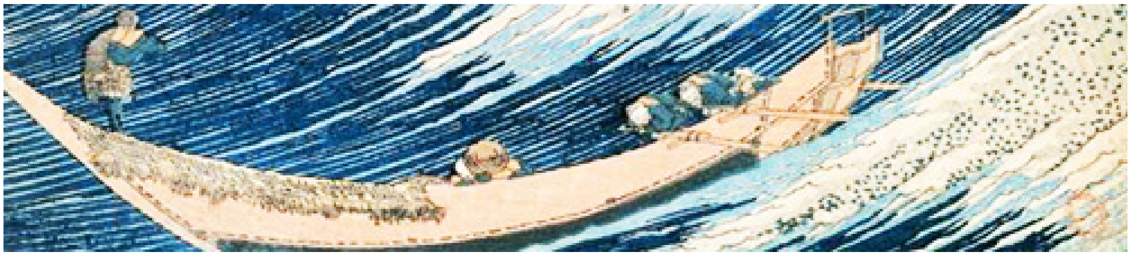
Mathematical Foundations of Data Sciences



Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

April 25, 2021

Chapter 1

Shallow Learning

In this chapter, we study the simplest example of non-linear parametric models, namely Multi-Layers Perceptron (MLP) with a single hidden layer (so they have in total 2 layers). Perceptron (with no hidden layer) corresponds to the linear models studied in the previous chapter. MLP with more layers are obtained by stacking together several such simple MLP, and are studied in Section **??**, since the computation of their derivatives is very suited to automatic-differentiation methods.

1.1 Recap on Supervised Learning

In this chapter, we denote $(x_i, y_i)_{i=1}^n$ the training data, with $x_i \in \mathcal{X}$ (typically $\mathcal{X} = \mathbb{R}^d$ in the following). Let (X, Y) be a pair of random variables with values in $\mathcal{X} \times \mathcal{Y}$. We observe $(x_i, y_i)_{i=1\dots n}$ an i.i.d. of the same law as (X, Y). For a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we write $R(g) \triangleq \mathbb{E}(\ell(g(X), Y))$ the risk of a function g. The goal is to approach $f \in \operatorname{argmin} R(f)$ as best as possible from only the knowledge of $(x_i, y_i)_{i=1,\dots,n}$. Given a space $\mathcal{F}$ of functions, this is achieved by minimizing the empirical risk

$$\hat{f} \triangleq \underset{g \in \mathcal{F}}{\operatorname{argmin}} \hat{R}(g) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i).$$

The goal of the theoretical analysis of supervised learning is to limit the excess risk of this estimator. To do this, we break down this excess risk into two terms:

$$R(\hat{f}) - R(f) = [R(\hat{f}) - \min_{g \in \mathcal{F}} R(g)] + [\min_{g \in \mathcal{F}} R(g) - R(f)] \tag{1.1}$$

The first term E_{est} quantifies the distance between the model $\hat{f}$ and the optimal model on $\mathcal{F}$, it is called the estimation error. This is a random error that can be controlled with tools such as Rademacher Complexity. Intuitively this requires that the set $\mathcal{F}$ is not too large.

The second term E_{app} is called the approximation error, it quantifies the quality of the optimal model on $\mathcal{F}$. If we suppose that ℓ is C_ℓ -Lipschitz with respect to its first variable(which is the case for classical cost functions), then we can limit

$$E_{\text{app}} \leqslant C_\ell \min_{g \in \mathcal{F}} \|g - f\|_{L^2(\mu)}$$

where we denote by μ the probability distribution of X and $\|h\|_{L^2(\mu)}^2 \triangleq \int_{\mathcal{X}} h(x)^2 d\mu(x)$. In part of this chapter, we focus on the study of this error. For it to be small, $\mathcal{F}$ must be large enough, but it will also be necessary to make assumptions about f.

1.2 Multi-layer Perceptron

We consider the case $\mathcal{X} = \mathbb{R}^d$. A MLP with a single hidden layer having q neurons defines a function as

$$f_q(x) \triangleq a_0 + \sum_{k=1}^{q} a_k \rho(\langle x,\, b_k \rangle + c_k), \tag{1.2}$$

where $a_k \in \mathbb{R}$ are outer neurons weights, and $w_k = (b_k, c_k) \in \mathbb{R}^d \times \mathbb{R}$ are the neurons weights which compose the hidden layers. To ease the notation, we denote $\bar{x} = (x, 1) \in \mathbb{R}^{d+1}$ and $w_k = (b_k, c_k) \in \mathbb{R}^{d+1}$, so that $\langle \bar{x},\, w_k \rangle = \langle x,\, b_k \rangle + c_k$. With this notation, we conveniently write

$$f_q(x; (W, a)) \triangleq a_0 + \sum_{k=1}^{q} a_k \rho(\langle \bar{x},\, w_k \rangle) = a_0 + \langle \rho(W\bar{x}),\, a \rangle_{\mathbb{R}^q},$$

where the neurons w_k as the rows of $W \in \mathbb{R}^{q \times (d+1)}$ and the non-linearity is implicitly assumed to be applied component-wise. In the following, we often denote $f_q(x)$ in place of $f_q(x; (W, a))$.

This function f_q is thus a weighted sum of q "ridge functions" $\rho(\langle \cdot,\, w_k \rangle)$. These functions are constant in the direction orthogonal to the neuron w_k and have a profile defined by ρ.

The most popular non-linearities are sigmoid functions such as

$$\rho(r) = \frac{e^r}{1 + e^r} \quad \text{and} \quad \rho(r) = \frac{1}{\pi}\text{atan}(r) + \frac{1}{2}$$

and the rectified linear unit (ReLu) function $\rho(r) = \max(r, 0)$.

Expressiveness. In order to define function of arbitrary complexity when q increases, it is important that ρ is non-linear. Indeed, if $\rho(s) = s$, then $f_q(x) = \langle Wx,\, a \rangle = \langle x,\, W^\top a \rangle$. It is thus a linear function with weights $W^\top a$, whatever the number q of neurons. Similarly, if ρ is a polynomial on $\mathbb{R}$ of degree d, then f_q is itself a polynomial of degree d in $\mathbb{R}^p$, which is a linear space V of finite dimension $\dim(V) = O(p^d)$. So even if q increases, the dimension $\dim(V)$ stays fixed and f_q cannot approximate an arbitrary function outside V. In sharp contrast, one can show that if ρ is not polynomial, then f_q can approximate any continuous function, as studied in Section 1.5.

1.3 Training a MLP

Given pairs of features and data values $(x_i, y_i)_{i=1}^n$, and as usual storing the features $(x_k, 1)$ in the rows of $X \in \mathbb{R}^{n \times (d+1)}$, we consider the following least square regression function (similar computation can be done for classification losses) which corresponds to a minimization of the empirical risk

$$\min_{W,a} \mathcal{E}(W, a) \overset{\text{def.}}{=} \frac{1}{2} \sum_{i=1}^{n} (f_q(x_i; (W, a)) - y_i)^2 = \frac{1}{2} \|\rho(XW^\top)a - y\|^2.$$

Note that here, the parameters being optimized are $(W, a) \in \mathbb{R}^{q \times (d+1)} \times \mathbb{R}^q$.

Optimizing with respect to a. This function $\mathcal{E}$ is convex with respect to a, since it is a quadratic function. Its gradient with respect to a can be computed as in (**??**) and thus (here we do not consider the derivative with respect to a_0, which is treated after separately)

$$\nabla_a \mathcal{E}(W, a) = \rho(XW^\top)^\top (a_0 + \rho(XW^\top)a - y) \in \mathbb{R}^q$$

and one can compute in closed form the solution (assuming $\ker(\rho(XW^\top)) = \{0\}$) as

$$a^\star = [\rho(XW^\top)^\top \rho(XW^\top)]^{-1} \rho(XW^\top)^\top (y - a_0) = [\rho(WX^\top)\rho(XW^\top)]^{-1} \rho(WX^\top)(y - a_0)$$

When $W = \mathrm{Id}_p$ and $\rho(s) = s$ one recovers the least square formula (**??**). The derivative with respect to the output bias a_0 is

$$\nabla_{a_0} \mathcal{E}(W, a) = \langle \mathbb{1}_n,\, \rho(XW^\top)a - y \rangle \in \mathbb{R}$$

Optimizing with respect to W. The function $\mathcal{E}$ is non-convex with respect to W because the function ρ is itself non-linear. Training a MLP is thus a delicate process, and one can only hope to obtain a local minimum of $\mathcal{E}$. It is also important to initialize correctly the neurons $(w_k)_k$ (for instance as unit norm random vector, but bias terms might need some adjustment), while u can be usually initialized at 0.

To compute its gradient with respect to W, we first note that for a perturbation $\varepsilon \in \mathbb{R}^{q\times(d+1)}$, one has

$$\rho(X(W+\varepsilon)^\top) = \rho(XW^\top + X\varepsilon^\top) = \rho(XW^\top) + \rho'(XW^\top) \odot (X\varepsilon^\top)$$

where we have denoted "$\odot$" the entry-wise multiplication of matrices, i.e. $U \odot V = (U_{i,j}V_{i,j})_{i,j}$. One thus has,

$$\begin{aligned}\mathcal{E}(W+\varepsilon, a) &= \frac{1}{2}\|e + [\rho'(XW^\top) \odot (A\varepsilon^\top)]y\|^2 \quad \text{where} \quad e \overset{\text{def.}}{=} \rho(XW^\top)a - y \in \mathbb{R}^n \\ &= \mathcal{E}(W,a) + \langle e,\, [\rho'(XW^\top) \odot (A\varepsilon^\top)]y\rangle + o(\|\varepsilon\|) \\ &= \mathcal{E}(W,a) + \langle A\varepsilon^\top,\, \rho'(XW^\top) \odot (eu^\top)\rangle \\ &= \mathcal{E}(W,a) + \langle \varepsilon^\top,\, A^\top \times [\rho'(XW^\top) \odot (ea^\top)]\rangle.\end{aligned}$$

The gradient thus reads

$$\nabla_W \mathcal{E}(W,a) = [\rho'(WX^\top) \odot (ae^\top)] \times A \in \mathbb{R}^{q\times(d+1)}.$$

1.4 Controlling the Estimation Error

Regarding the error E_{est}, defined in (1.1), it is well studied in the literature. Under the assumptions that the weights of the network are bounded, this error is increased with great probability, by a term of the order of $\frac{1}{\sqrt{n}}$. This corresponds to a decay rate independent of the number p of neurons.

Denoting $\sigma^2 \triangleq \int \|x\|^2 \mathrm{d}\mu(x)$ the variance of the data, if we impose that φ is G_φ-Lipschitz, that

$$\sum_k |a_k| \leqslant K_a, \quad \|b_k\|_2, |c_k|/\sigma \leqslant K_{b,c}, \tag{1.3}$$

and that the loss ℓ is G_ℓ-Lipschitz, then as explained in [2, Proposition 9.1], with high probability,

$$E_{\text{est}} \leqslant \frac{2G_\ell G_\varphi K_a K_{b,c}\sigma}{\sqrt{n}} \tag{1.4}$$

which is a speed independent of the number q of neurons. We now study how the error E_{app} decreases as a function of q.

1.5 Universality

In the following we assume that $\rho : \mathbb{R} \to \mathbb{R}$ is a bounded measurable function such that

$$\rho(r) \overset{r\to-\infty}{\longrightarrow} 0 \quad \text{and} \quad \rho(r) \overset{r\to+\infty}{\longrightarrow} 1. \tag{1.5}$$

Note in particular that such a function cannot be a polynomial and that the ReLu function does not satisfy these hypothesis (universality for the ReLu is more involved to show). We denote a ridge function as

$$\varphi_w(x) \triangleq \rho(\langle w,\, \bar{x}\rangle) \quad \text{where} \quad w = (b,c) \in \mathbb{R}^d \times \mathbb{R}. \tag{1.6}$$

The goal is to show the following theorem.

Theorem 1 (Cybenko [4])**.** *For any compact set $\Omega \subset \mathbb{R}^d$, the space spanned by the functions $\{\varphi_w\}_w$ is dense in $\mathcal{C}(\Omega)$ for the uniform convergence. This means that for any continuous function f and any $\varepsilon > 0$, there exists $q \in \mathbb{N}$ and weights $(w_k, a_k)_{k=1}^q$ such that*

$$\forall\, x \in \Omega, \quad |f(x) - f_q(x,(W,a))| \leqslant \varepsilon.$$

In a typical ML scenario, this implies that one can "overfit" the data, since using a q large enough ensures that the training error can be made arbitrary small. Of course, there is a bias-variance tradeoff, and q needs to be cross-validated to account for the finite number n of data, and ensure a good generalization properties.

Proof in dimension $p = 1$. In 1D, the approximation f_q can be thought as an approximation using smoothed step functions. Indeed, introducing a parameter $\varepsilon > 0$, one has (assuming the function is Lipschitz to ensure uniform convergence),

$$\varphi_{w/\varepsilon} \overset{\varepsilon \to 0}{\longrightarrow} 1_{[-c/b,+\infty[}$$

This means that

$$f(\cdot;(W/\varepsilon,a)) \overset{\varepsilon \to 0}{\longrightarrow} \sum_k a_k 1_{[-c_k/b_k,+\infty[},$$

which is a piecewise constant function. Inversely, any piecewise constant function can be written this way. Indeed, if f assumes the value d_k on each interval $[t_k, t_{k+1}[$, then it can be written as

$$h = \sum_k d_k (1_{[t_k,+\infty[} - 1_{[t_k,+\infty[}).$$

Since the space of piecewise constant functions is dense in continuous function over an interval, this proves the theorem.

Proof in arbitrary dimension d. We start by proving the following dual characterization of density, using bounded Borel measure $\nu \in \mathcal{M}(\Omega)$ i.e. such that $\nu(\Omega) < +\infty$.

Proposition 1. *If ρ is such that for any Borel measure $\nu \in \mathcal{M}(\Omega)$*

$$\left(\forall\, w, \int \rho(\langle \bar{x},\, w \rangle) \mathrm{d}\nu(x) = 0 \right) \quad \Longrightarrow \quad \nu = 0, \tag{1.7}$$

then Theorem 1 holds.

Proof. We consider the linear space

$$\mathcal{S} \overset{\text{def.}}{=} \left\{ \sum_{k=1}^q a_k \varphi_{w_k} \; ; \; q \in \mathbb{N}, w_k \in \mathbb{R}^{d+1}, a_k \in \mathbb{R} \right\} \subset \mathcal{C}(\Omega).$$

Let $\bar{\mathcal{S}}$ be its closure in $\mathcal{C}(\Omega)$ for $\|\cdot\|_\infty$, which is a Banach space. If $\bar{\mathcal{S}} \neq \mathcal{C}(\Omega)$, let us pick $g \neq 0$, $g \in \mathcal{C}(\Omega) \backslash \bar{\mathcal{S}}$. We define the linear form L on $\bar{\mathcal{S}} \oplus \mathrm{span}(g)$ as

$$\forall\, s \in \bar{\mathcal{S}},\, \forall\, \lambda \in \mathbb{R}, \qquad L(s + \lambda g) = \lambda$$

so that $L = 0$ on $\bar{\mathcal{S}}$. L is a bounded linear form, so that by Hahn-Banach theorem, it can be extended in a bounded linear form $\bar{L} : \mathcal{C}(\Omega) \to \mathbb{R}$. Since $L \in \mathcal{C}(\Omega)^*$ (the dual space of continuous linear form), and that this dual space is identified with Borel measures, there exists $\nu \in \mathcal{M}(\Omega)$, with $\nu \neq 0$, such that for any continuous function h, $\bar{L}(h) = \int_\Omega h(x) \mathrm{d}\nu(x)$. But since $\bar{L} = 0$ on $\bar{\mathcal{S}}$, $\int \rho(\langle \bar{\cdot},\, w \rangle) \mathrm{d}\nu = 0$ for all w and thus by hypothesis, $\nu = 0$, which is a contradiction. □

The theorem now follows from the following proposition.

Proposition 2. *If ρ is continuous and satisfies* (1.5)*, then it satisfies* (1.7).

Proof. One has, for $w = (b, c) \in \mathbb{R}^{d+1}$ and $t \in \mathbb{R}$,

$$\varphi_{\frac{b}{\varepsilon}, \frac{c}{\varepsilon}+t}(x) = \rho\left(\frac{\langle \bar{x},\, w\rangle}{\varepsilon} + t\right) \overset{\varepsilon\to 0}{\longrightarrow} \gamma(x) \overset{\text{def.}}{=} \begin{cases} 1 & \text{if} \quad x \in H_w, \\ \rho(t) & \text{if} \quad x \in P_w, \\ 0 & \text{if} \quad \langle \bar{x},\, w\rangle < 0, \end{cases}$$

where we defined $H_w \overset{\text{def.}}{=} \{x\ ;\ \langle \bar{x},\, w\rangle > 0\}$ and $P_w \overset{\text{def.}}{=} \{x\ ;\ \langle w,\, \bar{x}\rangle = 0\}$. By Lebesgue dominated convergence (since the involved quantities are bounded uniformly on a compact set)

$$\int \varphi_{\frac{b}{\varepsilon}, \frac{c}{\varepsilon}+t} \mathrm{d}\nu \overset{\varepsilon\to 0}{\longrightarrow} \int \gamma \mathrm{d}\nu = \varphi(t)\nu(P_w) + \nu(H_w).$$

Thus if ν is such that all these integrals vanish, then

$$\forall\, (w, t), \quad \varphi(t)\nu(P_w) + \nu(H_w) = 0.$$

By selecting (t, t') such that $\varphi(t) \neq \varphi(t')$, one has that

$$\forall\, w, \quad \nu(P_w) = \nu(H_w) = 0.$$

We now need to show that $\nu = 0$. For a fixed $b \in \mathbb{R}^d$, we consider the function

$$h \in L^\infty(\mathbb{R}), \quad F(h) \overset{\text{def.}}{=} \int_\Omega h(\langle b,\, x\rangle)\mathrm{d}\nu(x).$$

$F : L^\infty(\mathbb{R}) \to \mathbb{R}$ is a bounded linear form since $|F(\nu)| \leqslant \|h\|_\infty \nu(\Omega)$ and $\nu(\Omega) < +\infty$. One has, for $w = (b, c) \in \mathbb{R}^{d+1}$,

$$F(1_{[-c,+\infty[} = \int_\Omega 1_{[-c,+\infty[}(\langle b,\, x\rangle)\mathrm{d}\nu(x) = \nu(P_w) + \nu(H_w) = 0.$$

By linearity, $F(h) = 0$ for all piecewise constant functions, and F is a continuous linear form, so that by density $F(h) = 0$ for all functions $h \in L^\infty(\mathbb{R})$. Applying this for $h(r) = e^{\mathrm{i}r}$ one obtains

$$\hat{\nu}(b) \overset{\text{def.}}{=} \int_\Omega e^{\mathrm{i}\langle x,\, b\rangle}\mathrm{d}\nu(x) = 0.$$

This means that the Fourier transform of ν is zero, so that $\nu = 0$. □

1.6 Approximation Rates

This section details in a mostly informal way the results of [3].

1.6.1 Barron's spaces

For an integrable function f, its Fourier transform is defined, for any $\omega \in \mathbb{R}^d$ by

$$\hat{f}(\omega) \triangleq \int_{\mathbb{R}^d} f(\omega)e^{\mathrm{i}\langle \omega,\, x\rangle}\mathrm{d}x.$$

The Barron's space [3] is the set of functions such as the semi-norm

$$\|f\|_B \triangleq \|b\| \int_{\mathbb{R}^d} \|\omega\||\hat{f}(\omega)|\mathrm{d}\omega$$

is finite. If we impose that $f(0)$ is fixed, we can show that this defines a norm and that the Barron space is a Banach space. One has

$$\|f\|_B = \int_{\mathbb{R}^d} \|\widehat{\nabla f}(\omega)\| \mathrm{d}\omega,$$

this shows that the functions of the Barron space are quite regular. Here are some example of function classes with the corresponding Barron's norm.

- *Gaussians*: for $f(x) = e^{-\|x\|^2/2}$, one has $\|f\|_B \leqslant 2\sqrt{d}$
- *Ridge function*: let $f(x) = \psi(\langle x,\, b\rangle + c)$ where $\psi : \mathbb{R} \to \mathbb{R}$ then one has

  $$\|f\|_B \leqslant \int_{\mathbb{R}} |u\hat{\psi}(u)| \mathrm{d}u.$$

  In particular, if ψ is $\mathcal{C}^{2+\delta}$ for $\delta > 0$ then f is in the Barron space. If ρ satisfies this hypothesis, the "neurons" functions are in Barron space.
- *Regular functions with s derivatives*: for all $s > d/2$, one has $\|f\|_B \leqslant C(d,s)\, \|f\|_{H^s}$ where the Sobolev norm is

  $$\|f\|_{H^s}^2 \triangleq \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 (1 + \|\omega\|^{2s}) \mathrm{d}\omega \sim \|f\|_{L^2(\mathrm{d}x)}^2 + \sum_{k=1}^{d} \|\partial_{x_k} f\|_{L^2(\mathrm{d}x)}^2,$$

  and $C(d,s) < \infty$ is a constant. This shows that if f has at least $d/2$ derivatives in L^2, it is in Barron space. Beware that the converse is false, the Barron space can contain less regular functions as seen in the previous examples. This somehow shows that the Barron space is larger than RKHS space of fixed smoothness degree.

1.6.2 Barron's Theorem

The main result is as follows.

Theorem 2 (Barron [3]). *We assume μ is supported on $B(0,R)$. For all q, there exists an MLP f_q of the form (1.2) such that*

$$\|f_q - f\|_{L^2(\mu)} \leqslant \frac{2R\|f\|_B}{\sqrt{q}}.$$

This result shows that if f is in Barron space, the decrease of the error does not depend on the dimension: this is often referred to as "overcoming the curse of dimensionality". Be careful however, the constant $\|f\|_B$ can depend on the dimension, this is the case for Gaussian functions (where it is $2\sqrt{d}$) but not for ridges functions.

The proof of this theorem shows that one can impose $c_0 = f(0)$ and $\sum_k |a_k| \leqslant \alpha_{\max} \triangleq 2R\|f\|_B$. This ℓ^1 normalization is crucial since it is the one required in (**??**) to control the estimation error. Controlling this error also requires to bound $\|w_k\|$, which can be achieved assuming more constraint on ρ. The estimation error is thus bounds as $O(1/\sqrt{n})$ while the approximation one is $O(1/\sqrt{q})$, so that a balance is obtained by selecting $q \sim n$, and the overall error has a dimension-free decay rate $0(1/\sqrt{n})$.

We will give the main ingredients of the proof of Barron's theorem.

Integral representation using a probability distribution The most important point is a formula for representing functions in the Barron space as an average of ridge functions, this average being computed using a probability distribution Λ over the space of neurons. For $\theta = (w, \alpha) \in \mathbb{R}^{d+1} \times \mathbb{R}$, we denote

$$\xi_\theta(x) \overset{\text{def.}}{=} \alpha\rho(\langle w,\, \bar{x}\rangle).$$

We can interpret the MLP as a finite average

$$f_q = a_0 + \frac{1}{q}\sum_{k=1}^{p} \xi_{\theta_k} \tag{1.8}$$

where $\theta_k \triangleq (w_k, p\, a_k)$. The central idea of Barron's proof is that the functions f of the Barron space are those where we replace this finite empirical mean by a "continuous average". That is, for any f in the Barron space, there exists a probability density Γ on $\mathbb{R}^{d+1} \times \mathbb{R}$ such that

$$f(x) = f(0) + \int_{\mathbb{R}^{d+1}\times\mathbb{R}} \xi_\theta \mathrm{d}\Gamma(\theta). \tag{1.9}$$

Barron shows in the proof of this result that we can impose that Γ is supported in $\mathbb{R}^{d+1} \times B(0, \alpha_{\max})$ where $\alpha_{\max} \triangleq 2R\|f\|_B$. For example, we get the (1.8) decomposition by taking $\Gamma = 1/p\sum_k \delta_{\theta_k}$ and $c_0 = f(0)$.

The proof of this decomposition is the crux of the proof. Using the inverse Fourier transform and the fact that $f(x)$ is real, one has

$$\begin{aligned} f(x) - f(0) &= \Re\left(\int_{\mathbb{R}^d} \hat{f}(\omega)(e^{\mathrm{i}\langle\omega,\, x\rangle} - 1)\mathrm{d}\omega\right) = \Re\left(\int_{\mathbb{R}^d} |\hat{f}(\omega)|e^{\mathrm{i}\Theta(\omega)}(e^{\mathrm{i}\langle\omega,\, x\rangle} - 1)\mathrm{d}\omega\right) \\ &= \int_{\mathbb{R}^d} (\cos(\langle\omega,\, x\rangle + \Theta(\omega)) - \cos(\Theta(\omega)))|\hat{f}(\omega)|\mathrm{d}\omega \\ &= \int_{\mathbb{R}^d} \frac{\|f\|_B}{|\omega|}(\cos(\langle\omega,\, x\rangle + \Theta(\omega)) - \cos(\Theta(\omega)))\frac{\|\omega\||\hat{f}(\omega)|}{\|f\|_B}\mathrm{d}\omega = \int_{\mathbb{R}^d} g_\omega(x)\mathrm{d}\Gamma(\omega) \end{aligned}$$

$$\text{where} \quad g_\omega(x) \triangleq \frac{\|f\|_B}{\|\omega\|}(\cos(\langle\omega,\, x\rangle + \Theta(\omega)) - \cos(\Theta(\omega))) \quad \text{and} \quad \mathrm{d}\Lambda(\omega) \triangleq \frac{\|\omega\||\hat{f}(\omega)|}{\|f\|_B}\mathrm{d}\omega$$

Note that $\|g_\omega\|_\infty \leqslant 1$. This calculation shows that the previous decomposition (1.11) is true but with sigmoid functions g_ω instead of functions ξ_θ. One then proceeds by showing that the function cos can be written using translates and dilates of the function ρ to obtain thought after integral formula.

Discretization To go from the integral representation (1.11) to f_q, which is a discretized version (1.8), Barron performs a probabilistic proof by Monte Carlo sampling. We consider i.i.d. realizations $\Theta \triangleq (\theta_1, \dots \theta_q)$ according to the law Γ and we define f_q according to the definition (1.8). Beware that f_q is now a random function. By expanding the squares and inner product, and using the fact that $\mathbb{E}_{\theta_k}(\varphi_{\theta_k}(x)) = f(x)$, one has

$$\mathbb{E}_\Theta(\|f - f_q\|^2_{L^2(\mu)}) = \frac{\mathbb{E}_\theta(\|\xi_\theta - \mathbb{E}_{\theta'}(\xi_{\theta'})\|^2_{L^2(\mu)})}{q} = \frac{\mathbb{E}_\theta(\|\xi_\theta\|^2_{L^2(\mu)}) - \|\mathbb{E}_\theta(\xi_\theta)\|^2_{L^2(\mu)}}{q}. \tag{1.10}$$

An important point is that one can bound

$$\mathbb{E}(\|\xi_\theta\|^2_{L^2(\mu)}) \leqslant \alpha^2_{\max} \quad \text{where} \quad \alpha_{\max} \triangleq 2\|f\|_B R$$

because the realizations (w, α) from Γ satisfy $|\alpha| \leqslant \alpha_{\max}$ because of the support localization of Γ. This shows that $\mathbb{E}_\Theta(\|f - f_q\|^2_{L^2(\mu)}) \leqslant 2\|f\|_B R/q$. This shows that the event $\|f - f_q\|^2_{L^2(\mu)} \leqslant 2\|f\|_B R/q$ has a positive probability so that the set of MLP such that this bound holds is non-empty.

Furthermore, since $|\alpha| \leqslant \alpha_{\max}$ on the support of Γ, the $(a_k)_k$ in the decomposition of the MLP, which are of of the form $a_k = \alpha_k/q$ with α_k in the support de Γ, safisfy $\sum_k |a_k| = 1/q\sum_k |\alpha_k| \leqslant \alpha_{\max}$.

1.6.3 Integral representation over w

Integral representation using a signed measure. It is possible to develop an alternate proof, which relies on a simpler and more natural representation of Barron functions using a signed measure. Indeed, one can re-write (1.11) as (using a slight abuse of notation)

$$f(x) = f(0) + \int_{\mathbb{R}^{d+1}\times\mathbb{R}} \alpha\rho(\langle w,\, \bar{x}\rangle \mathrm{d}\Gamma(w,\alpha) = f(0) + \int_{\mathbb{R}^{d+1}} \varphi_w(x)\mathrm{d}\gamma(w) \quad \text{where} \quad \gamma(\cdot) = \int_{\mathbb{R}} \alpha \mathrm{d}\Gamma(\cdot,\alpha) \tag{1.11}$$

where we recall that the ridge functions $\varphi_w(x) \triangleq \rho(\langle w,\, \bar{x}\rangle)$ were defined in (1.6). Furthermore, we recall that we can assume that Λ is supported on $\mathbb{R}^{d+1} \times B(0,\alpha_{\max})$ so that the total variation of γ is bounded as

$$\|\gamma\|_{\mathrm{TV}} = |\gamma|(\mathbb{R}^{d+1}) \leqslant \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}} |\alpha| \mathrm{d}\Gamma(w,\alpha) \leqslant \alpha_{\max}. \tag{1.12}$$

Assuming $f(0) = 0$ for simplicity, we denote the relation (1.11) as $f = \Phi\gamma$ where Φ is a linear operator between $\mathcal{M}(\mathbb{R}^{d+1})$ (the space of signed Radon measures) and $\mathcal{C}(\mathbb{R}^d)$ the space of continuous functions (here we assume ρ is continuous for simplicity).

Convex learning. In the limit $q \to +\infty$, one can thus replace the MLP f_q by functions of the form $\Phi(\gamma)$ with $\|\gamma\|_{\mathrm{TV}} \leqslant \alpha_{\max}$. This allows one to replace in the large q limit the non-convex minimization of the risk

$$\inf_{W,a} \frac{1}{n}\sum_{i=1}^{n} \ell(f_q(x_i;(W,a)),y_i)$$

by the following convex but infinite dimensional optimization problem [1]

$$\inf_{\gamma\in\mathcal{M}(\mathbb{R}^{d+1})} \mathcal{E}_0(\gamma) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell((\Phi\gamma)(x_i),y_i). \tag{1.13}$$

possibly with the additional constraint $\|\gamma\|_{\mathrm{TV}} \leqslant \alpha_{\max}$.

Note that this type of TV constraint corresponds to defining a norm on the space of functions

$$\|f\|_* \triangleq \inf_{\gamma\in\mathcal{M}(\mathbb{R}^{d+1})} \left\{ \|\gamma\|_{\mathrm{TV}} \;;\; f = \Phi\gamma \right\}.$$

One can show under some hypotheses that $\|\cdot\|_*$ is actually equivalent to $\|\cdot\|_B$.

Revisiting Barron's proof using Frank-Wolfe Barron gives an alternate proof of his theorem using a greedy algorithm. This is very similar to using the celebrated Frank-Wolfe algorithm, that we now detail in the case of the space of measure (this algorithm is very general and work on arbitrary vector space, not necessarily with Banach structure). This follows [1]. The idea is to consider the following "fake" approximation problem

$$\inf_{\|\gamma\|_{\mathrm{TV}}\leqslant\alpha_{\max}} \mathcal{E}_1(\gamma) = \frac{1}{2}\|\Phi\gamma - f\|^2_{L^2(\mu)} \tag{1.14}$$

for which we know from (1.12) that the minimum value is 0.

More generally, we consider the problem of constraint minimization (this can be generalized to penalized minimization)

$$\inf_{\gamma\in\mathcal{C}} \mathcal{E}(\gamma)$$

where $\mathcal{C}$ is a compact set of contraints for the strong topology and $\mathcal{E}$ is some function. We assume that this function has directional derivatives, that we write for ρ another measure

$$\mathcal{E}(\gamma + \varepsilon\lambda) = \mathcal{E}(\gamma) + \varepsilon\langle \delta\mathcal{E}(\gamma),\, \lambda\rangle + o(\varepsilon)$$

where $\delta\mathcal{E}(\gamma) \in \mathcal{C}(\mathbb{R}^{d+1})$ should be thought as the gradient of $\mathcal{E}$ at γ, when one replace a hilbertian inner product by the duality pairing between measures and continuous functions,

$$\forall (g,\lambda) \in \mathcal{C}(\mathbb{R}^{d+1}) \times \mathcal{M}(\mathbb{R}^{d+1}), \quad \langle g,\, \lambda\rangle \triangleq \int_{\mathbb{R}^{d+1}} g(w)\mathrm{d}\lambda(w).$$

For instant, in the cases (1.13) and (1.14), if we assume that ℓ is smooth, one has

$$\delta\mathcal{E}_0(\gamma) : w \in \mathbb{R}^{d+1} \mapsto \frac{1}{n}\sum_{i=1}^{n} \ell'((\Phi\gamma)(x_i), y_i)\varphi_w(x_i) \in \mathbb{R} \tag{1.15}$$

$$\delta\mathcal{E}_1(\gamma) : w \in \mathbb{R}^{d+1} \mapsto \int_{\mathbb{R}^d} (\Phi\gamma(x) - f(x))\varphi_w(x)\mathrm{d}\mu(x) \in \mathbb{R}. \tag{1.16}$$

Starting from $\gamma_0 \in \mathcal{C}$, the algorithm proceeds at step k by extracting an extremal point of the constraint set

$$\zeta_k \in \underset{\zeta\in\mathcal{M}(\mathbb{R}^{d+1})}{\operatorname{argmin}} \langle \delta\mathcal{E}(\gamma_k),\, \zeta\rangle \tag{1.17}$$

and then update

$$\gamma_{k+1} = (1-t_k)\gamma_k + t_k\zeta_k$$

where the step size can be for instance chosen as $t_k \sim \frac{1}{k}$ or using a line search. In the special case where

$$\mathcal{C} = \left\{\gamma \in \mathcal{M}(\mathbb{R}^{d+1}) \;;\; \|\gamma\|_{\mathrm{TV}} \leqslant \alpha_{\max}\right\}$$

then the extremal point are of the form $\zeta_k = \pm\alpha_{\max}\delta_{w_k}$ where the optimal weight thus solves an optimization over $\mathbb{R}^{d+1}$ of a non-convex function

$$w_k \in \underset{w\in\mathbb{R}^{d+1}}{\operatorname{argmax}} |[\delta\mathcal{E}(\gamma_k)](w)|$$

Unfortunately, even simply finding this sole neurone with theoretical guarantees is computationally hard.

In order to ensure the convergence (in value) of Frank-Wolfe, the function needs somehow to be L-smooth (Lipschitz gradient). Since we are not optimizing over a vector space, this property is controlled by upper-bounding the directional Hessian, which corresponds to assuming a curvature bound (bounding the function by above by a parabola along 1-D segments)

$$\forall\,(\gamma,\zeta) \in \mathcal{C}^2,\ \forall t \in [0,1], \quad \mathcal{E}((1-t)\gamma + t\zeta) \leqslant \mathcal{E}(\gamma) + t\langle\delta\mathcal{E}(\gamma),\, \zeta-\gamma\rangle + \frac{L}{2}t^2.$$

Under this hypothesis, one has the following convergence result

Proposition 3. *One has for the choice* $t_k = \frac{2}{2+k}$

$$\mathcal{E}(\gamma_k) - \inf_{\gamma\in\mathcal{C}} \mathcal{E}(\gamma) \leqslant \frac{L}{k+2}.$$

In the case of $\mathcal{E} = \mathcal{E}_0$ defines in (1.14), which is quadratic, one has

$$\mathcal{E}((1-t)\gamma + t\zeta) - \mathcal{E}(\gamma) - t\langle\delta\mathcal{E}(\gamma),\, \zeta-\gamma\rangle = \frac{t^2}{2}\|\Phi(\gamma-\zeta)\|^2_{L^2(\mu)}$$

and then denoting $\xi \triangleq \gamma - \zeta$, since $\|\varphi_w\|_\infty \leqslant 1$ and $\|\gamma-\zeta\|_{\mathrm{TV}} \leqslant 2\gamma_{\max}$ (because ζ and γ are in $\mathcal{C}$),

$$\frac{1}{2}\|\Phi(\gamma-\zeta)\|^2_{L^2(\mu)} = \frac{1}{2}\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^{d+1}} \varphi_w(x)\mathrm{d}(\gamma-\zeta)(w)\right)\mathrm{d}\mu(x) \leqslant \frac{1}{2}\int_{\mathbb{R}^d} \|\varphi_w\|^2_\infty\|\gamma-\zeta\|^2_{\mathrm{TV}}\mathrm{d}\mu(x) \leqslant \gamma_{\max}$$

so that $L \leqslant 2\gamma^2_{\max}$. Using the convergence speed of one recovers the approximation result of Barron for $f_q = \Phi\gamma_q$ computed at step q of F-W

$$\|f - f_q\|^2_{L^2(\mu)} = \mathcal{E}(\gamma_q) - \inf_{\gamma\in\mathcal{C}} \mathcal{E}(\gamma) \leqslant \frac{2\gamma^2_{\max}}{q+2}.$$

Bibliography

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

[2] Francis Bach. Learning theory from first principles. 2021.

[3] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[4] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.