

# Math of ML : Exercises 9 \*

November 17, 2025

This exercise sheet follows the notation introduced in the previous week's exercises.

**Exercise 1** (Rademacher complexity of two-layer neural networks). *We will prove an upper bound on the Rademacher complexity of the space of functions  $\mathcal{F}_1(c) = \{f \in \mathcal{F}_1 : \|f\|_{\mathcal{F}_1} \leq c\}$ , where  $c > 0$ . In this exercise, in the definition<sup>1</sup> of the function space  $\mathcal{F}_1$  we let the activation function  $\sigma$  be the relu  $\sigma(x) = \max(0, x)$  and we define  $\mathcal{S} = \{w \in \mathbb{R}^d : \|w\|_2 \leq c_w\} \times \{b \in \mathbb{R} : |b| \leq c_b\}$  for some constants  $c_w, c_b > 0$ .*

1. Let  $g \in \mathbb{R}^n$  and let  $x_1, \dots, x_n \in \mathcal{X}_r$  be fixed. Prove that

$$\sup_{f \in \mathcal{F}_1(c)} \frac{1}{n} \sum_{i=1}^n g_i f(x_i) = c \sup_{(w,b) \in \mathcal{S}} \left| \frac{1}{n} \sum_{i=1}^n g_i \sigma(x_i^\top w + b) \right|.$$

2. Let  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$  be a sequence of i.i.d. Rademacher random variables (i.e.,  $\{\pm 1\}$  valued symmetric random variables). Let  $x_1, \dots, x_n \in \mathcal{X}_r$  be fixed. Prove that

$$\mathbf{E}_\varepsilon \left[ \sup_{f \in \mathcal{F}_1(c)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right] \leq \frac{c(c_w r + c_b)}{\sqrt{n}}.$$

In particular, the Rademacher complexity of neural networks with weights  $(w_j, b_j) \in \mathcal{S}$  is independent of the number of neurons. Instead, it depends on the  $\ell_1$  norm of the output layer weights  $\|\eta\|_1$ .

**Exercise 2** (Weight decay regularizes the the total variation). *In this exercise, fix the following setup: let  $\mathcal{S} = \{(w, b) : \|w\|_2^2 + b^2 = 1\}$  and let  $\sigma(u) = \max(u, 0)$  be the relu activation. The purpose of this exercise is to show that when training both layers of a two-layer neural network, ridge-type regularization (called weight decay in neural networks literature) penalizes the  $\|\cdot\|_{\mathcal{F}_1}$  norm.*

Let  $(x_1, y_1), \dots, (x_n, y_n)$  be the observed dataset. For any neural network weights  $w = (w_1, \dots, w_m)$ ,  $b = (b_1, \dots, b_m)$  and  $\eta = (\eta_1, \dots, \eta_m)$ , define the two-layer neural network by

$$f_{(w,b,\eta)}(x) = \sum_{j=1}^m \eta_j \sigma(w_j^\top x + b_j).$$

Consider the following regularized empirical risk objective:

$$\widehat{L}_\lambda(w, b, \eta) = \frac{1}{n} \sum_{i=1}^n \ell(f_{(w,b,\eta)}(x_i), y_i) + \lambda \left( \sum_{j=1}^m \|w_j\|_2^2 + b_j^2 + \eta_j^2 \right), \quad (1)$$

---

\*Lénaïc Chizat EPFL [lenaic.chizat@epfl.ch](mailto:lenaic.chizat@epfl.ch)

<sup>1</sup>Recall that the normed function space  $(\mathcal{F}_1, \|\cdot\|_{\mathcal{F}_1})$  is defined via

$$\|f\|_{\mathcal{F}_1} = \inf_{\mu \in \mathcal{M}(\mathcal{S})} \left\{ \|\mu\|_{\text{TV}} : \forall x \in \mathcal{X}_r, f(x) = \int_{\mathcal{S}} \sigma(w^\top x + b) d\mu(w, b) \right\}.$$

where  $\ell$  is an arbitrary the loss function.

Suppose that  $(w^*, b^*, \eta^*)$  is any global minimizer of the objective  $\widehat{L}_\lambda$  defined above. Prove that

$$\|f_{(w^*, b^*, \eta^*)}\|_{\mathcal{F}_1} \leq \frac{1}{2} \sum_{j=1}^m \|w_j^*\|_2^2 + (b^*)_j^2 + (\eta^*)_j^2.$$

In particular, when we train both layers of a neural network, the ridge-type regularization in (1) penalizes the  $\|\cdot\|_{\mathcal{F}_1}$  norm.

**Exercise 3** (Simulations). Implement (from scratch) code for training two-layer relu neural networks with stochastic gradient descent. Using your code, reproduce the figures in [Bach, 2024][Section 9.4].

## References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Francis Bach. *Learning theory from first principles*. MIT press, 2024. URL [https://www.di.ens.fr/~fbach/lfp\\_book.pdf](https://www.di.ens.fr/~fbach/lfp_book.pdf).