Math of ML : Exercises 7 *

November 3, 2025

Before attempting the exercises, read the following blog post by Francis Bach: https://francisbach.com/quest-for-adaptivity/.

In this exercise sheet, let $\tau \in \mathcal{P}(\mathcal{S})$ be the law of the first neural network layer initialization weights (w_j, b_j), where $\mathcal{S} \subseteq \mathbb{R}^d \times \mathbb{R}$ is the support of τ. Further, let σ be the activation function and let the input space be $\mathcal{X}_r = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ for some constant $r > 0$.

Let $\mathcal{M}(\mathcal{S})$ denote the set of signed measures on $\mathcal{S}$ equipped with Borel sigma algebra. The total variation norm of a measure $\mu \in \mathcal{M}(\mathcal{S})$ is defined by

$$\|\mu\|_{\mathrm{TV}} = \sup_{|f| \leq 1} \int_{\mathcal{S}} f(w,b) d\mu(w,b),$$

where the above supremum runs over all measurable functions f.

Following the 6-th week lecture notes, define the normed function spaces $(\mathcal{F}_1, \|\cdot\|_{\mathcal{F}_1})$ and $(\mathcal{F}_2, \|\cdot\|_{\mathcal{F}_2})$ by

$$\|f\|_{\mathcal{F}_1} = \inf_{\mu \in \mathcal{M}(\mathcal{S})} \left\{ \|\mu\|_{\mathrm{TV}} : \forall x \in \mathcal{X}_r \, f(x) = \int_{\mathcal{S}} \sigma(w^\top x + b) d\mu(w,b) \right\} \tag{1}$$

and

$$\|f\|_{\mathcal{F}_2}^2 = \inf_{\eta \in L_2(\tau)} \left\{ \int_{\mathcal{S}} \eta(w,b)^2 d\tau(w,b) : \forall x \in \mathcal{X}_r \, f(x) = \int_{\mathcal{S}} \eta(w,b)\sigma(w^\top x + b) d\tau(w,b) \right\}. \tag{2}$$

Exercise 1 (On the total variation norm)**.**

1. *Let $\mu \in \mathcal{M}(\mathcal{S})$ be given by the finite combination of atoms $\mu = \sum_{j=1}^n \eta_j \delta_{(w_j,b_j)}$, where $\delta_{(w_j,b_j)}$ denotes the Dirac measure. Prove that $\|\mu\|_{\mathrm{TV}} = \sum_{j=1}^n |\eta_j|$.*

2. *Suppose $\mu \in \mathcal{M}(S)$ has a density η with respect to the measure τ. Prove that $\|\mu\|_{\mathrm{TV}} = \int_{\mathcal{S}} |\eta(w,b)| d\tau(w,b)$.*

3. *Define the normed space $(\mathcal{F}_1', \|\cdot\|_{\mathcal{F}_1'})$ by*

$$\|f\|_{\mathcal{F}_1'} = \inf_{\eta \in L_1(\tau)} \left\{ \int_{\mathcal{S}} |\eta(w,b)| d\tau(w,b) : \forall x \in \mathcal{X}_r \, f(x) = \int_{\mathcal{S}} \eta(w,b)\sigma(w^\top x + b) d\tau(w,b) \right\}.$$

Suppose that $f \in \mathcal{F}_1'$. Prove that $f \in \mathcal{F}_1$ and $\|f\|_{\mathcal{F}_1} \leq \|f\|_{\mathcal{F}_1'}$.

Solution 1.

1. *By the definition of the total variation norm we have*

$$\|\mu\|_{\mathrm{TV}} = \sup_{|f| \leq 1} \int_{\mathcal{S}} f(w,b) d\mu(w,b) = \sup_{f(w_j,b_j) \in [-1,1], j=1,\ldots,m} \sum_{j=1}^n \eta_j f(w_j,b_j) = \sum_{j=1}^n |\eta_j|.$$

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

2. *We have*

$$\begin{aligned}\|\mu\|_{\mathrm{TV}} &= \sup_{|f|\leq 1} \int_{\mathcal{S}} f(w,b) d\mu(w,b) \\ &= \sup_{|f|\leq 1} \int_{\mathcal{S}} f(w,b)\eta(w,b) d\tau(w,b) \\ &\leq \int_{\mathcal{S}} |\eta(w,b)| d\tau(w,b). \end{aligned} \tag{3}$$

It remains to show that we can choose a measurable function $|f| \leq 1$ that realized the upper bound (3). Let $\mathcal{S}_+ = \{(w,b) \in \mathcal{S} : \eta(w,b) \geq 0\}$ and $\mathcal{S}_- = \{(w,b) \in \mathcal{S} : \eta(w,b) < 0\}$. Because η is a measurable function, the sets $\mathcal{S}_+$ and $\mathcal{S}_-$ are measurable. Hence, we can define the measurable function $f^* = \mathbf{1}_{\mathcal{S}_+} - \mathbf{1}_{\mathcal{S}_-}$ taking value $+1$ on $\mathcal{S}_+$ and -1 on $\mathcal{S}_-$. It follows that

$$\int_{\mathcal{S}} f^*(w,b)\eta(w,b) d\tau(w,b) = \int_{\mathcal{S}_+} \eta(w,b) d\tau(w,b) - \int_{\mathcal{S}_-} \eta(w,b) d\tau(w,b) = \int_{\mathcal{S}} |\eta(w,b)| d\tau(w,b),$$

which completes the proof.

3. *The claim is immediate from the previous part of this exercise.*

Exercise 2 (On spaces $\mathcal{F}_1$ and $\mathcal{F}_2$)**.** *In this exercise, we explore some basic differences between the spaces $\mathcal{F}_1$ and $\mathcal{F}_2$.*

1. *Explain briefly why $(\mathcal{F}_2, \|\cdot\|_{\mathcal{F}_2})$ is a Hilbert space.*

2. *Let $\mathcal{S} = [-1,1] \times \{0\}$ and let $\sigma(x) = \max(0,x)$ be the relu activation. Consider the functions $f_1 = \sigma(x)$ and $f_2 = -\sigma(-x)$. Prove that $\|f_1 + f_2\|_{\mathcal{F}_1} \geq 2$ and $\|f_1 - f_2\|_{\mathcal{F}_1} \geq 2$. Deduce that the space $(\mathcal{F}_1, \|\cdot\|_{\mathcal{F}_1})$ is not a Hilbert space.*

   Hint: *for $(\mathcal{F}_1, \|\cdot\|_{\mathcal{F}_1})$ to be a Hilbert space it is necessary that the parallelogram law holds for the norm $\|\cdot\|_{\mathcal{F}_1}$.*

3. *Prove that $\mathcal{F}_2 \subseteq \mathcal{F}_1$.*

4. *Let m be a finite natural number and for $j = 1, \ldots, m$ let $(w_j, b_j) \in \mathcal{S}$. Prove that the function f defined by a finite combinations of neurons $f(x) = \sum_{j=1}^m \eta_j \sigma(w_j^\top x + b_j)$ belongs to the space $\mathcal{F}_1$. Give an upper bound on its $\mathcal{F}_1$ norm.*

   Remark: *typically, even a single neuron – a function of the form $x \mapsto \sigma(w^\top x + b)$ – is not a member of the space $\mathcal{F}_2$. However, this fact is non-trivial to prove.*

Solution 2.

1. *As explained in Lecture 5, the definition of $\mathcal{F}_2$ (2) proves that $\mathcal{F}_2$ is in fact a RKHS with kernel $k(x,x') = \int_{\mathcal{S}} \sigma(w^\top x + b)\sigma(w^\top x' + b) d\tau(w,b)$. In particular, $\mathcal{F}_2$ is a Hilbert space.*

2. *Observe that $(f_1 + f_2)(x) = x$. Let μ be a signed measure so that for any $x \in \mathcal{X}_r$ we have*

$$(f_1 + f_2)(x) = \int_{-1}^{1} \sigma(wx) d\mu(w). \tag{4}$$

   For $x > 0$, the identity (4) yields

$$0 < x = (f_1 + f_2)(x) = x \int_{-1}^{1} \max(0,w) d\mu(w) = x \int_0^1 w d\mu(w).$$

In particular, the above implies that

$$\int_0^1 w d\mu(w) = 1. \tag{5}$$

By considering the case $x < 0$, we similarly have

$$0 < x = (f_1 + f_2)(x) = x \int_{-1}^1 \max(0, -w) d\mu(w) = x \int_{-1}^0 (-w) d\mu(w).$$

In particular, we have

$$\int_{-1}^0 (-w) d\mu(w) = 1. \tag{6}$$

Combining (5) and (6) yields

$$\|\mu\|_{\mathrm{TV}} = \sup_{|f| \le 1} \int_{-1}^1 f(w) d\mu(w) \ge \int_{-1}^0 (-w) d\mu(w) + \int_0^1 w d\mu(w) = 2.$$

Thus $\|f_1 + f_2\|_{\mathcal{F}_1} \ge 2$. Similarly, we can show that $\|f_1 - f_2\|_{\mathcal{F}_1} \ge 2$. It follows that

$$2\|f_1\|_{\mathcal{F}_1}^2 + 2\|f_2\|_{\mathcal{F}_1}^2 \le 4 < 8 \le \|f_1 + f_2\|_{\mathcal{F}_1}^2 + \|f_1 - f_2\|_{\mathcal{F}_1}^2.$$

Thus, the functions f_1, f_2 violate the parallelogram identity for the norm $\|\cdot\|_{\mathcal{F}_1}$, establishing that $(\mathcal{F}_1, \|\cdot\|_{\mathcal{F}_1})$ is not a Hilbert space.

3. *Suppose $f \in \mathcal{F}_2$ and suppose that $\|f\|_{\mathcal{F}_2} < a$ for some $a > 0$. Then, there exists some signed measure $d\mu(w, b) = \eta(w, b) d\tau(w, b)$ such that*

$$\int_{\mathbb{S}} \eta(w, b)^2 d\tau(w, b) < a \text{ and for any } x \in \mathcal{X}_r, f(x) = \int_{\mathbb{S}} \sigma(w^\top x + b) \eta_n(w, b) d\tau(w, b).$$

Because the measure μ is feasible for the optimization problem in (1), it follows by Jensen's inequality that

$$\|f\|_{\mathcal{F}_1} \le \int_{\mathbb{S}} |\eta(w, b)| d\tau(w, b) \le \sqrt{\int_{\mathbb{S}} \eta(w, b)^2 d\tau(w, b)} < a < \infty.$$

In particular, $f \in \mathcal{F}_1$ and $\|f\|_{\mathcal{F}_1} \le \|f\|_{\mathcal{F}_2}$.

4. *Take $\mu = \sum_{j=1}^m \eta_j \delta_{(w_j, b_j)}$ where $\delta_{(w_j, b_j)}$ is the Dirac distribution located at (w_j, b_j). Then $f(x) = \int_{\mathbb{S}} \sigma(x^\top w + b) d\mu(w, b)$ and $\|f\|_{\mathcal{F}_1} \le \|\mu\|_{\mathrm{TV}} = \sum_{j=1}^m |\eta_j|$.*