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

December 16, 2024

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 Multi-layer Perceptron

1.1.1 Multi-layer

Let us first consider the general case of an arbitrary number of layers to defined mapping $f_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$. From $x_0 := x \in \mathbb{R}^{d_s} = \mathbb{R}^d$, they iterate along the depth indexes $s = 0, \ldots, S-1$

$$x_{s+1} + \sigma(W_s x_s + b_s)$$

where $W_s \in \mathbb{R}^{d_{s+1} \times d_s}$ and the bias is $b_s \in \mathbb{R}^{d_{s+1}}$. Here σ is a non-linear (and in fact non-polynomiial) function applied component wise, i.e. we denote $\sigma(z) = (\sigma(z_i))_i$.

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 $\rho(r) = \max(r, 0)$. There is an important difference both in practice and in theory on these two class of activation (bounded vs. un-bounded). ReLu works better in practice because there is less saturaration effect, so that gradient are not zero if the values commputed by the networks are large. Also the ReLu is positively 1-homogeneous, which allows to rescale the weights and for some proof, consider that these weight are on a unit sphere. A difficulty however is that the ReLu is not differentiable at 0, which makes some rigorous proof difficult to do (but in practice, this non smoothness seems harmless).

In order to define function of arbitrary complexity when width (number of neuron per layer) increases, it is important that σ is non-polynomial. Otherwise, f_θ would be a polynimial of degree proportional to s, so these functions would for instance not be dense in continuous functions. Note however that the linear case $\sigma = \text{Id}$ is of independent to compute matrix factorization, but this does not corresponds to supervised learning problem (but rather dimensionality reduction using PCA or non-negative matrix factorization).

1.1.2 2-layers MLPs

We consider two-layer neural networks of the form:

$$f_\theta(x) := \sum_{k=1}^{n} u_k \sigma(\langle v_k, x\rangle + b_k), \quad \forall x \in \mathbb{R}^d, \tag{1.1}$$

where σ is the activation function. The parameters of the network are denoted as $\theta_k = (u_k \in \mathbb{R}^{d'}, v_k \in \mathbb{R}^d, b_k \in \mathbb{R})$ for $k = 1, \dots, n$. In most of the following, for the sake of simplicity, we consider $d' = 1$, i.e. real-valued output.

In practice, neural network are designed by doing gradient descent, i.e. we consider a loss (for the sake of simplicity here quadratic

$$\min_\theta E(\theta) := \int \|f_\theta(x) - y\| \mathrm{d}\rho(x, y) \tag{1.2}$$

and use the gradient (it is of course possible to use SGD)

$$\theta_{t+1} = \theta_t - \tau_t \nabla E(\theta).$$

Gradient computation Ignoring the bias b_k for simplicity, we can write in matrix form $f_\theta(x) = U\sigma(V^\top x)$ where $U \in \mathbb{R}^{d' \times n}$, $V \in \mathbb{R}^{d \times n}$. For the sake of simplicity, we assume there is a finite number N of data points $X = (x_i)_{i=1}^N \in \mathbb{R}^{d \times N}$ and $Y = (y_i)_{i=1}^N \in \mathbb{R}^{d' \times N}$ Training with a ℓ^2 loss thus reads

$$\min_{U,V} E(U, V) := \frac{1}{2}\|U\sigma(V^\top X) - Y\|^2.$$

If we denote $Z := \sigma(V^\top X)$ (which can be thought as applying the feature map $x \to \sigma(V^\top x)$ to the data), then training U is a classical least square $\frac{1}{2}\|UZ - Y\|^2$ and the gradient reads

$$\nabla_U E(U, V) = (UZ - Y)Z^\top.$$

We perform a Taylor expansion to compute the gradient with respect to V, denoting $R := U\sigma(V^\top X) - Y$ and $S := \sigma'(V^\top X)$

$$E(U, V + \varepsilon D) = \frac{1}{2}\|R + \varepsilon U[S \odot (D^\top X)]\|^2 = E(U, V) + \langle R,\, H\rangle + O(\varepsilon^2)$$

where we denoted $A \odot B = (A_{i,j}B_{i,j})$ and where

$$\langle R,\, H\rangle = \langle R,\, U[S \odot (D^\top X)]\rangle = \langle D^\top X,\, (U^\top R) \odot S\rangle = \langle X^\top D,\, (R^\top U) \odot S^\top\rangle$$

which leads to

$$\nabla_V E(U, V) = X[(R^\top U) \odot S^\top].$$

This computation is quite painful, and the advice is not to use this derivation for deeper network, because not only they are overly complicated, but they are vastly sub-optimal. The correct way to compute this gradient is to use the back-propagation method, which corresponds to revserse mode automatic differentiation.

1.2 L^∞ non-quantitative universality

If σ is a sigmoid function, George Cybenko's theorem, later refined by Kurt Hornik, Maxwell Stinchcombe, and Halbert White, demonstrates that the functions f_θ can approximate any continuous function uniformly on a compact domain. So this means here we insist on doing an L^∞ approximation, which is strictly stronger (and more difficult) that doing an L^2 error control as consider during training (1.2).

Proposition 1. *If σ is an increasing (not necessarily continuous) function satisfying:*

$$\lim_{s\to-\infty}\sigma(s)=0 \quad \textit{and} \quad \lim_{s\to+\infty}\sigma(s)=1,$$

and $K\subset\mathbb{R}^d$ is compact, then for any continuous function f on K and any $\varepsilon>0$, there exist n and parameters $(\theta_k)_{k=0}^n$ such that:

$$\sup_{x\in K}|f(x)-f_\theta(x)|\leqslant\varepsilon.$$

This theorem establishes the universal approximation property of two-layer neural networks. However, it does not provide bounds on the number of neurons n required as a function of ε. Furthermore, the proof does not constructively specify how to determine the parameters of the approximating network f_θ. The first proof was done by Cybenko [2] using a duality argument. We detail next the proof due to Hornik et al. which is a bit more constructive, and rely on Stone-Weierstrass theorem to perform a Fourier-type approximation. On contrary to a direct Fourier series expansion, this leads to a uniform approximation of a continuous function, whereas Fourier series do not lead to a uniform approximation.

Proof. It first considers the activation $\sigma=\cos$ (note that the initial density argument would also work with $\sigma=\exp$ which interestingly is a non-bounded activation). Consider the function space:

$$A:=\left\{\sum_{k=1}^n u_k\cos(\langle v_k,x\rangle+b_k):n\in\mathbb{N},(u_k,b_k,v_k)_k\right\}.$$

This space is an algebra of continuous functions on the compact set K. It contains the constant functions and separates points; that is, for $x\neq x'$, there exists w such that $\cos(\langle w,x\rangle)\neq\cos(\langle w,x'\rangle)$. By the Stone-Weierstrass theorem, A is dense in the space of continuous functions on K.

Let $r=\max_k(|v_k|\cdot\mathrm{Radius}(K)+|b_k|)$. To approximate functions on K, it thus suffices by the previous density to approximate $\cos(s)$ on the interval $[-r,r]$. Splitting the interval into subintervals where $\cos(s)$ is monotonic, this can be replaced by just approximating the rectified cosine squashing function :

$$\cos_+(s)=\begin{cases}0, & s\leqslant 0,\\ 1, & s\geqslant\pi/2,\\ 1-\cos(s), & s\in[0,\pi/2].\end{cases}$$

The goal is to construct σ-based functions of the form:

$$\left|\sum_k u_k\sigma(v_k s+b_k)-\cos_+(s)\right|\leqslant\varepsilon,$$

where $u_k,b_k,v_k\in\mathbb{R}$.

Divide $[0,\pi/2]$ into Q subintervals $[s_k,s_{k+1}]$, where $s_k=\cos_+^{-1}(k/Q)$. Choose $M>0$ large enough such that:

$$\sigma(-M)<\frac{\varepsilon}{2Q},\quad \sigma(M)>1-\frac{\varepsilon}{2Q}.$$

Define v_k and b_k such that the affine map $v_k s+b_k$ sends $[s_k,s_{k+1}]$ to $[-M,M]$. Set the weights $u_k=1/Q$. For each subinterval, the construction ensures that:

$$|\sigma(v_k s+b_k)-\cos_+(s)|\leqslant\frac{\varepsilon}{Q}.$$

Summing over all subintervals gives the desired approximation:

$$\left|a_0+\sum_k u_k\sigma(v_k s+b_k)-\cos_+(s)\right|\leqslant\varepsilon,$$

provided $Q>2/\varepsilon$. Combining the results, the network f_θ can approximate any continuous function f on K to within ε, completing the proof. □

1.3 L^2 Quantitative Approximation (Barron's theorem)

In constrast to the uniform error control of the previous section, we consider here L^2 approximation as consider in the initial loss (1.2). We only focuss on approximation error, so that we consider that the data satisfy exactly $y = f(x)$ for some function f to approximate and x is distributed according to some $\rho(x)$. Another limitation of the theory we detail next is that we assume ρ is compactly supported on a ball of radius R. Without any hypothesis beside convexity, it is not possible to show any rate (i.e. approximation by a network can be arbitrary slow). The functional space to obtain fast rate (independent of the dimension) is called the Barron's space, and was introduced by Andrew Barron.

1.3.1 Barron's space

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

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

The Barron's space [1] is the set of functions such as the semi-norm

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

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

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

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

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

  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

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

  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.

1.3.2 Barron's Theorem

The main result is as follows.

Theorem 1 (Barron [1]). *We assume ρ is supported on $B(0,R)$. For all n, there exists f_θ with n neurons such that*

$$\|f(0) + f_\theta - f\|_{L^2(\rho)} \leqslant \frac{2R\|f\|_B}{\sqrt{n}}.$$

Furthermore, one can impose that $\sum_k |u_k| \leqslant 2R\|f\|_B$

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.

1.3.3 Mean field representation.

The proof of Barron's theorem involves rescaling the coefficients u_k by $1/n$ and rewriting the neural network in Equation (1.1) as:

$$f_\theta(x) := \frac{1}{n}\sum_{k=1}^{n} \varphi(x, \omega_k),$$

where $\theta = (\omega_k)_{k=1}^n$, $\omega_k = (u_k, v_k, b_k) \in \mathbb{R}^{d'} \times \mathbb{R}^{d+1}$ and $\varphi(x,\omega) := u\sigma(\langle v, x\rangle + b)$. Introducing the empirical measure:

$$\hat{\mu} := \frac{1}{n}\sum_{k=1}^{n} \delta_{\omega_k},$$

this neural network can be expressed as an integral:

$$f_\theta(x) := \int_\Omega \varphi(x,\omega)\,\mathrm{d}\hat{\mu}(\omega)$$

where $\Omega \subset \mathbb{R}^{d'} \times \mathbb{R}^{d+1}$ is the set of considered parameter (we will see bellow that it is important to be able to restrict u to belong to a compact domain). An advantage of this integral representation is that it is linear in the measure μ. This eliminates the need to restrict to discrete measures and allows for a general probabilistic interpretation of μ.

For the sake of simplicity, we consider the 1-D ouput case, $d = 1$. The core of Barron's theorem demonstrates that if the Barron norm of f, $\|f\|_B$, is finite, then f can be represented by a measure.

Proposition 2. *If $\|f\|_B < +\infty$, there exists a probability measure μ such that:*

$$f(x) = \Phi(\mu)(x),$$

where:

$$\Phi(\mu)(x) := \int_\Omega \varphi(x,\omega)\,\mathrm{d}\mu(\omega). \tag{1.3}$$

Furthermore, the measure μ can be restricted to a compact support on the outer weights, $\mathrm{supp}(\mu) \subset \Omega$ where

$$\Omega := [-M, M] \otimes \mathbb{R}^{d+1},$$

where $M := R\|f\|_B$, and R is the radius of the domain K on which the approximation is performed.

Proof. We only sketch the construction. Using the inverse Fourier transform and the fact that $f(x)$ is real, one has

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

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

Note that

$$|g_\xi(x)| \leqslant \frac{\|f\|_B}{\|\xi\|}|\langle \xi,\, x\rangle| \leqslant \|f\|_B R$$

so that g_ξ are similar to bounded sigmoid functions. This calculation shows that the previous decomposition (**??**) is true but with sigmoid functions g_ξ instead of functions g_ω. One then proceeds by showing that the function cos can be written using translates and dilates of the function ρ to obtain the thought after integral formula. □

1.3.4 Probabilistic proof

A first proof used the so-called "probabilistic method", which relies on drawing a random neural network and showing that the probaiblity of reaching the desired $O(1/n)$ error is non zero, thus showing the existence of a network with this error bound. We thus consider $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$, where the $(\omega_i)_i$ are now random vector, independent one from each other, and with law $\omega_i = \mu$, where μ is the measure so that $\Phi(\mu) = f$ constructed above. Beward that now $\Phi(\hat{\mu})$ is a random function, and note that

$$\mathbb{E}_{\hat{\mu}}(\Phi(\hat{\mu}))(x) = \frac{1}{n}\sum_i \mathbb{E}_{\omega_i}(\varphi(x,\omega_i)) = f(x)$$

i.e. $\mathbb{E}_{\hat{\mu}}(\Phi(\mu)) = f$. In the following, we denote $\varphi_\omega(x) = \varphi(x,\omega)$ for the ease of writing. We consider the average error according to the data distribution $\rho(x)$ on the x variable. This corresponds to the classical error in a Monte-Carlo estimation of an integral (excepted here that the value of the integral is a function and not just a scalar as it is usually the case). In the following, we use the short-hand notation $\|\cdot\| = \|\|_{L^2(\rho)}$ and the inner product are also for $L^2(\rho)$

$$\mathbb{E}_{\hat{\mu}}\|\Phi(\hat{\mu}) - f\|^2 = \mathbb{E}_{\hat{\mu}}\|\Phi(\hat{\mu})\|^2 - 2\langle \mathbb{E}_{\hat{\mu}}\Phi(\hat{\mu}),\, f\rangle + \|f\|_{L^2}^2 = \mathbb{E}_{\hat{\mu}}\|\Phi(\hat{\mu})\|^2 - \|f\|^2.$$

We now compute the first expectation, using the fact that for $i \neq j$, ω_i and ω_j are indepentent

$$\mathbb{E}_{\hat{\mu}}\|\Phi(\hat{\mu})\|^2 = \frac{1}{n^2}\sum_i \mathbb{E}_{\omega_i}\|\varphi_{\omega_i}\|^2 + \frac{1}{n^2}\sum_{i\neq j}\langle \mathbb{E}_{\omega_i}\varphi_{\omega_i},\, \mathbb{E}_{\omega_j}\varphi_{\omega_j}\rangle = \frac{1}{n}\mathbb{E}_\omega\|\varphi_\omega\|^2 + (1-\frac{1}{n})\|f\|^2.$$

Putting all this together leads to the bound

$$\mathbb{E}_{\hat{\mu}}\|\Phi(\hat{\mu}) - f\|^2 = \frac{\mathbb{E}_\omega\|\varphi_\omega\|^2 - \|f\|^2}{n} \leqslant \frac{\mathbb{E}_\omega\|\varphi_\omega\|^2}{n}$$

One has $\mathbb{E}_\omega\|\varphi_\omega\|^2 \leqslant \|\varphi\|_{K\times\Omega}^2 \leqslant C := R^2\|f\|_B^2\|\sigma\|_\infty^2$. So this means that the probability of the event $\|\Phi(\hat{\mu}) - f\|^2 \leqslant C/n$ holds is non zero, hence the proof of the theorem.

1.3.5 Proof by optimization

A second proof is fully deterministic and relies on using n step of an optimization algorithm (Frank-Wolfe method), for which an $O(1/n)$ convergence rate is known. To prove the existence of such a discrete measure achieving the desired error, we thus consider the following approximation problem over the space of probability measures $\mathcal{P}(\Omega)$:

$$\inf_{\mu\in\mathcal{P}(\Omega)} E(\mu) := \frac{1}{2}\int_K (\Phi(\mu)(x) - f(x))^2 \,\mathrm{d}x, \tag{1.4}$$

where $\mathrm{d}x$ is the integration measure with support on K. This optimization problem is infinite-dimensional.

The classical way to solve it is to restrict the previous optimization to discrete measure $\hat{\mu} = \sum_i \delta_{\omega_i}$ with n neurons and perform gradient descent on the neuron's parameter $(\omega_i)_i$. Since the function is non-convex this might be trapped in a local minima. A recent breakthrough was recently obtained by Chizat and Bach. They made the remark that this flow is equialent to a Wasserstein gradient flow (a gradient flow for the optimal transport distance). This allows one to consider the mean field limit when $n \to +\infty$, and in this limit, provided that the initialization has a density, they showed that this flow can never be trapped in a local minimizer. This in turn ensure that, if the number of neurons n is large enough, and if these are initialized at random according to some distribution with a density, then the usual gradient descent cannot be trapped in a local minimum (if it converges, it converges to the global minimizer, hence to a 0 loss). Note however that it is not possible to know how many neurons are needed for this conclusion to holds, so it is not known wehter it is possible to reach the $O(1/n)$ rate with a gradient descent algorithm.

To make the proof, one has to rely on another algorithm with known convergence guarantees, which relies on classical convex optimization. The advantage is that it leads to a constructive proof, but the issue is that this algorithm relies on the computatino of an oracle which is a priori not tractable (exact optimization of a single neurons). So this algorithm cannot be used in practice in high dimensions.

First order variations. To derive this algorithm, we have to rely on linearization, which we detail in the general context of a Banach space (but it can in fact be done even more abstractly without a norm structure by only relying on directional derivative) and can be applied for our concern over the space of probabilty equipped with the total variation norm. In the following, to ease the description, we denote the integration as a pairing between functions and measure using an inner product notation

$$\langle f,\, \mu \rangle := \int f(x) \mathrm{d}\mu(x).$$

Let $\mu + \varepsilon\rho$ be a small perturbation of μ, where ρ is another measure. Then the first variation $\nabla E(\mu)$ is a function defined using the Frechet directional derivative rule

$$E(\mu + \varepsilon\rho) = E(\mu) + \varepsilon\langle \nabla E(\mu),\, \rho \rangle + o(\varepsilon),$$

so that $\nabla E(\mu)$ is the FrÃ©chet derivative of E, also called the first variation. In our case, we have:

$$E(\mu + \varepsilon\rho) = \frac{1}{2} \int_K \left(\Phi(\mu)(x) - f(x) + \varepsilon\varphi(\rho)(x)\right)^2 \,\mathrm{d}x,$$

which expands to:

$$E(\mu + \varepsilon\rho) = E(\mu) + \varepsilon \int_K \varphi(\rho)(x) \left(\Phi(\mu)(x) - f(x)\right) \,\mathrm{d}x + O(\varepsilon^2).$$

Rewriting this in terms of φ, we find:

$$\nabla E(\mu)(\omega) = \int_K \varphi(x, \omega) \left(\Phi(\mu)(x) - f(x)\right) \,\mathrm{d}x.$$

which is a continuous function.

Frank-Wolfe algorithm. The Frank-Wolfe algorithm seeks to minimize a function on a convex sub-set of a Banach space,

$$\min_{\mu \in \mathcal{C}} E(\mu).$$

It operates by successive linearization of the objective function $E(\mu)$. It initializes μ_0 arbitrarily (e.g., as a Dirac measure). At each iteration k, for a step size τ_k, the measure is updated as:

$$\mu_{k+1} = (1 - \tau_k)\mu_k + \tau_k \nu_k^*,$$

where ν_k^* is a measure minimizing the linearized functional:

$$\nu_k^* \in \arg\min_{\nu \in \mathcal{P}(\Omega)} \langle \nabla E(\mu_k),\, \nu \rangle$$

We call this computation of ν_k an "oracle" since a priori it is not always simple to obtain. In finite dimension, if the measure are on a grid, this can be carried over, but as we will see, in the general setting, it requires the resolution of a non-convex optimization over the space Ω of (single) neurons. In our specific case, $\mathcal{C} = \mathcal{P}(\Omega)$ is endowed with the total variation norm $\|\mu\|_{\mathrm{TV}} = |\mu|(\Omega)$ (it is the extension to measure of the L^1 norm of functions). In this special case, a key property of the algorithm is that the solution ν_k^* can always be taken as a Dirac measure since, denoting $g_k := \nabla E(\mu_k)$,

$$\nu_k^* = \delta_{\omega_k^*}, \quad \text{where } \omega_k^* \in \arg\min_{\omega \in \Omega} g_k(\omega).$$

This holds because for any $\nu \in \mathcal{P}(\Omega)$:

$$\int g_k(\omega)\, \mathrm{d}\nu(\omega) \geqslant \min(g_k),$$

and equality is achieved when $\nu = \delta_{\omega_k^*}$. Therefore, if μ_0 is initialized as a Dirac measure, each iteration of the algorithm ensures that μ_k remains a sum of at most $k+1$ Dirac masses.

Convergence Rate The following theorem establishes the convergence rate of the Frank-Wolfe algorithm. In our case, $\|\cdot\| = \|\cdot\|_{\mathrm{TV}}$ is the total variation norm of measure, and the dual norm $\|\cdot\|_* = \|\cdot\|_\infty$ is the L^∞ norm on function. We first recall that a function with a Lipschitz gradient has a quadratic upper bound.

Lemma 1. *Let $E : \mathcal{C} \to \mathbb{R}$ be a differentiable function with L-Lipschitz gradient with respect to the norm $\|\cdot\|$. That is, for all $\mu, \nu \in \mathcal{C}$:*

$$\|\nabla E(\mu) - \nabla E(\nu)\|_* \leqslant L\|\nu - \mu\|,$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. Then, for all $\mu, \nu \in \mathcal{C}$:

$$E(\nu) \leqslant E(\mu) + \langle \nabla E(\mu), \nu - \mu \rangle + \frac{L}{2}\|\nu - \mu\|^2.$$

Proof. By the fundamental theorem of calculus, we can express $E(\nu)$ as:

$$E(\nu) = E(\mu) + \int_0^1 \langle \nabla E(\mu + t(\nu - \mu)), \nu - \mu \rangle\, dt.$$

Adding and subtracting $\nabla E(\mu)$ inside the integrand:

$$E(\nu) = E(\mu) + \langle \nabla E(\mu), \nu - \mu \rangle + \int_0^1 \langle \nabla E(\mu + t(\nu - \mu)) - \nabla E(\mu), \nu - \mu \rangle\, dt.$$

Using the L-Lipschitz property of the gradient, we bound the difference:

$$\|\nabla E(\mu + t(\nu - \mu)) - \nabla E(\mu)\|_* \leqslant Lt\|\nu - \mu\|.$$

Substitute this bound into the integral:

$$\left| \int_0^1 \langle \nabla E(\mu + t(\nu - \mu)) - \nabla E(\mu), \nu - \mu \rangle\, dt \right| \leqslant \int_0^1 Lt\|\nu - \mu\|^2\, dt.$$

Evaluate the integral:

$$\int_0^1 Lt\, dt = \frac{L}{2}.$$

Thus:

$$E(\nu) \leqslant E(\mu) + \langle \nabla E(\mu), \nu - \mu \rangle + \frac{L}{2}\|\nu - \mu\|^2.$$

□

Theorem 2. *Let E be convex and assume that $\nabla E(\mu)$ is L-Lipschitz, i.e.,*

$$\|\nabla E(\mu) - \nabla E(\mu')\|_* \leqslant L\|\mu - \mu'\|.$$

For the step size $\tau_k = \frac{2}{k+2}$, the F-W to optimize F on a set of radius

$$r := \sup_{\mu,\mu' \in \mathcal{C}^2} \|\mu - \mu'\|$$

satisfies, denoting $E^* := \inf_{\mu \in \mathcal{C}} E(\mu)$,

$$E(\mu_k) - E^* \leqslant \frac{2Lr^2}{k+1},$$

Proof. Using the L-Lipschitz gradient property with respect to a Banach norm $\|\cdot\|$, using Lemma 1, we have the following quadratic upper bound:

$$E(\nu) \leqslant E(\mu) + \langle \nabla E(\mu), \nu - \mu \rangle + \frac{L}{2}\|\nu - \mu\|^2, \quad \forall \mu, \nu \in \mathcal{C}.$$

One-Step Improvement The Frank-Wolfe update is:

$$\mu_{k+1} = \mu_k + \tau_k(\nu_k - \mu_k),$$

where $\tau_k = \frac{2}{k+1}$ and $\nu_k = \arg\min_{\nu \in \mathcal{C}} \langle \nabla E(\mu_k), \nu \rangle$. By smoothness of F, we have:

$$E(\mu_{k+1}) \leqslant E(\mu_k) + \tau_k \langle \nabla E(\mu_k), \nu_k - \mu_k \rangle + \frac{L}{2}\tau_k^2 \|\nu_k - \mu_k\|_{\mathrm{TV}}^2.$$

Furthermore, the boundedness of $\mathcal{C}$ ensures $\|\nu_k - \mu_k\| \leqslant r$. Substituting, we get:

$$E(\mu_{k+1}) \leqslant E(\mu_k) + \tau_k g_k + \frac{L}{2}\tau_k^2 r^2,$$

where $g_k := \langle \nabla E(\mu_k), \nu_k - \mu_k \rangle$ Defining $h_k = E(\mu_k) - E^*$ as the suboptimality at iteration k, we have:

$$h_{k+1} \leqslant h_k - \tau_k g_k + \frac{L}{2}\tau_k^2 r^2. \tag{1.5}$$

We now bound g_k, using the optimality of ν_k

$$g_k := \langle \nabla E(\mu_k), \nu_k - \mu_k \rangle = \min_{\nu \in \mathcal{C}} \langle \nabla E(\mu_k),\, \nu - \mu_k \rangle$$

and by convexity,

$$E(\nu_k) \geqslant E(\mu_k) + \langle \nabla E(\mu_k),\, \nu_k - \mu_k \rangle$$

so that $\langle \nabla E(\mu_k),\, \nu - \mu_k \rangle \leqslant E(\nu) - E(\mu_k)$ so

$$g_k \leqslant \min_{\nu \in \mathcal{C}} E(\nu) - E(\mu_k) = E^* - E(\mu_k) = -h_k.$$

Plugging this into (1.5), we obtained the fundamental descent property

$$h_{k+1} \leqslant h_k - \tau_k h_k + \frac{L}{2}\tau_k^2 r^2.$$

Substituting $\tau_k = \frac{2}{k+1}$:

$$h_{k+1} \leqslant h_k \left(1 - \frac{2}{k+1}\right) + \frac{2Lr^2}{(k+1)^2}.$$

Recursion Argument Assume the inductive hypothesis:

$$h_k \leqslant \frac{2Lr^2}{k+1}.$$

We will prove that:

$$h_{k+1} \leqslant \frac{2Lr^2}{k+2}.$$

Using the inductive hypothesis in the recursive relation for h_{k+1}:

$$h_{k+1} \leqslant \frac{2Lr^2}{k+1}\left(1 - \frac{2}{k+1}\right) + \frac{2Lr^2}{(k+1)^2}.$$

Simplify the coefficient:

$$\frac{2Lr^2}{k+1}\left(1 - \frac{2}{k+1}\right) = \frac{2Lr^2}{k+1} \cdot \frac{k-1}{k+1}.$$

Substitute back:

$$h_{k+1} \leqslant \frac{2Lr^2(k-1)}{(k+1)^2} + \frac{2Lr^2}{(k+1)^2}.$$

Combine terms:

$$h_{k+1} \leqslant \frac{2Lr^2\left((k-1)+1\right)}{(k+1)^2} = \frac{2Lr^2}{k+2}.$$

□

In the case of MLP training, where E is defined in (1.4), the proposition bellow shows that $L \leqslant M^2\|\sigma\|_\infty^2$, where $M = R\|f\|_B$, and we have $r = 2$ (radius of the space of probability for TV). Recall that$\|f\|_B$ is the Barron norm of the target function f. Furthermore, we know that $E(\mu^*) = 0$, as the existence of a valid representative measure was established in Equation (1.3). By applying the Frank-Wolfe algorithm, we deduce the existence of a discrete measure μ_k consisting of at most $k+1$ Dirac masses. This discrete measure achieves an approximation error:

$$E(\mu_k) = O\left(\frac{1}{k}\right).$$

Thus, the Frank-Wolfe algorithm constructs a sparse representation of the target function with a provably decreasing error bound as the number of iterations k increases.

Proposition 3. *The first variation $\nabla E(\mu)$ of the functional $E(\mu)$ defined in (1.4), which is*

$$\nabla E(\mu)(\omega) = \int_K \varphi(x,\omega)\left(\Phi(\mu)(x) - f(x)\right)\,\mathrm{d}x,$$

where $\Phi(\mu)(x) = \int_\Omega \varphi(x,\omega)\,\mathrm{d}\mu(\omega)$ and $\varphi(x,\omega) = a\sigma(\langle w, x\rangle + b)$, is L-Lipschitz with respect to the total variation norm. Specifically, for any $\mu, \mu' \in \mathcal{P}(\Omega)$:

$$\|\nabla E(\mu) - \nabla E(\mu')\|_\infty \leqslant L\|\mu - \mu'\|_{TV},$$

where $L = \|\sigma\|_\infty^2 M^2$.

Proof. The difference of $\nabla E(\mu)$ and $\nabla E(\mu')$ is:

$$\nabla E(\mu)(\omega) - \nabla E(\mu')(\omega) = \int_K \varphi(x,\omega)\left(\Phi(\mu)(x) - \varphi(\mu')(x)\right)\,\mathrm{d}x.$$

$$\Phi(\mu)(x) - \varphi(\mu')(x) = \int_\Omega \varphi(x,\omega)\,\mathrm{d}(\mu - \mu')(\omega).$$

Substitute the above into the expression for $\nabla E(\mu)$:

$$\nabla E(\mu)(\omega) - \nabla E(\mu')(\omega) = \int_K \varphi(x,\omega) \left(\int_\Omega \varphi(x,\omega') \, \mathrm{d}(\mu - \mu')(\omega') \right) \mathrm{d}x.$$

Using Fubini's theorem, introducing $k(\omega,\omega') := \int_K \varphi(x,\omega)\varphi(x,\omega') \, \mathrm{d}x$,

$$\nabla E(\mu)(\omega) - \nabla E(\mu')(\omega) = \int_\Omega k(\omega,\omega') \mathrm{d}(\mu - \mu')(\omega').$$

Take the L^∞ norm with respect to ω:

$$\|\nabla E(\mu) - \nabla E(\mu')\|_\infty = \sup_{\omega \in \Omega} \left| \int_\Omega k(\omega,\omega') \mathrm{d}(\mu - \mu')(\omega') \right|.$$

Using the triangle inequality:

$$\|\nabla E(\mu) - \nabla E(\mu')\|_\infty \leqslant \|k\|_{L^\infty(K\times K)} \|\mu - \mu'\|_{\mathrm{TV}}.$$

One has

$$\|k\|_{L^\infty(K\times K)} = \sup_{(\omega,\omega')\in\Omega^2} | \int_K \varphi(x,\omega)\varphi(x,\omega') \mathrm{d}x| \leqslant \|\varphi\|^2_{L^\infty(K\times\Omega)} \leqslant M^2 \|\sigma\|^2_\infty.$$

□

Bibliography

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

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