Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA, École Normale Supérieure

CHAPTER 16

Shallow Learning

September 9, 2026

Networks with one hidden layer provide a setting in which to ask which functions neural models can represent and how their size controls approximation error. These questions distinguish expressive power from the difficulty of fitting a model to data. We derive the training gradients, prove universal approximation, and develop quantitative bounds through Barron's theorem, a representation by measures, and the Frank–Wolfe algorithm.

16.1 Multilayer Perceptrons

A multilayer perceptron (MLP) with one hidden layer has two trainable layers: a hidden layer and an output layer. Models with no hidden layer recover the linear predictors studied earlier. Deeper MLPs compose several layers; their derivatives are efficiently computed by the automatic differentiation methods of the multilayer-perceptron discussion.

16.1.1 Networks of Arbitrary Depth

We first define a network of arbitrary depth representing a map $f_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$. Starting from $x_0 = x \in \mathbb{R}^{d_0} = \mathbb{R}^d$, the network computes, for $s = 0, \dots, S-1$

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

where $W_s \in \mathbb{R}^{d_{s+1} \times d_s}$ and $b_s \in \mathbb{R}^{d_{s+1}}$. The output is $f_\theta(x) = x_S$, with $d_S = d'$ and parameters $\theta = ((W_s, b_s))_{s=0}^{S-1}$. The final activation can be omitted to obtain a linear output layer. The activation σ is a nonlinear, nonpolynomial function applied componentwise: $\sigma(z) = (\sigma(z_i))_i$.

Common activations include the sigmoid functions

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

and the rectified linear unit (ReLU), $\sigma(r) = \max(r, 0)$. Sigmoids are bounded, whereas ReLU is unbounded and does not saturate on the positive half-line: its derivative is one there and zero on the negative half-line. The positive 1-homogeneity of ReLU allows weights to be rescaled, which is useful in proofs that restrict their directions to a unit sphere. At zero, ReLU is not differentiable. Implementations choose a derivative convention, while mathematical arguments must account for the kink.

For increasing width to yield universal approximation at fixed depth, the activation σ must be nonpolynomial. Otherwise, f_θ has degree at most m^S for an activation of degree m and fixed depth S, so increasing the width cannot produce a dense class of continuous functions. The choice $\sigma = \mathrm{Id}$ remains useful for matrix factorization and supervised learning, but its input-output map is affine (linear when the biases vanish).

16.1.2 Two-layer MLPs

We now specialize to networks with one hidden layer and a linear output layer:

$$f_\theta(x) := \sum_{k=1}^{n} u_k \sigma(\langle v_k, x \rangle + b_k), \quad \forall x \in \mathbb{R}^d, \tag{16.1}$$

Here σ is the activation, and the neuron parameters are $\theta_k = (u_k \in \mathbb{R}^{d'}, v_k \in \mathbb{R}^d, b_k \in \mathbb{R})$ for $k = 1, \dots, n$. We usually take $d' = 1$ for simplicity, so the output is scalar.

In practice, neural networks are trained by gradient methods. For example, consider a quadratic loss

$$\min_\theta E(\theta) := \frac{1}{2} \int \|f_\theta(x) - y\|^2 \mathrm{d}\rho(x, y) \tag{16.2}$$

where ρ is the joint distribution of the inputs and outputs. Taking ρ to be the empirical distribution gives the training loss. Gradient descent, or its stochastic counterpart, updates the parameters by

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

Gradient computation Omitting the biases b_k, the network has the matrix form $f_\theta(x) = U\sigma(V^\top x)$, where $U \in \mathbb{R}^{d' \times n}$ and $V \in \mathbb{R}^{d \times n}$. For a training set with N observations, collect the inputs as the columns of $X = (x_i)_{i=1}^N \in \mathbb{R}^{d \times N}$ and the outputs as the columns of $Y = (y_i)_{i=1}^N \in \mathbb{R}^{d' \times N}$. Here N counts observations and n counts hidden neurons. All matrix norms and inner products in this computation are Frobenius norms and inner products. Omitting the constant factor $1/N$, training with squared loss gives

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

Set $Z := \sigma(V^\top X)$, the matrix obtained by applying the feature map $x \to \sigma(V^\top x)$ to the data. With these features fixed, training U is the least-squares problem $\frac{1}{2}\|UZ - Y\|^2$, whose gradient is

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

Assume σ is differentiable at the entries of $V^\top X$. To differentiate with respect to V, expand the objective in a perturbation D and set $R := U\sigma(V^\top X) - Y$ and $S := \sigma'(V^\top X)$:

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

where $H = U[S \odot (D^\top X)]$, the product $A \odot B = (A_{i,j} B_{i,j})$ is coordinatewise, and

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

which leads to

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

For deeper networks, writing a fully expanded derivative becomes cumbersome and can duplicate computations. Backpropagation organizes the same chain rule into an efficient reverse pass.

16.2 L^∞ Nonquantitative Universal Approximation

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. Uniform approximation is stronger than the mean-square approximation used in the training loss (16.2).

Proposition 16.1. *Let $K \subset \mathbb{R}^d$ be compact, and let σ be nondecreasing, not necessarily continuous, with*

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

Then, for any continuous function f on K and any $\varepsilon > 0$, there exist n and parameters $(\theta_k)_{k=1}^n$ such that

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

The theorem establishes universal approximation by two-layer networks. It gives no quantitative bound on the number of neurons n needed for accuracy ε, nor an efficient procedure for finding the parameters of f_θ. Cybenko [2] proved the result by duality. We present the more constructive argument of Hornik and collaborators, based on Stone–Weierstrass approximation by trigonometric functions. This density argument avoids relying on uniform convergence of Fourier partial sums, which need not hold for an arbitrary continuous function.

Proof. First establish density for networks with cosine activation; we will then approximate each cosine using the prescribed sigmoid σ. 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: the product-to-sum identity expresses a product of cosines as a sum of two cosines. It contains constants and separates points. Indeed, for $x \neq x'$, choose v so that $\langle v,\, x' - x \rangle = \pi$ and set $b = -\langle v,\, x \rangle$; the resulting cosine takes values 1 and -1 at these points. The Stone–Weierstrass theorem therefore implies that A is dense in $C(K)$.

It remains to approximate each univariate cosine on a bounded interval by finite linear combinations of σ. We give an argument that applies to any continuous function q on $[a, b]$. Choose a fine partition $a = t_0 < \cdots < t_Q = b$ and set $\Delta_j = q(t_j) - q(t_{j-1})$. Uniform continuity makes every $|\Delta_j|$ and the oscillation of q on each interval small. The staircase

$$q(t_0) + \sum_{j=1}^{Q} \Delta_j 1_{[t_j, +\infty)}(s)$$

approximates q uniformly, up to this oscillation. Replace the step functions by $\sigma(M(s - t_j))$. Choose disjoint small neighborhoods of the finitely many t_j. Outside these neighborhoods, convergence to the steps is uniform as $M \to \infty$; inside any one neighborhood, only one transition contributes a potentially large discrepancy, bounded by the small jump $|\Delta_j|$. Since $0 \leqslant \sigma \leqslant 1$, the resulting approximation is uniform. A constant is itself a neuron with zero inner weight and an appropriately scaled nonzero activation value.

Apply this construction to each cosine in a finite approximation of f, choosing the individual errors small enough after weighting by the coefficients. Substitution of $s = \langle v_k,\, x \rangle + b_k$ gives a finite network with uniform error at most ε. □

16.3 L^2 Quantitative Approximation (Barron's Theorem)

We now study L^2 approximation, as measured by the loss (16.2). To isolate approximation error, assume noiseless observations $y = f(x)$ for a target f, with inputs x distributed according to $\rho(x)$. We also assume that ρ is supported in a ball of radius R. Without additional regularity, no uniform approximation rate follows: the error can decay arbitrarily slowly. Barron introduced a function class for which the approximation rate has an exponent independent of the ambient dimension.

16.3.1 Barron Space

For an integrable function f, this section uses the following normalization of the Fourier transform at $\xi \in \mathbb{R}^d$:

$$\hat{f}(\xi) \triangleq (2\pi)^{-d} \int_{\mathbb{R}^d} f(x) e^{-\mathrm{i}\langle \xi, x \rangle} \mathrm{d}x.$$

The Barron space [1] is the class of functions for which the Fourier seminorm

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

is finite. Throughout this section, f denotes its continuous Fourier-inversion representative, with the convention $f(x) = \int \hat{f}(\xi) e^{\mathrm{i}\langle \xi, x \rangle} \mathrm{d}\xi$. For integrable f, finiteness of the displayed seminorm ensures that $\hat{f}$ is integrable: it is bounded near the origin, and its first moment controls the tail. This representative matters when the sampling measure has atoms. Constants can be handled separately; $|f(0)| + \|f\|_B$ removes the ambiguity due to constants in the corresponding Fourier-measure space. This Fourier-defined class should be distinguished from activation-dependent Barron spaces. One has

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

which relates the Fourier seminorm to first-order regularity. The following examples illustrate the seminorm and its dependence on the function class.

- *Gaussians*: for $f(x) = e^{-\|x\|^2/2}$, one has $\|f\|_B = \mathbb{E}\|Z\| \leqslant \sqrt{d}, \quad Z \sim \mathcal{N}(0, \mathrm{Id}_d)$

- *Ridge functions*: for $f(x) = \psi(\langle x,\, b\rangle + c)$, the Fourier transform may be a measure supported on a line rather than an integrable density. In the Fourier-measure extension of the seminorm,

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

  A sufficient condition is that ψ is compactly supported and belongs to $C^{2,\delta}(\mathbb{R})$ for some $\delta > 0$. Smoothness alone without decay is not enough for this Fourier integral.
- *Sobolev functions*: if $s > d/2 + 1$, the Cauchy–Schwarz inequality gives

$$\|f\|_B \leqslant C(d,s)\|f\|_{H^s}, \qquad \|f\|_{H^s}^2 \overset{\text{def.}}{=} \int_{\mathbb{R}^d} (1 + \|\xi\|^2)^s |\hat{f}(\xi)|^2 \mathrm{d}\xi.$$

  Indeed, $\int \|\xi\|^2 (1 + \|\xi\|^2)^{-s} \mathrm{d}\xi$ is finite precisely when $s > d/2 + 1$. For integer s, this squared Sobolev norm is equivalent, up to Fourier-normalization constants, to the sum of the squared L^2 norms of all weak derivatives of order at most s. This is a sufficient condition; the Fourier-measure class also contains ridge functions outside these ambient-space Sobolev classes.

16.3.2 Barron's Theorem

The main result is as follows.

Theorem 16.2 (Barron [1])**.** *Assume ρ is a probability measure supported on $B(0,R)$, f is the continuous Fourier-inversion representative described above with $\|f\|_B < \infty$, and σ is a nondecreasing sigmoid with limits 0 and 1. Allow an output bias c_0, so the approximation is $c_0 + f_\theta$. For every $n \geqslant 1$,*

$$\inf_{c_0,\theta} \|c_0 + f_\theta - f\|_{L^2(\rho)} \leqslant \frac{2R\|f\|_B}{\sqrt{n}}.$$

The infimum can be restricted to networks with $\sum_{k=1}^n |u_k| \leqslant 2R\|f\|_B$. In particular, for every $\varepsilon > 0$, a finite network attains the displayed bound plus ε.

The exponent $1/2$ in the approximation rate does not depend on the dimension. The constant $\|f\|_B$ can still depend on it: the Gaussian example gives growth of order $\sqrt{d}$, whereas the bound for a fixed ridge profile depends on the ridge direction and profile.

16.3.3 Mean field representation.

Absorb a factor n into the outer coefficients u_k to express the same network as an average:

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

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 admissible parameter set. A uniform bound on u will be essential for the approximation estimate. The representation is linear in the measure μ. It therefore extends naturally from empirical measures to general probability measures μ.

We now restrict to scalar outputs, $d' = 1$. The Fourier condition yields a bounded dictionary representation in a closure sense. This distinction matters: the assumptions do not automatically provide an exact probability measure on finite sigmoid parameters.

Proposition 16.3. *Let $M = 2R\|f\|_B$. On $K \subset B(0,R)$, the function $f - f(0)$ belongs to the closed convex hull, in $L^2(\rho)$, of the centered neurons*

$$\varphi_\omega(x) = u\big(\sigma(\langle v,\, x\rangle + b) - \sigma(b)\big), \qquad \omega = (u,v,b), \quad |u| \leqslant M.$$

Every dictionary element has $L^2(\rho)$ norm at most M. For general measures μ on parameters, write

$$\Phi(\mu)(x) = \int \varphi_\omega(x)\mathrm{d}\mu(\omega). \tag{16.3}$$

The proposition asserts membership in the closure of such averages, rather than existence of one exact average.

Proof sketch. Fourier inversion gives

$$f(x) - f(0) = \int \big(\cos(\langle \xi,\, x\rangle + \Theta(\xi)) - \cos\Theta(\xi)\big)|\hat f(\xi)|\mathrm{d}\xi,$$

where Θ is the phase of $\hat f$. For $a = R\|\xi\|$ and $|t| \leqslant a$,

$$\cos(t+\Theta) - \cos\Theta = \int_{-a}^{a} -\sin(s+\Theta)\big(1_{t\geqslant s} - 1_{0\geqslant s}\big)\mathrm{d}s.$$

The total variation of these coefficients is at most $2a$. After integration over ξ, it is therefore at most $2R\|f\|_B = M$. Step functions are limits of dilated sigmoids; choosing thresholds away from atoms of the projected data distribution, dominated convergence gives the same closed convex hull in $L^2(\rho)$. A zero atom absorbs any unused coefficient mass. Centering the neurons accounts for an output bias when the sum is rewritten in the form (16.1). □

For the next argument, write $g = f - f(0)$. First suppose $g = \Phi(\mu)$ exactly for a probability measure over this bounded dictionary. The same bound extends to g in its closed convex hull by approximating g arbitrarily closely before sampling.

16.3.4 Probabilistic proof

Draw independent parameters $\omega_1, \ldots, \omega_n$ with law μ and set $g_n = n^{-1}\sum_i \varphi_{\omega_i}$. In the Hilbert space $L^2(\rho)$, independence yields

$$\mathbb{E}g_n = g, \qquad \mathbb{E}\|g_n - g\|^2 = \frac{1}{n}\left(\mathbb{E}\|\varphi_\omega\|^2 - \|g\|^2\right) \leqslant \frac{M^2}{n}.$$

Consequently at least one realization has squared error at most M^2/n. For a target in the closed convex hull, choose an average g_ε with $\|g - g_\varepsilon\| \leqslant \varepsilon$ and apply the same argument. The resulting network has error at most $M/\sqrt{n} + \varepsilon$. Each outer coefficient has magnitude at most M/n, so their absolute sum is at most M. This proves the approximation theorem, including its coefficient bound.

16.3.5 Proof by optimization

A deterministic argument uses n Frank–Wolfe steps and its $O(1/n)$ objective convergence rate. For the centered target $g = f - f(0)$, minimize over probability measures $\mathcal{P}(\Omega)$:

$$\inf_{\mu\in\mathcal{P}(\Omega)} E(\mu) := \frac{1}{2}\int_K (\Phi(\mu)(x) - g(x))^2 \,\mathrm{d}\rho(x), \tag{16.4}$$

where ρ is the probability measure supported on K. This optimization problem is infinite-dimensional.

Ordinary network training instead restricts μ to an empirical measure and descends in the neuron parameters. This parameterization is nonconvex. Chizat and Bach relate suitable mean-field limits to Wasserstein gradient flows and prove global convergence under additional structural and initialization assumptions. Such results do not imply that every sufficiently wide finite network, initialized from an arbitrary density, avoids all nonglobal stationary points.

The deterministic approximation argument below uses convex optimization over measures. Its linear minimization oracle is a global optimization problem over a single neuron's parameters and may itself be computationally difficult in high dimensions. Since the parameter domain need not be compact, exact oracle minimizers may fail to exist; approximate oracles or closure arguments then replace exact minimization.

First order variations. We derive the algorithm by linearizing the objective on measures equipped with the total variation norm. The same construction applies to differentiable convex objectives on a Banach space. We write integration as a pairing between a function and a measure:

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

Let $\mu + \varepsilon\eta$ be a small perturbation of μ, where η is a finite signed measure. The first variation $\nabla E(\mu)$ represents the derivative through the expansion

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

Thus $\nabla E(\mu)$ represents the Fréchet derivative of E. For the present quadratic objective,

$$E(\mu + \varepsilon\eta) = \frac{1}{2}\int_K (\Phi(\mu)(x) - g(x) + \varepsilon\Phi(\eta)(x))^2 \,\mathrm{d}\rho(x),$$

which expands to:

$$E(\mu + \varepsilon\eta) = E(\mu) + \varepsilon \int_K \Phi(\eta)(x)\,(\Phi(\mu)(x) - g(x))\,\mathrm{d}\rho(x) + O(\varepsilon^2).$$

Rewriting this in terms of φ, we find:

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

It is a bounded function of ω; continuity follows when the activation and parameterization are continuous.

Frank-Wolfe algorithm. Frank–Wolfe minimizes a differentiable convex function over a convex subset of a Banach space:

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

Each step linearizes the objective $E(\mu)$. Choose an initial measure μ_0, for example a Dirac mass. At iteration k, update with step size τ_k:

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

where ν_k^* minimizes the linearized objective:

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

This linear minimization step is called an oracle. On a finite parameter grid it amounts to comparing finitely many values. Over a continuous neuron parameter space Ω, it is generally a nonconvex optimization problem. Here $C = \mathcal{P}(\Omega)$ carries the total variation norm $\|\mu\|_{\mathrm{TV}} = |\mu|(\Omega)$, the measure analogue of the L^1 norm. In this special case, a key property of the algorithm is that the solution ν_k^* can be taken as a Dirac measure when the minimum over Ω is attained 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 next theorem gives the Frank–Wolfe convergence rate. In our case, $\|\cdot\| = \|\cdot\|_{\mathrm{TV}}$ is the total variation norm of measures. Bounded functions paired with these measures carry the supremum norm $\|\cdot\|_\infty$. Throughout this argument, L^∞ denotes the pointwise supremum norm, so values at atoms are retained. We first establish the quadratic upper bound implied by a Lipschitz gradient.

Lemma 16.4. *Let C be convex, and let $E : C \to \mathbb{R}$ be differentiable with L-Lipschitz gradient with respect to the norm $\|\cdot\|$. That is, for all $\mu, \nu \in C$,*

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

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

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

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 16.5. *Let C be a bounded convex set of diameter $D = \sup_{\mu,\nu\in C}\|\mu - \nu\|$, and let E be convex with L-Lipschitz derivative in the corresponding dual norm. Assume the linear minimization oracle is attained. With steps $\tau_k = 2/(k+2)$ for $k \geqslant 0$, the Frank–Wolfe iterates satisfy, for $k \geqslant 1$,*

$$E(\mu_k) - E^* \leqslant \frac{2LD^2}{k+2}, \qquad E^* = \inf_{\mu\in C} E(\mu).$$

Proof. Write $h_k = E(\mu_k) - E^*$ and $g_k = \min_{\nu\in C}\langle \nabla E(\mu_k), \nu - \mu_k\rangle$. Convexity implies $g_k \leqslant -h_k$. Lemma 16.4 therefore gives

$$h_{k+1} \leqslant h_k + \tau_k g_k + \frac{L}{2}\tau_k^2 D^2 \leqslant (1 - \tau_k)h_k + \frac{L}{2}\tau_k^2 D^2. \tag{16.5}$$

Since $\tau_0 = 1$, $h_1 \leqslant LD^2/2 \leqslant 2LD^2/3$, establishing the base case. If $h_k \leqslant 2LD^2/(k+2)$, then

$$h_{k+1} \leqslant \frac{2LD^2(k+1)}{(k+2)^2} \leqslant \frac{2LD^2}{k+3},$$

because $(k+1)(k+3) \leqslant (k+2)^2$. This proves the claim by induction. □

For the objective E in (16.4), the proposition below gives $L \leqslant M^2$ with $M = 2R\|f\|_B$. The total variation diameter of probability measures is $D = 2$, and $\|f\|_B$ denotes the Barron seminorm of the target f. For the centered target $g = f - f(0)$, Proposition 16.3 gives $\inf E = 0$, even if the infimum is not attained by a measure on finite sigmoid parameters. The rate therefore concerns approximation of g. When the linear oracle attains its minimum, Frank–Wolfe constructs a measure μ_k supported on at most $k + 1$ points, with objective error

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

Thus each iteration adds at most one neuron and gives an $O(k^{-1/2})$ bound on the L^2 error. If the oracle is not attained, choosing an $O(1/k)$-accurate linear oracle adds an $O(1/k^2)$ term to the descent recurrence and preserves this order of approximation.

Proposition 16.6. *The first variation of the objective in* (16.4) *is*

$$\nabla E(\mu)(\omega) = \int_K \varphi(x,\omega)\,(\Phi(\mu)(x) - g(x))\,\mathrm{d}\rho(x),$$

where $\Phi(\mu)(x) = \int_\Omega \varphi(x,\omega)\,\mathrm{d}\mu(\omega)$ *and* $\varphi(x,\omega) = u\big(\sigma(\langle v, x\rangle + b) - \sigma(b)\big)$ *with* $|u| \leqslant M$. *This first variation is L-Lipschitz from the total variation norm to the supremum norm: for any* $\mu, \mu' \in \mathcal{P}(\Omega)$,

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

where $L = M^2$.

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

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

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

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

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

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

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

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:

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

One has

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

□

References

[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.