *is optimal for* (4)*, and* $p^* = d^*$ *(strong duality holds)* if and only if *the pair* (x^*, λ^*) *satisfies the Karush-Kuhn-Tucker (KKT) conditions:*

   (a) *(Stationarity)* $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) = 0$*.*

   (b) *(Primal feasibility)* $f_i(x^*) \leq 0$ *for* $i = 1, \dots, m$*.*

   (c) *(Dual feasibility)* $\lambda_i^* \geq 0$ *for* $i = 1, \dots, m$*.*

   (d) *(Complementary slackness)* $\lambda_i^* f_i(x_i^*) = 0$ *for* $i = 1, \dots, m$*.*

   Remark: Strong duality can be ensured under so-called constraint qualification conditions. One such condition, called Slater's condition*, is the existence of* $x \in \mathbb{R}^d$ *such that* $f_i(x) < 0$ *for* $i = 1, \dots, m$ *(i.e., the existence of a feasible solution to* (2) *such that all constraints are satisfied with strict inequalities).*

References

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.