Monte Carlo, reproducing kernels and neural networks: explicit integral representations and quantitative bounds for two-layer ReLU networks

Anthony Lee, University of Bristol

arXiv:2604.23260

ETH Zurich, June 2026

Funded by EPSRC grant ProbAI: Mathematical and Computational Foundations of Probabilistic AI

Outline

Outline

A neural network

- Let $\varsigma = \max\{x, 0\}$ denote the ReLU activation function.
- An L layer ReLU neural network, viewed as a function $f_{\mathrm{NN}} : \mathbb{R}^d \to \mathbb{R}$, may be written as

$$f_{\mathrm{NN}} = f_L \circ \cdots \circ f_1,$$

where for $\ell \in \{1, \ldots, L-1\}$, with W_ℓ a matrix and b_ℓ a vector,

$$f_\ell(x) = \varsigma\left(W_\ell x + b_\ell\right),$$

and $f_L(x) = a + u^T x$.

- Empirically, these have very impressive representational capacity and amenability to optimization.
- Today: inspired by Peyré [2025]'s survey, consider networks as Monte Carlo approximations of a function.

Two-layer network

- We focus on the case $L = 2$, the two-layer ReLU network, which can be written as

$$f_n(x) = a + \frac{1}{n}\sum_{i=1}^{n} u_i \varsigma\left(\langle x, v_i\rangle + b_i\right), \qquad x \in \mathbb{R}^d,$$

  where the parameters of the function are $(a, v_{1:n}, b_{1:n}, u_{1:n})$ with $a, b, u_i \in \mathbb{R}$ and $v_i \in \mathbb{R}^d$.

- Idea of Barron [1993]: represent

$$f(x) = a + \int_Z u(v,s)\varsigma\left(\langle x, v\rangle + b(v,s)\right)\nu(\mathrm{d}v, \mathrm{d}s), \qquad x \in \mathbb{R}^d.$$

  The variable x appears only in the inner product with v.

- Think of $\phi(x; v, s) = u(v,s)\varsigma\left(\langle x, v\rangle + b(v,s)\right)$ as a neuron.

- f is represented by an average of infinitely many neurons.

Random two-layer networks

- We can construct a random n neuron network by sampling $Z_i = (V_i, S_i) \sim \pi \gg \nu$ IID,

$$f_n(x) = a + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{d}\nu}{\mathrm{d}\pi}(Z_i)\phi(x; Z_i)\,.$$

- We are interested in functions with small $L^2(\mathcal{D}) = L^2(\mathbb{R}^d, \mathcal{D})$ error where $\mathcal{D}$ is a data distribution.

Lemma. *The prob. measure π minimizing $\mathbb{E}_\pi\left[\|f_n - f\|^2_{L^2(\mathcal{D})}\right]$ satisfies*

$$\mathbb{E}_{\pi^\star}\left[\|f_n - f\|^2_{L^2(\mathcal{D})}\right] = \frac{1}{n}\left\{\left[\int_Z \nu(\mathrm{d}z)\,\|\phi(\cdot; z)\|_{L^2(\mathcal{D})}\right]^2 - \|f - a\|^2_{L^2(\mathcal{D})}\right\}.$$

Hence, there exists an $f_n^\star$ such that

$$\|f_n^\star - f\|_{L^2(\mathcal{D})} \leq \frac{1}{\sqrt{n}} \int_Z \nu(\mathrm{d}z)\,\|\phi(\cdot; z)\|_{L^2(\mathcal{D})}\,.$$

Comments on $\pi^\star$

- This is analogous to the derivation of the optimal importance sampling distribution.
- We have

$$\pi^\star(\mathrm{d}z) = \nu(\mathrm{d}z)\frac{\|\phi(\cdot;z)\|_{L^2(\mathcal{D})}}{\int \|\phi(\cdot;z')\|_{L^2(\mathcal{D})}\,\nu(\mathrm{d}z')}.$$

- For an optimized π, every neuron has the same $L^2(\mathcal{D})$ norm,

$$\|\phi_{\pi^\star}(\cdot;z)\|_{L^2(\mathcal{D})} = \int_Z \nu(\mathrm{d}z)\,\|\phi(\cdot;z)\|_{L^2(\mathcal{D})}$$

  for ν-almost all z, where

$$\phi_\pi(\cdot;z) = \frac{\mathrm{d}\nu}{\mathrm{d}\pi}(z)\phi\,(\cdot;z)\,.$$

Finding a representation

- If we have a ReLU representation (ν, a, b, u), we can "optimize" it to get a bound on $\|f_n^\star - f\|_{L^2(\mathcal{D})}$.
- All that remains is to find a ReLU representation...
- Barron considered the Fourier transform, with $\tilde{f} : \mathbb{R}^d \to \mathbb{C}$,

$$f(x) = \int_{\mathbb{R}^d} \exp\left(i \langle v, x \rangle\right) \tilde{f}(v) \mathrm{d}v,$$

  from which one can proceed by defining an appropriate change of measure.
- The class of functions is quite rich, at least if we are interested only in f in a ball.
- It is hard to understand how any specific function is represented: even polynomials need to be modified.

Outline

Functions we consider

- Let $\mathcal{F}$ be the class of functions f such that

$$f(x) = \sum_{\alpha \in \mathbb{N}_0^d} c_\alpha x^\alpha, \qquad \|f\|_{\mathcal{F}}^2 = \sum_\alpha c_\alpha^2 \alpha! < \infty.$$

- For example, this could be from a Taylor expansion

$$f = \sum_{\alpha : |\alpha|} \frac{\partial^\alpha f(0)}{\alpha!} x^\alpha.$$

- Or from some other analytic approximation of a target φ that is not necessarily smooth.
- We define P to be the operator (H_α are Hermite polynomials),

$$Pf = \sum_\alpha c_\alpha H_\alpha, \qquad \text{where} \qquad f(x) = \sum_\alpha c_\alpha x^\alpha.$$

- P is an isometric isomorphism from $\mathcal{F}$ to $L^2(\gamma)$, $\gamma = N(0, I_d)$.

Heat equation

- In fact, we can write $P = \exp(-\frac{1}{2}\Delta)$.
- P is the inverse of the Markov operator $\exp\left(\frac{1}{2}\Delta\right)$ corresponding to

$$\exp\left(\frac{1}{2}\Delta\right) g(x) = \int g(y) N(y; x, I_d) \mathrm{d}y.$$

- If $f \in \mathcal{F}$ then $f^{\#} = Pf \in L^2(\gamma)$ is well-defined and

$$\mathbb{E}\left[f^{\#}(x + Z)\right] = f(x), \qquad Z \sim \gamma.$$

- $f^{\#}$ is sharpened so that smoothing it gives f.

- From the properties of P alone, we can verify some interesting representations!

Smoothing property and exponential kernels

Lemma. *Let $f \in \mathcal{F}$. Then $f^{\#} = Pf \in L^2(\gamma)$ and*

$$f(x) = \exp\left(-\frac{1}{2}\|x\|^2\right) \int_{\mathbb{R}^d} f^{\#}(v) \exp\left(\langle x, v\rangle\right) \gamma(\mathrm{d}v).$$

Proof. From the smoothing property,

$$\begin{aligned} f(x) &= \int_{\mathbb{R}^d} f^{\#}(y) N(y; x, I_d) \mathrm{d}y \\ &= \int_{\mathbb{R}^d} f^{\#}(y) \exp\left(\langle x, y\rangle - \frac{1}{2}\|x\|^2\right) N(y; 0, I_d) \mathrm{d}y \\ &= \exp\left(-\frac{1}{2}\|x\|^2\right) \int_{\mathbb{R}^d} f^{\#}(v) \exp\left(\langle x, v\rangle\right) \gamma(\mathrm{d}v). \end{aligned}$$

□

- $\mathcal{F}$ is the RKHS associated with $K(x, y) = \exp\left(\langle x, y\rangle\right)$, and

$$f(x) = \langle f, K_x\rangle_{\mathcal{F}} = \langle Pf, PK_x\rangle_{L^2(\gamma)} = \left\langle f^{\#}, PK_x\right\rangle_{L^2(\gamma)}.$$

A simple representation

Theorem. *Let $f \in \mathcal{F}$, $f^{\#} = Pf$. Then*

$$f(x) = \exp\left(-\frac{1}{2}\|x\|^2\right) \int_{\mathbb{R}^d} \int_{\mathbb{R}} f^{\#}(v) \exp(s) \varsigma\left(\langle x, v\rangle - s\right) \mathrm{d}s\gamma(\mathrm{d}v).$$

Proof. Combine the Lemma with

$$\int_{\mathbb{R}} \varsigma\left(t - s\right) \exp(s)\mathrm{d}s = \int_{-\infty}^{t} \left(t - s\right) \exp(s)\mathrm{d}s = \exp\left(t\right),$$

and take $t = \langle x, v\rangle$. □

- We use a standard RKHS, but approximate the kernel using ς.
- $\nu = \gamma \otimes \mathrm{Leb}(\mathbb{R})$ is universal; dependence on f in the "base" representation is $u(v, s) = f^{\#}(v) \exp(s)$.
- Not quite an integral rep. due to $G(x) = \exp(-\frac{1}{2}\|x\|^2)$.
  - This is a fixed function independent of f, however.

Quantitative bound

Theorem. *Let $\mathcal{D}$ be a distribution, sub-Gaussian with mean 0 and variance proxy $\sigma_X^2 < 1/2$. Let $f \in \mathcal{F}$. Then there exists a two-layer ReLU network q_n such that $f_n = G \cdot q_n$ satisfies*

$$\begin{aligned}\|f - f_n\|_{L^2(\mathcal{D})} &\leq \frac{1}{\sqrt{n}} \left\{ C(d, \sigma_X)^2 \|f\|_{\mathcal{F}}^2 - \|f\|_{L^2(\mathcal{D})}^2 \right\}^{1/2} \\ &\leq \frac{1}{\sqrt{n}} C(d, \sigma_X) \|f\|_{\mathcal{F}} .\end{aligned}$$

If $\sigma_X = d^{-1/2}$, $d \geq 3$, then $C(d, \sigma_X) \leq 2e\sqrt{2\pi} \cdot \exp\left(\frac{6}{d-2}\right)$.

- If we used an $\exp(\langle x, v \rangle)$-network, the error bound would be $C(d, \sigma_X) \leq e^{1/2} \exp\left(\frac{1/2}{(d-2)}\right)$: not much smaller!
- The main difficulty in high dimensions is neurons to represent the many directions v.

Outline

Representation without G

Theorem. *Let $f = \sum_{k \geq 0} f_k$ where each $f_k \in \mathcal{F}$ is a homogeneous, degree k polynomial. Then*

$$f(x) = f(0) + \int_{\mathbb{R}^d} \int_{\mathbb{R}^+} u(v,s) \varsigma\left(\langle x, v \rangle - s\right) \varrho(\mathrm{d}s) \gamma(\mathrm{d}v),$$

where $\varrho = \delta_0 + \mathrm{Leb}(\mathbb{R}_+)$, $\gamma = N(0, I_d)$ and

$$u(v,s) = \begin{cases} 2 f_1^{\#}(v) & s = 0 \\ 2 \sum_{k \geq 2} f_k^{\#}(v) \frac{s^{k-2}}{(k-2)!} & s > 0, \end{cases}$$

where $f_k^{\#} = P f_k$ for each $k \in \mathbb{N}_0$.

Proof I

- The Lemma says

$$\exp\left(\frac{1}{2}\|x\|^2\right) f_k(x) = \int_{\mathbb{R}^d} f_k^{\#}(v) \exp\left(\langle x, v\rangle\right) \gamma(\mathrm{d}v),$$

  which can be written as $\sum_{i=0}^{\infty} L_i(x) = \sum_{i=0}^{\infty} R_i(x)$, where

$$L_i(x) = f_k(x) \frac{\left\{\frac{1}{2}\|x\|^2\right\}^i}{i!}, \qquad R_i(x) = \int_{\mathbb{R}^d} f_k^{\#}(v) \frac{\langle x, v\rangle^i}{i!} \gamma(\mathrm{d}v).$$

- L_i and R_i are homogeneous polynomials of degree $k + 2i$ and degree i, respectively.
- Homogeneous polynomials of different degrees are linearly independent, so $L_i = R_{k+2i}$.
- In particular

$$f_k = L_0 = R_k = \int_{\mathbb{R}^d} f_k^{\#}(v) \frac{\langle x, v\rangle^k}{k!} \gamma(\mathrm{d}v).$$

Proof II

- Now observe that if $t \geq 0$, $k \geq 2$,

$$\int_0^\infty \frac{s^{k-2}}{(k-2)!} \varsigma(t-s)\,\mathrm{d}s = \int_0^t \frac{s^{k-2}}{(k-2)!}(t-s)\,\mathrm{d}s = \frac{t^k}{k!}.$$

- Using symmetry properties of $f_k^{\#}$ and γ, we can deduce that

$$f_k(x) = \int_{\mathbb{R}^d} f_k^{\#}(v)\frac{\langle x,v\rangle^k}{k!}\gamma(\mathrm{d}v) = 2\int f_k^{\#}(v)\frac{\langle x,v\rangle^k}{k!}1_{\langle x,v\rangle>0}\gamma(\mathrm{d}v),$$

and so for $k \geq 2$,

$$f_k(x) = 2\int_{\mathbb{R}^d} f_k^{\#}(v)\int_{\mathbb{R}_+}\frac{s^{k-2}}{(k-2)!}\varsigma(\langle x,v\rangle - s)\,\mathrm{d}s\gamma(\mathrm{d}v).$$

For $k=1$, we simply have $\varsigma(\langle x,v\rangle) = \langle x,v\rangle 1_{\langle x,v\rangle>0}$, and we can deduce

$$2\int_{\mathbb{R}^d} f_1^{\#}(v)\varsigma(\langle x,v\rangle)\gamma(\mathrm{d}v) = f_1(x).$$

RKHS perspective

- $\mathcal{F}_k$ is the RKHS with kernel $K^{(k)}(x,y) = \frac{1}{k!}\langle x, y\rangle^k$, $\|f_k\|_{\mathcal{F}_k} = \|f_k\|_{\mathcal{F}}$.
- This is a space of homogeneous degree k polynomials with

$$\|f_k\|^2_{\mathcal{F}_k} = \sum_{|\alpha|=k} c_\alpha^2 \alpha!, \qquad f_k = \sum_{|\alpha|=k} c_\alpha x^\alpha.$$

- As before the representation also corresponds to the reproducing property

$$f_k(x) = \left\langle f_k, K_x^{(k)} \right\rangle = \left\langle Pf_k, PK_x^{(k)} \right\rangle_{L^2(\gamma)} = \left\langle f_k^{\#}, PK_x^{(k)} \right\rangle_{L^2(\gamma)},$$

RKHS perspective

- $\mathcal{F}_k$ is the RKHS with kernel $K^{(k)}(x,y) = \frac{1}{k!}\langle x,y\rangle^k$, $\|f_k\|_{\mathcal{F}_k} = \|f_k\|_{\mathcal{F}}$.
- This is a space of homogeneous degree k polynomials with

$$\|f_k\|^2_{\mathcal{F}_k} = \sum_{|\alpha|=k} c_\alpha^2 \alpha!, \qquad f_k = \sum_{|\alpha|=k} c_\alpha x^\alpha.$$

- As before the representation also corresponds to the reproducing property

$$f_k(x) = \left\langle f_k, K_x^{(k)} \right\rangle = \left\langle Pf_k, PK_x^{(k)} \right\rangle_{L^2(\gamma)} = \left\langle f_k^{\#}, PK_x^{(k)} \right\rangle_{L^2(\gamma)},$$

- In the homogeneous setting, we have

$$\left\langle f_k^{\#}, PK_x^{(k)} \right\rangle_{L^2(\gamma)} = \left\langle f_k^{\#}, K_x^{(k)} \right\rangle_{L^2(\gamma)},$$

even though $PK_x^{(k)} \neq K_x^{(k)}$. So we use the ς-integral representation of $K_x^{(k)}$.

Quantitative bound

Theorem. *Let $\mathcal{D}$ be a distribution, sub-Gaussian with mean 0 and variance proxy $\sigma_X^2 < 1/2$. Let $f_k \in \mathcal{F}_k$ and $f = \sum_{k \geq 0} f_k$. Then there exists a two-layer ReLU network f_n such that*

$$\|f - f_n\|_{L^2(\mathcal{D})} \leq \frac{1}{\sqrt{n}} \sum_{k \geq 1} C_{d,k} \|f_k\|_{\mathcal{F}} .$$

If $\sigma_X = d^{-1/2}$ then $C_{d,1} \leq 2^{3/2}$ and $C_{d,k} \leq \sqrt{2k} \left(2e \cdot \frac{d+k-1}{dk}\right)^{k/2}$.

- Here the triangle inequality split the $\|f_k\|_{\mathcal{F}}$, but $C_{d,k} \to 0$ as $k \to \infty$ quickly for $d \geq 6$.

Outline

How did we get here?

- The two representations considered so far are very simple.
  - Validity could be proved in an undergraduate course!
- Good approximation properties for $\mathcal{D}$ that are concentrated on the unit sphere.
  - Or functions that scale inputs by $d^{-1/2}$ or more.
- In fact, there are a huge number of (often complicated) integral representations.

Fundamental Theorem of Calculus

- The second representation actually arose from studying

$$T_\mu g(x) = \int g\left(\langle x, v\rangle\, v\right) \mu(\mathrm{d}v),$$

  for spherically symmetric μ.
- Why? Because if $f = T_\mu g$ then

$$f(x) = \int_{\mathbb{R}^d} g\left(\langle x, v\rangle\, v\right) \mu(\mathrm{d}v) = 2 \int_{\mathbb{R}^d} g\left(\langle x, v\rangle\, v\right) 1_{\langle x,v\rangle > 0} \mu(\mathrm{d}v),$$

  and for $t \geq 0$, [inspired by Telgarsky, 2023]

$$g\left(tv\right) = g_v(0) + \varsigma\left(t\right) g_v'(0) + \int_0^\infty \varsigma\left(t - s\right) g_v''(s) \mathrm{d}s.$$

- Here $g_v(s) = g(sv)$, and we are integrating along lines.

Solutions of $f = T_\mu g$

- In fact, if μ is spherically symmetric, T_μ is invertible.
- The unique solution $g = T_\mu^{-1} f$ is not necessarily nice, however.
- We have

$$T_\mu^{-1} f = \sum_{k \geq 0} \sum_{j=0}^{\lfloor k/2 \rfloor} \lambda_{d,k-2j,j,\mu}^{-1} f_{kj},$$

  where $f = \sum f_{kj}$ is the harmonic decomposition of f.
- If f is not harmonic, the eigenvalues do not behave nicely with dimension.
- We also find that in high dimensions, $\mu = \gamma$ is a good idea!

Getting a nicer "solution"

- The solution is unique for a given polynomial $f : \mathbb{R}^d \to \mathbb{R}$.
- We would prefer harmonic functions.
- Obvious idea: extend f to $\mathcal{H}(f) : \mathbb{R}^{d+1} \to \mathbb{R}$ such that $\mathcal{H}(f)$ is harmonic and $\mathcal{H}(f)(x, 0) = f(x)$.
- There is an extension that does exactly this! We obtain

$$\mathcal{H}(f) = \sum_{k \geq 0} h_k, \qquad T_\mu^{-1} \mathcal{H}(f) = \sum_{k \geq 0} \frac{1}{k!} h_k.$$

- Our integral representation for f_k is,

$$f_k(x) = \int_{\mathbb{R}^{d+1}} h_k(v, w) \frac{\langle (x, 0), (v, w) \rangle^k}{k!} \gamma(\mathrm{d}v, \mathrm{d}w),$$

  and we find $\int_{\mathbb{R}} h_k(v, w) \gamma(\mathrm{d}w) = f_k^{\#}(v)$.
- Note that $g_v(s) = h_k(sv, sw) = s^k h_k(v, w)$ gives $g_v''(s) = k(k-1)s^{k-2} h_k(v, w)$.

Transformations

- So we obtained the second representation by thinking about integrating along lines from 0.
- We improved the most obvious representation by harmonic extension and integration of the extra variable.
- Transformations more generally: we could use any function $\tilde{f}$ such that
$$f(x) = \tilde{f}(A(x - x_0)).$$
- Harmonic extension corresponds to A being $(d + 1) \times d$, $x_0 = 0$, $\tilde{f} = \mathcal{H}(f)$.
- More obvious things would be just taking A invertible and x_0 a shift.
- It's not obvious what the "right" transformations are.

Transformations

- After a transformation, we have

$$f(x) = f(x_0) + \int_{\mathbb{R}^{d'}} \int_{\mathbb{R}} u(v,s) \varsigma\left(\left\langle x, A^T v \right\rangle - \left\langle x_0, A^T v \right\rangle - s\right) \nu(\mathrm{d}v, \mathrm{d}s),$$

  where (ν, u) are from the integral representation of $\tilde{f}$.
- Biases can look less regular, centering has an effect.
- Even without transformations, the optimized distribution for v is not typically Gaussian.

Outline

Comments

- The representations are surprisingly simple and accessible.
- Is $\mathcal{F}$ the right function class?
  - It seems that the local nature does correspond to what a linear combination of ReLUs could approximate.
  - One can approximate non-smooth functions on a compact set using analytic functions.
  - $\mathcal{F}$ does not constrain the set of compactly support functions that much.
- Do the representations reflect features of trained networks?
- Can we use them to analyze deep networks?
  - A deep network can indeed be viewed as a composition of two-layer networks.
  - But what are the intermediate functions!?

Different activations

- For the exponential kernel approach, one can use any activation σ such that

$$Z := \int_{\mathbb{R}} \sigma(u) \exp(-u) \mathrm{d}u \in (0, \infty),$$

since then

$$\int_{\mathbb{R}} \sigma(z-s) \frac{\exp(s)}{Z} \mathrm{d}s = \exp(z).$$

- This cannot hold for σ a polynomial or sigmoid.
- For the homogeneous polynomial approach, we can use any activation such that there is a representation

$$\int_{\mathbb{R}} \sigma(z-s) \varrho(\mathrm{d}s) = z^k, \qquad z \geq 0.$$

There are many but these are less easy to invert to obtain ϱ.

Related work

- There are representations starting with the Fourier transform.
- At least pedagogically, but perhaps generally, these simpler representations seem useful.
- There are also papers introducing kernels involving ς.
  - Here we use the exponential / homogeneous polynomial kernel.
  - The ReLU function enters as part of the Monte Carlo approximation of the infinite-neuron, exponential/polynomial kernel network.
- There are deep network constructions, but usually involving careful assembly of sparse networks.
  - Exponential error decay, but do these reflect trained networks? Can we do better?
- What about mean-field optimization? [Chizat and Bach, 2018, Mei et al., 2019]

References I

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Y. Cho and L. Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

D. Hsu, C. H. Sanford, R. Servedio, and E. V. Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random ReLUs. In *Conference on Learning Theory*, pages 2423–2461. PMLR, 2021.

Z. Ji, M. Telgarsky, and R. Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*, 2019.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

References II

A. Petrosyan, A. Dereventsov, and C. G. Webster. Neural network integral representations with the ReLU activation function. In *Mathematical and Scientific Machine Learning*, pages 128–143. PMLR, 2020.

G. Peyré. The mathematics of artificial intelligence. *arXiv preprint arXiv:2501.10465*, 2025.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.

S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

M. Telgarsky. *Deep learning theory*. 2023. URL `https://mjt.cs.illinois.edu/dlt/two.pdf`. Draft compiled September 25 2023.