Topics in Math of ML, Lecture 7
Neural Networks (II)

Lénaïc Chizat*

November 3, 2025

1 Reminders from last week

We fix a radius $R > 0$ and define

$$\|f\|_{\mathcal{F}_1} \coloneqq \min_{\mu \in \mathcal{M}(S)} \left\{ \|\mu\|_{\mathrm{TV}} \ ; \ f(x) = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b) \mathrm{d}\mu(w,b), \forall x \in \mathbb{R}^d \text{ s.t. } \|x\| \leq R \right\} \quad (1)$$

where $\|\mu\|_{\mathrm{TV}}$ is the total variation norm of the measure μ, $S = \{(w,b) \in \mathbb{R}^d \times \mathbb{R} \ ; \ \|w\| = 1 \text{ and } b \in [-R,R]\}$ and for $\sigma(u) = (u)_+$ the ReLU activation.

Last week, we obtained for a function f that is differentiable on $[-R,R]$,

$$\|f\|_{\mathcal{F}_1} \leq \frac{|f(0)|}{R} + 2|f'(0)| + \int_{-R}^{R} |f''(b)| \mathrm{d}b \quad (2)$$

which corresponds to the decomposition of f in Eq. (1) with an explicit measure $\mu_f \in \mathcal{M}(\{-1,1\} \times [-R,R])$.

2 Approximation in higher dimension

Let us assume that $f : \mathbb{R}^d \to \mathbb{R}$ is continuous on the centered ball of radius R. Then we can write

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^\top x} \mathrm{d}\omega, \qquad \hat{f}(\omega) \coloneqq \int_{\mathbb{R}^d} f(x) e^{-i\omega^\top x} \mathrm{d}x. \quad (3)$$

Taking the real part in the integrand we get

$$f(x) = \frac{1}{(2\pi)^d} \left(\int_{\mathbb{R}^d} \Re(\hat{f}(\omega)) \cos(\omega^\top x) \mathrm{d}\omega - \int_{\mathbb{R}^d} \Im(\hat{f}(\omega)) \sin(\omega^\top x) \mathrm{d}\omega \right), \quad (4)$$

By the subadditivity property of norms, we have

$$\|f\|_{\mathcal{F}_1} \leq \frac{2}{(2\pi)^d} \int_{\mathbb{R}^d} |\hat{f}(\omega)| N(\omega) \mathrm{d}\omega, \qquad N(\omega) \coloneqq \max\{\|\cos(\omega^\top \cdot)\|_{\mathcal{F}_1}, \|\sin(\omega^\top \cdot)\|_{\mathcal{F}_1}\}. \quad (5)$$

(There are various ways to extend f outside of the ball of radius R, which lead to different $\hat{f}$: we can minimize over these extensions to get tighter upper bounds).

*EPFL lenaic.chizat@epfl.ch

We can explicitly build an (infinite width) neural network that represents $\cos(\omega^\top \cdot)$ by taking the definition of μ used in Eq. (2) and replacing $\{-1, 1\}$ by the points in the unit sphere $\{-\omega/\|\omega\|_2, \omega/\|\omega\|_2\}$. By the previous section (Eq. (2)), it thus follows

$$N(\omega) \leq \frac{1}{R} + 2\|\omega\|_2 + 2R\|\omega\|_2^2 \leq \frac{1}{R}(1 + 2R\|\omega\|_2 + 2R^2\|\omega\|_2^2) \leq \frac{2}{R}(1 + 2R^2\|\omega\|_2^2)$$

where we have used that $2ab \leq a^2 + b^2$. Thus we obtain

$$\|f\|_{\mathcal{F}_1} \leq \frac{1}{(2\pi)^d} \frac{4}{R} \int |\hat{f}(\omega)|(1 + 2R^2\|\omega\|_2^2)\mathrm{d}\omega, \tag{6}$$

which is a (non-standard) measure of smoothness of f.

Link with Sobolev regularity For $s > d/2$ we have

$$\begin{aligned}
\|f\|_{\mathcal{F}_1} &\leq \frac{1}{(2\pi)^d} \frac{4}{R} \int |\hat{f}(\omega)|(1 + 2R^2\|\omega\|_2^2)^{1+s/2}(1 + 2R^2\|\omega\|_2^2)^{-s/2}\mathrm{d}\omega \\
&\leq \frac{1}{(2\pi)^d} \frac{4}{R} \sqrt{\int |\hat{f}(\omega)|^2(1 + 2R^2\|\omega\|_2^2)^{2+s}\mathrm{d}\omega} \sqrt{\int (1 + 2R^2\|\omega\|_2^2)^{-s}\mathrm{d}\omega} \\
&\lesssim C\|f\|_{H^{2+s}}
\end{aligned}$$

where H^{2+s} is the Sobolev space of order $2 + s > 2 + d/2$ (the constraint $s > d/2$ is chosen so that the rightmost integral is finite). Thus the approximation properties of such Sobolev spaces apply, similarly to the case of kernel methods with translation invariant kernels. However, we will see next that $\mathcal{F}_1$ has in fact much richer than such RKHS.

3 Adaptivity to linear structures.

So far we did not observe any advantage of NNs over kernel methods. Let us now discuss a very powerful *adaptivity* to linear structure property which is specific to the $\mathcal{F}_1$ space and NNs.

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = g(V^\top x)$$

where $V \in \mathbb{R}^{d \times r}$ has orthonormal columns and $r \leq d$. In other words, f only depends on a smaller dimensional projection of the input. If $g \in \mathcal{F}_1(\mathbb{R}^r)$, it can be written

$$g(z) = \int_{\mathbb{R}^{r+1}} (w^\top z + b)_+ \mathrm{d}\mu(w, b),$$

with $\mathrm{d}\mu$ supported on $\{(w, b) \in \mathbb{R}^{r+1}, \|w\|_2 = 1, |b| \leq R\}$ and $\|g\|_{\mathcal{F}_1} = \|\mu\|_{\mathrm{TV}}$. We then have

$$f(x) = g(V^\top x) = \int_{\mathbb{R}^{r+1}} ((Vw)^\top x + b)_+ \mathrm{d}\mu(w, b) = \int_{\mathbb{R}^{d+1}} (\tilde{w}^\top x + b)_+ \mathrm{d}\tilde{\mu}(\tilde{w}, b)$$

where $\tilde{\mu} \in \mathcal{M}(\mathbb{R}^{d+1})$ is the pushforward measure[1] of μ with respect to the map $(w, b) \mapsto (Vw, b) \in \mathbb{S}_{d-1} \times [-R, R]]$ and has the same total variation norm as μ (because the map is

[1] Given a measurable map $T : \mathcal{X} \to \mathcal{Y}$ and a measure $\mu \in \mathcal{M}(\mathcal{X})$, the pushforward $T_\sharp \mu$ is the measure $\nu \in \mathcal{M}(\mathcal{Y})$ characterized by $\int \phi \mathrm{d}T_\sharp \mu = \int \phi \circ T \mathrm{d}\mu$ for all $\phi : \mathcal{Y} \to \mathbb{R}$ continuous.

injective, but in fact we only need $\|\tilde{\mu}\|_{TV} \leq \|\mu\|_{TV}$ which is a general property of pushforward measures). We have shown that

$$\|f\|_{\mathcal{F}_1(\mathbb{R}^d)} \leq \|g\|_{\mathcal{F}_1(\mathbb{R}^r)}.$$

Thus in presence of this low-dimensional structure, we can get a bound of the form $\|f\|_{\mathcal{F}_1(\mathbb{R}^d)} \leq \|g\|_{H^{(r+5)/2}}$, which is much better than $\|g\|_{H^{(d+5)/2}}$ (in absence of such a structure), in particular if $r \ll d$.

In contrast, kernel methods do not benefit from this additional "adaptivity" to low dimensional structure and would pay the price of the dimension d. More precise results below.

4 Summing-up the story (without optimization)

Let us finally study bounds on the excess risk, with the following NN estimator

$$\hat{f}_D \in \arg\min_f \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) \text{ subject to } \|f\|_{\mathcal{F}_1} \leq D. \tag{7}$$

Let $\mathcal{F}_D = \{f \in \mathcal{F}_1 \; ; \; \|f\|_{\mathcal{F}_1} \leq D\}$. In the exercises, you will show that there exists $C_R > 0$ such that the Rademacher complexity of $\mathcal{F}_D$ is bounded as

$$\mathrm{Rad}_n(\mathcal{F}_D) \leq \frac{C_R D}{\sqrt{n}}.$$

This bound is similar to the one we have for balls in a RKHS (proved in Lecture 3). Assuming ρ has a bounded density, $\ell(y,\cdot)$ is Lipschitz continuous (uniformly in y) and for the ideal choice of D, we have using our analysis from Lecture 5 (for some $C, C' > 0$ independent of n):

$$\inf_{D\geq 0} \mathbf{E}[\mathcal{R}(\hat{f}_D)] - \mathcal{R}(f^*) \lesssim \sqrt{\tilde{A}(C/n, f^*)}$$

where,

$$\tilde{A}(\lambda, f^*) := \inf_f \{\|f - f^*\|^2_{L^2(\mathrm{d}x)} + \lambda \|f\|^2_{\mathcal{F}_1}\} \leq C' \|f^*\|^2_{H^t} \lambda^{2t/(d+5)} \tag{8}$$

if $0 \leq t \leq (d+5)/2$ (the last inequality can be proved as an exercise, along the same lines as in Lecture 5). It follows:

- if $f^* \in H^t$, that is (roughly) t bounded derivatives, then the error decays as $O(n^{-t/(d+5)})$ (this suffers from the curse of dimensionality, as when learning with a translation invariant kernel, e.g. for the exponential kernel we get $O(n^{-t/(d+1)})$);
- if $f^*(x) = g(Px)$ where P is an orthonormal projection of rank $r \leq d$ and $g \in H^t$, then in Eq. (8) we can replace d by r (following the reasoning of the previous section). Then the excess risk decays as $O(n^{-t/(r+5)})$. This means that NNs are able to break the curse of dimensionality.[2].

This section is well complemented by the reading of the following blog post https://francisbach.com/quest-for-adaptivity/, that sums up part of what we have seen in the class until now (i.e. the pre-optimization story).

[2]This is a bit over-optimistic because in fact the ERM problem with $\mathcal{F}_1$-norm regularization is much harder to optimize in general than the one with $\mathcal{F}_2$-norm (which is efficiently approximated with random features, and tractable with convex optimization).

5 Approximation with a finite number of neurons

We consider an infinite width 2NN

$$f(x) = \int_S \sigma(w^\top x) \mathrm{d}\mu(w)$$

(NB: now the intercept b does not play a special role, so we set $x \leftarrow [x; 1]$ and $w \leftarrow [w; b]$ and one may take S as the unit sphere in $\mathbb{R}^{d+1}$). We ask wether it can be well approximated by a finite width 2NN

$$\hat{f}_m(x) = \frac{1}{m} \sum_{i=1}^m \eta_j \sigma(w_j^\top x) = \int \sigma(w^\top x) \mathrm{d}\hat{\mu}_m(w)$$

with $\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^m \eta_j \delta_{w_j}$. For simplicity, we consider approximation in L^2 norm for some given data distribution $\rho \in \mathcal{P}(\mathbb{R}^{d+1})$.

Monte-Carlo sampling Let $\mu = \mu_+ - \mu_-$ be the Jordan decomposition of μ (as the difference of two non-negative measures), which is such that $\|\mu\| = \|\mu_+\| + \|\mu_-\|$ and we assume that these two terms are positive (the other case being in fact easier).

Let (s, w) be a couple of random variables where $s \in \{-1, +1\}$ is such that $\mathbf{P}(s = +1) = \|\mu_+\|/\|\mu\|$ and $w|s = +1 \sim \mu_+/\|\mu_+\|$ and $w|s = -1 \sim \mu_-/\|\mu_-\|$. Remark that the law of w is, by construction, $|\mu|/\|\mu\|_{TV}$ where $|\mu| = \mu_+ + \mu_- \in \mathcal{M}_+(S)$ is the *variation* of μ.

We sample $(w_1, s_1), \dots, (w_m, s_m)$ independently from this random variable. Letting $\eta_j = s_j \|\mu\|$, this defines an empirical measure $\hat{\mu} = \frac{1}{m} \sum_{j=1}^m \eta_j \delta_{w_j}$ and a corresponding predictor $\hat{f}_m$. This method enjoys the usual Monte-Carlo rate.

Proposition 5.1. *It holds*

$$\mathbf{E}\|f - \hat{f}_m\|^2_{L^2(\rho)} \leq \frac{\|\mu\|_{TV} \int \|x \mapsto \sigma(w^\top x)\|^2_{L^2(\rho)} \mathrm{d}|\mu|(w)}{m} \leq C \frac{\|\mu\|^2_{TV}}{m}. \tag{9}$$

with $C = \sup_{w \in S} \|x \mapsto \sigma(w^\top x)\|^2_{L^2(\rho)}$. *In particular, for any* $f \in \mathcal{F}_1$, *there exists a 2NN of width* m *such that* $\|f - \hat{f}_m\|^2_{L_2(\rho)} \leq C\|f\|^2_{\mathcal{F}_1}/m$.

Proof. Let $\Phi : \mathbb{R}^{d+1} \to L^2(\rho)$ be defined by $\Phi(w) = (x \mapsto \sigma(w^\top x))$ and define the random variable $Z = s\|\mu\|\Phi(w)$, which expectation is, by construction, $\mathbf{E}Z = \int \Phi(w) \mathrm{d}\mu(w) = f$. Now Eq. (9) is precisely a bound on the (centered) variance of $\bar{Z} = \frac{1}{m} \sum_i Z_i$ for independent samples of Z, that is

$$\mathbf{E}\|\hat{f} - f\|^2_{L^2(\rho)} = \mathbf{E}\Big\|\frac{1}{m} \sum_{j=1}^m Z_j - \mathbf{E}Z\Big\|^2 = \frac{1}{m} \mathbf{E}\,\|Z - \mathbf{E}Z\|^2 \leq \frac{1}{m} \mathbf{E}\|Z\|^2 \tag{10}$$

which is the first bound in Eq. (9). Indeed, since $\mathrm{Law}(w) = |\mu|/\|\mu\|_{TV}$,

$$\mathbf{E}\|Z\|^2 = \mathbf{E}\|s\|\mu\|_{TV}\Phi(w)\|^2 = \|\mu\|^2_{TV} \int \|\Phi(w)\|^2 \mathrm{d}(|\mu|(w)/\|\mu\|_{TV}) = \|\mu\|_{TV} \int \|\Phi(w)\|^2 \mathrm{d}|\mu|(w).$$

Note that by the central limit theorem, using independent samples, it holds

$$\sqrt{m} \left(\frac{1}{m} \sum_{j=1}^m Z_j - \mathbf{E}[Z] \right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\Sigma = \mathrm{Cov}(Z - \mathbf{E}[Z])$ so, as usual for Monte-Carlo methods, this bound has a tight rate (unless Φ is constant μ-a.e.). □

- **Tractability via over-parameterization.** This result could mislead us to think that it is sufficient to train a NN of width $O(\|f^*\|_{\mathcal{F}_1}/\epsilon^2)$ to get a final $L^2(\rho)$ error of ϵ. This would be forgetting about the problem of optimization: in practice μ is unknown and cannot be sampled from (the goal of optimization is precisely to find μ). In fact, recent theory suggests that over-parameterization (using more neurons than approximation theory would suggest) helps optimization.

- **Pruning.** The previous method can be used in practice if one wants to reduce the size of a trained NN (e.g., to upload it on a small device such as an embedded system or a phone). The problem of reducing the size of a trained NN is called *pruning* and many other methods have been proposed.