

Expressivity

We will start by briefly discussing *expressivity* of neural networks. The question of expressivity asks: given a distribution $\mathcal{D}$ over inputs x , a function class $\mathcal{F}$, for example the set of two layer neural networks of width m , and a target function f^* , when does there exist $\hat{f} \in \mathcal{F}$ such that

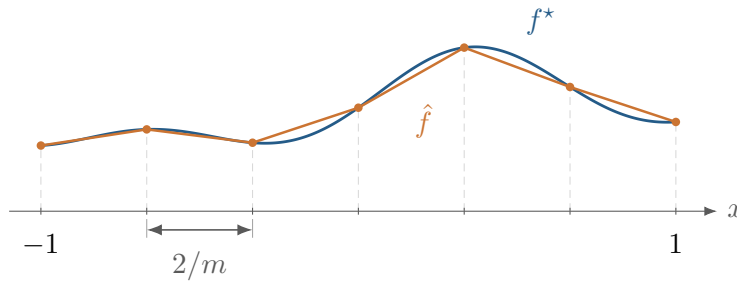
$$\mathbb{E}_x[(\hat{f}(x) - f^*(x))^2] \leq \epsilon?$$

We will focus on the function class $\mathcal{F}_m^\sigma$ of two layer neural networks of width m and activation σ :

$$\mathcal{F}_m^\sigma = \left\{ x \mapsto \sum_{j=1}^m a_j \sigma(w_j \cdot x + b_j) + c \right\}, \quad \text{where } \sigma : \mathbb{R} \rightarrow \mathbb{R}.$$

While many of the results can be extended to arbitrary activations, we will focus for concreteness on $\sigma(z) = \text{ReLU}(z) := \max(0, z)$.

We will begin by discussing the 1D setting. Note that a ReLU network is piecewise linear, and every neuron adds at most one kink. Therefore, $\mathcal{F}_m^\sigma$ is piecewise linear with at most $m + 1$ components. The natural way to approximate a univariate target function f^* is therefore linear interpolation:



For simplicity, let us restrict f^* to $[-1, 1]$ and split the interval into m segments of length $2/m$. The following lemma bounds the error from approximating f^* by a line on each segment:

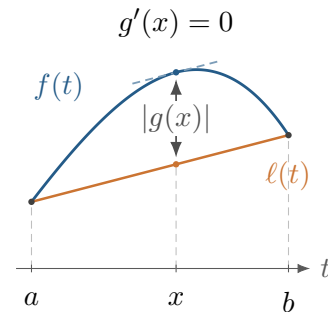
Lemma 1. Let $f \in C^2([a, b])$ and $\ell(x) := f(a) + (x - a) \frac{f(b) - f(a)}{b - a}$ be the secant line. Then

$$\max_{x \in [a, b]} |f(x) - \ell(x)| \leq \frac{(b - a)^2}{8} \|f''\|_\infty.$$

Proof. Let $g = f - \ell$ and let x be a maximizer of $|g|$. Without loss of generality let x be closer to a than b . Because x is either a minimizer or a maximizer of g , $g'(x) = 0$, so Taylor's theorem gives:

$$0 = g'(x) = g'(x) \frac{(a - x)}{(a - x)} + \frac{1}{2} g''(\xi)(a - x)^2.$$

and rearranging and using $(a - x)^2 \leq \frac{(b - a)^2}{4}$ completes the proof. $\square$



These are informal course notes and have not undergone peer review. AI tools assisted with simplifying proofs, editing, and typesetting. I take responsibility for the content, and corrections are welcome.

Because each interval has length $2/m$, [Lemma 1](#) implies the following corollary.

Corollary 1. *If $f \in C^2([-1, 1])$, $\inf_{\hat{f} \in \mathcal{F}_m^\sigma} \|\hat{f} - f\|_\infty \leq \frac{\|f''\|_\infty}{2m^2}$.¹*

What about higher dimensions? In 1D, we used the m neurons to create a grid over the space which gave the $1/m^2$ rate. However, in order to create an ϵ net in high dimensions, m has to grow exponentially large in the dimension, so we will need a new idea.

We expect the error to vanish as $m \rightarrow \infty$, which means that it should be possible to represent f using “infinitely many neurons.” Explicitly, f should admit a decomposition of the form

$$f(x) = \mathbb{E}_{(a,w,b) \sim \mu} [a\sigma(w \cdot x + b)],$$

where μ is a distribution over neurons which characterizes this infinite-width limit.

Once we have such an “infinite width” decomposition of f , we can try to sparsify this infinite width network to get a good finite width approximation by sub-sampling neurons:

$$f(x) = \mathbb{E}_{(a,w,b) \sim \mu} [a\sigma(w \cdot x + b)] \xrightarrow{\text{sub-sampling}} f_m(x) = \frac{1}{m} \sum_{j=1}^m a_j \sigma(w_j \cdot x + b_j).$$

This is formalized by Maurey’s lemma. To simplify notation, we will use $\phi_j(x) = a_j \sigma(w_j \cdot x + b_j)$ to denote the j th neuron and $\phi \sim \mu$ to denote the random map $x \mapsto a\sigma(w \cdot x + b)$, where $(a, w, b) \sim \mu$.

Lemma 2 (Maurey; [Pisier, 1981](#)). *Let $f(x) = \mathbb{E}_{\phi \sim \mu} [\phi(x)]$. There exist m neurons $\phi_1, \dots, \phi_m$ whose average $f_m(x) := \frac{1}{m} \sum_{j=1}^m \phi_j(x)$ satisfies*

$$\mathbb{E}_x [(f_m(x) - f(x))^2] \leq \frac{\mathbb{E}_{x, \phi \sim \mu} [\phi(x)^2]}{m}$$

Proof. Sample $\phi_1, \dots, \phi_m \sim \mu$. Then,

$$\mathbb{E}_{\phi_1, \dots, \phi_m} [(f(x) - f_m(x))^2] = \frac{1}{m} \text{Var}_{\phi \sim \mu} [\phi(x)] \leq \frac{1}{m} \mathbb{E}_{\phi \sim \mu} [\phi(x)^2].$$

Taking expectations with respect to x and swapping the expectations gives:

$$\mathbb{E}_{\phi_1, \dots, \phi_m} \mathbb{E}_x [(f(x) - f_m(x))^2] \leq \frac{1}{m} \mathbb{E}_{x, \phi \sim \mu} [\phi(x)^2].$$

Because $\mathbb{E}_x [(f(x) - f_m(x))^2] \geq 0$ for any choice of neurons, there must exist $\phi_1, \dots, \phi_m$ such that this quantity is at most the right hand side. $\square$

This motivates the following variational norm, which minimizes the right hand side over all possible representations $f = \mathbb{E}_\mu \phi$:

$$\|f\|_{\mathcal{V}(\phi, \mathcal{D})}^2 := \inf_{\mu: f = \mathbb{E}_\mu \phi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{\phi \sim \mu} [\phi(x)^2].$$

¹At the cost of a more complex argument, it is possible to generalize this to $\frac{\|D^2 f\|_{\text{TV}}}{m^2}$ for functions with second-order bounded variation using adaptive breakpoints. This allows [Corollary 1](#) to directly apply to piecewise-linear functions.

Note that different representations of the particles ϕ (e.g. different activations) and different distributions $\mathcal{D}$ can give different variational norms. For notational simplicity, we will fix $\sigma = \text{ReLU}$ and an arbitrary distribution $\mathcal{D}$ over x and simply write $\|f\|_{\mathcal{V}}$.

Minimizing over representations $f = \mathbb{E}_{\phi \sim \mu} \phi$, [Lemma 2](#) gives the following corollary.

Corollary 2. *If $\|f\|_{\mathcal{V}} < \infty$, there exists a width m neural network f_m such that*

$$\mathbb{E}[(f_m - f)^2] \leq \frac{\|f\|_{\mathcal{V}}^2}{m}.$$

Note. While the L^2 error decays like $m^{-1/2}$, there may still be terrible dependencies on the dimension of x in the variational norm $\|f\|_{\mathcal{V}}$. For some target functions this can be exponentially large in the dimension, or even infinite.

We will need the triangle inequality for the variational norm:

Lemma 3. *If $\{\phi\}$ is closed under scaling, $\|f + g\|_{\mathcal{V}} \leq \|f\|_{\mathcal{V}} + \|g\|_{\mathcal{V}}$. More generally,*

$$\left\| \int a(t) h_t dt \right\|_{\mathcal{V}} \leq \int |a(t)| \|h_t\|_{\mathcal{V}} dt.$$

Proof. Let $f = \mathbb{E}_{\mu} \phi$ and $g = \mathbb{E}_{\nu} \phi$ with representation costs A, B , respectively. Then if $p + q = 1$,

$$f + g = p \mathbb{E}_{\mu} [\phi/p] + q \mathbb{E}_{\nu} [\phi/q]$$

is a representation of $f + g$ with squared representation cost $\frac{A^2}{p} + \frac{B^2}{q}$. Setting $p = \frac{A}{A+B}$ and taking the infimum over valid μ, ν completes the proof. The same argument proves the integral version. $\square$

Since the variational norm is defined as an infimum over distributions μ , it is often difficult to compute exactly, but we can upper bound it. We will first need the following 1D lemma:

Lemma 4. *If $g \in C^2([-B, B])$ then $\|g\|_{\mathcal{V}} \lesssim |g(0)| + B|g'(0)| + B^2\|g''\|_{\infty}$.*

Proof. For $x \geq 0$, Taylor's theorem with integral remainder gives

$$g(x) - g(0) - g'(0)x = \int_0^x g''(t)(x-t) dt = \int_0^B g''(t)\sigma(x-t) dt.$$

The last equality holds because $\sigma(x-t) = x-t$ for $t \leq x$ and is zero for $t > x$. Applying the same argument to $g(-x)$ handles $x \leq 0$. Each ReLU integral vanishes on the opposite half of the interval, so for all $x \in [-B, B]$,

$$g(x) = g(0) + g'(0)x + \int_0^B g''(t)\sigma(x-t) dt + \int_{-B}^0 g''(t)\sigma(t-x) dt.$$

Since $1 = \sigma(1)$ and $x = \sigma(x) - \sigma(-x)$, the first two terms have costs at most $|g(0)|$ and $2B|g'(0)|$.

Each neuron $\sigma(x - t)$, $\sigma(t - x)$ appearing above takes values in $[0, B]$ on $[-B, B]$, so has cost at most B . By [Lemma 3](#), the last two terms therefore have total cost at most

$$B \int_{-B}^B |g''(t)| dt \leq 2B^2 \|g''\|_\infty.$$

Combining these inequalities gives the final bound. $\square$

We can now prove Barron's theorem for ReLU networks, which states that any function whose Fourier transform decays sufficiently quickly has a finite variational norm:

Theorem 1 ([Barron, 1993](#); [Klusowski and Barron, 2018](#)). *If $\mathcal{D}$ is supported on the unit ball,*

$$\|f\|_{\mathcal{V}} \lesssim \int (1 + \|\omega\|^2) |\hat{f}(\omega)| d\omega.$$

Proof. By the Fourier inversion formula,

$$\begin{aligned} f(x) &= \int \hat{f}(\omega) e^{i\omega \cdot x} d\omega && \text{Fourier inversion} \\ &= \int |\hat{f}(\omega)| e^{i(\omega \cdot x + \theta_\omega)} d\omega && \hat{f}(\omega) = |\hat{f}(\omega)| e^{i\theta_\omega} \\ &= \int |\hat{f}(\omega)| \cos(\omega \cdot x + \theta_\omega) d\omega && \text{Taking real parts} \end{aligned}$$

Note that this is already an infinite width neural network with cosine activations! Indeed, by [Lemma 3](#), it suffices to bound the cost of representing each cosine neuron with ReLU neurons:

$$\|f\|_{\mathcal{V}} \leq \int |\hat{f}(\omega)| \cdot \underbrace{\|x \mapsto \cos(\omega \cdot x + \theta_\omega)\|_{\mathcal{V}}}_{\text{cost of approximating cosine with ReLU}} d\omega.$$

As $\|\omega\|$ grows, the cosine oscillates faster, so a ReLU network needs more width to approximate it. To make this precise, apply [Lemma 4](#) to $g(z) = \cos(z + \theta_\omega)$ with $z = \omega \cdot x$ and note that $|z| \leq \|\omega\|$:

$$\|x \mapsto \cos(\omega \cdot x + \theta_\omega)\|_{\mathcal{V}} \lesssim 1 + \|\omega\| + \|\omega\|^2 \lesssim 1 + \|\omega\|^2.$$

Plugging this into the above triangle inequality finishes the proof. $\square$

Finally, while the proof is beyond the scope of this course, it's worth mentioning that there are many functions that cannot be easily expressed by 2 layer neural networks but that can be expressed by deeper neural networks. The following theorem, proved in [Appendix A](#), is an extension of results by [Eldan and Shamir \(2016\)](#), [Daniely \(2017\)](#), [Safran et al. \(2019\)](#), and [Safran and Lee \(2022\)](#):

Theorem 2. *Let $d \geq 2$, $x \sim \mathcal{N}(0, I_d)$ and $f^*(x) = \mathbf{1}_{\{\|x\|^2 \geq d\}}$. Then for any activation σ ,*

$$\inf_{f_m \in \mathcal{F}_m^\sigma} \|f^* - f_m\|_2 \geq \frac{c}{1 + (\log_d m)^{1/4}} \quad \text{for } m \leq e^{d^{0.99}}.$$

However, there is a depth 3 neural network with widths $(d, 1)$ that expresses it exactly:

$$x \xrightarrow{z \mapsto z^2} (x_1^2, \dots, x_d^2) \xrightarrow{\text{sum and threshold}} \mathbf{1}_{\{\|x\|^2 \geq d\}}.$$

Transformer Expressivity

There is a nice line of results that studies expressivity of Transformers through a very different lens. At a high level, a transformer takes in a sequence and then iteratively applies attention layers, which move information from one part of the sequence to another, and MLP layers, which independently process each token in the sequence in parallel. We can write a transformer layer as:

$$x_i \leftarrow \text{MLP}\left(x_i, \underbrace{\sum_j \alpha_{ij} x_j}_{\text{attention}}\right), \quad \text{where} \quad \underbrace{\alpha_{ij} = \frac{s(x_i, x_j)}{\sum_k s(x_i, x_k)}}_{\text{normalized weights}}.$$

where x_i is the token at position i in the sequence. Standard softmax attention corresponds to $s(x_i, x_j) = \exp(q(x_i) \cdot k(x_j))$ where $q(x_i), k(x_j)$ are the queries and keys.

For the purpose of the upper bound, we can allow MLP to be an arbitrary function since this only makes the Transformer more expressive. The bottleneck becomes moving information between tokens using the attention mechanism. Two common restrictions on attention for theory are:

- **Hardmax:** each token only attends to a single token, i.e. $\alpha_{i,:}$ is one-hot
- **Softmax:** keep full attention, but all computations are done with $\log(T)$ bits of precision where T is the sequence length

The following results show that constant depth Transformers are no more expressive than bounded depth circuits. Thus, while they may excel on problems that reward parallel computation, they cannot express functions that require deeper circuits to compute.

Theorem 3 (Informal). *If $\mathcal{F}$ is the set of Transformers with constant width and depth:*

- *With hardmax attention, $\mathcal{F} \subseteq \text{AC}^0$ (Hao et al., 2022).*
- *With softmax and $O(\log T)$ -bit precision, $\mathcal{F} \subseteq \text{TC}^0$ (Merrill and Sabharwal, 2023).*

Here AC^0 consists of functions computable by constant depth Boolean circuits with polynomial size and unbounded fan-in using only AND, OR, and NOT gates. TC^0 additionally allows majority gates. This distinction matters even for simple problems. For example, computing $\prod_i x_i$ from $x \in \{\pm 1\}^T$ (PARITY) can be done in TC^0 but not AC^0 (Furst et al., 1984). Similarly, graph connectivity can be computed in polynomial time, but is conjectured not to be computable in TC^0 .

A Proof of the depth separation theorem

Note. This proof is beyond the scope of the course and is included for completeness.

Proof of Theorem 2. We first handle the case of sufficiently large d . Let F_k, H_k be the normalized Hermite tensors of f^* and f_m . By Hermite Parseval (O'Donnell, 2014, Proposition 11.36),

$$\|f^* - f_m\|_2^2 = \sum_{k \geq 0} \|F_k - H_k\|_F^2.$$

Since f^* is radial, its odd coefficients vanish and F_{2k} is proportional to the unique rotation-invariant symmetric tensor $\mathbb{E}_{u \sim \text{Unif}(S^{d-1})}[u^{\otimes 2k}]$. Each neuron of f_m is a ridge function $x \mapsto \psi(u \cdot x)$, with rank-one Hermite tensors $\beta_k u^{\otimes k}$, so H_{2k} has symmetric tensor rank at most m . Intuitively, this rotation-invariant symmetric tensor should be hard to approximate by a rank m tensor for reasonably large k . Indeed, if $\text{rank}(B)$ denotes symmetric tensor rank, showing that

$$\inf_{\text{rank}(B) \leq m} \|F_{2k} - B\|_F^2 \geq \frac{1}{2} \|F_{2k}\|_F^2 \quad \text{for all } k \geq K \quad (\text{A.1})$$

would give, by the Hermite-tail theorem of Kindler et al. (2018) applied to $2f^* - 1$,

$$\|f^* - f_m\|_2^2 \geq \frac{1}{2} \sum_{k \geq K} \|F_{2k}\|_F^2 \gtrsim \frac{\text{Var}(f^*)}{\sqrt{K}} \gtrsim K^{-1/2}.$$

Thus it suffices to prove the tensor approximation bound (A.1) with $K \asymp 1 + \log_d m$.

Take $K = \lceil C(1 + \log_d m) \rceil$ for a sufficiently large universal constant C . The assumption $m \leq e^{d^{0.99}}$ ensures $K = o(d)$. For $k \geq K$, use the balanced flattening

$$A_k = \mathbb{E}_{u \sim \text{Unif}(S^{d-1})} [\text{Vec}(u^{\otimes k}) \text{Vec}(u^{\otimes k})^\top] \in \mathbb{R}^{d^k \times d^k}.$$

Write N_ℓ for harmonic dimensions and λ_ℓ for the eigenvalues of A_k at degrees $\ell = k, k-2, \dots$. Specializing Azevedo and Menegatto (2015, Section 2) to $(u \cdot v)^k$ gives

$$\frac{\lambda_{\ell+2}}{\lambda_\ell} = \frac{k - \ell}{d + k + \ell}, \quad p_\ell := \frac{N_\ell \lambda_\ell^2}{\|A_k\|_F^2}, \quad \sum_\ell p_\ell = 1.$$

For $\ell \leq K/16$, using $k \geq K$ and the dimension formula (Dai and Xu, 2013, Corollary 1.4) gives

$$\frac{p_{\ell+2}}{p_\ell} \geq \frac{d^2}{4(\ell+2)^2} \left(\frac{K - \ell}{d + K + \ell} \right)^2 \geq 4.$$

Summing this geometric bound gives $\sum_{\ell \geq s} p_\ell \geq 3/4$ for $s = \lfloor K/16 \rfloor$. Since N_ℓ increases with ℓ , each squared eigenvalue there is at most $\|A_k\|_F^2 / N_s$. Eckart–Young therefore gives

$$\inf_{\text{rank}(B) \leq m} \|A_k - B\|_F^2 \geq \left(\frac{3}{4} - \frac{m}{N_s} \right) \|A_k\|_F^2.$$

Choosing C sufficiently large gives $N_s \geq (d/(2s))^s \geq 4m$. Because a tensor of rank at most m has matrix rank at most m after flattening, this proves (A.1).

The result for the remaining bounded d can be absorbed into the universal constant so long as the approximation error is bounded away from zero. This is guaranteed by the fact that f^* has nonzero Hermite coefficients at arbitrarily high degrees whose balanced flattenings have arbitrarily high rank. The depth 3 construction is given in the theorem statement. $\square$

References

- D. Azevedo and V. A. Menegatto. Eigenvalues of dot-product kernels on the sphere. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 3(1), 2015. doi: 10.5540/03.2015.003.01.0039. URL <https://proceedings.sbmec.org.br/sbmec/article/view/398>.
- A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500. URL <https://doi.org/10.1109/18.256500>. The original result treats sigmoidal activations and a first Fourier moment condition.
- F. Dai and Y. Xu. Spherical harmonics, 2013. URL <https://arxiv.org/abs/1304.2585>.
- A. Daniely. Depth separation for neural networks. In *COLT*, pages 690–696, 2017. URL <https://proceedings.mlr.press/v65/daniely17a.html>.
- R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *COLT*, pages 907–940, 2016. URL <https://proceedings.mlr.press/v49/eldan16.html>.
- M. Furst, J. B. Saxe, and M. Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17:13–27, 1984. doi: 10.1007/BF01744431. URL <https://doi.org/10.1007/BF01744431>.
- Y. Hao, D. Angluin, and R. Frank. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810, 2022. URL <https://aclanthology.org/2022.tacl-1.46/>. Theorem 1.
- G. Kindler, N. Kirshner, and R. O’Donnell. Gaussian noise sensitivity and Fourier tails. *Israel Journal of Mathematics*, 225(1):71–109, 2018. doi: 10.1007/s11856-018-1646-8. URL <https://doi.org/10.1007/s11856-018-1646-8>.
- J. M. Klusowski and A. R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018. URL <https://arxiv.org/abs/1607.07819>. See Theorem 2 and its proof for the ReLU representation under a second Fourier moment condition. The formulation here is adapted to the unit Euclidean ball and our variational norm.
- W. Merrill and A. Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023. URL <https://aclanthology.org/2023.tacl-1.31/>. Theorem 1 and Corollary 1.1.
- R. O’Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. URL <https://www.cs.cmu.edu/~odonnell/papers/Analysis-of-Boolean-Functions-by-Ryan-ODonnell.pdf>.
- G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire d’analyse fonctionnelle, École Polytechnique*, (5):1–12, 1981. URL https://www.numdam.org/item/SAF_1980-1981____A5_0/. 1980–1981, Exposé no. 5. See Lemma 2 and its proof, pp. V.2–V.3, for Maurey’s empirical approximation argument.
- I. Safran and J. Lee. Optimization-based separations for neural networks. In *COLT*, pages 3–64, 2022. URL <https://proceedings.mlr.press/v178/safran22a.html>.
- I. Safran, R. Eldan, and O. Shamir. Depth separations in neural networks: What is actually being separated? In *COLT*, pages 2664–2666, 2019. URL <https://proceedings.mlr.press/v99/safran19a.html>. Full version: <https://arxiv.org/abs/1904.06984>.