3-layer networks, 2-layer networks

Administrative:

- Some project info:
  1. “Open problems” are good candidates for surveys.
  2. The “project proposal” will be done as follows.
     1. Everyone submits a list of three project ideas, along with at least one reference for each. Each idea shold be just a few sentences, and the whole thing should fit on one page. I’ll put an example on the schedule in a few days.
     2. I will skim the project ideas and then meet with everyone for 15 minutes, and help cook up a project.
  3. A proper project schedule and details for handin of this “milestone 0” will appear before class on 9/7.
- Please keep giving feedback.

Neural network conventions

The previous lecture gave an abstract treatment of neural networks. Today there will be two proofs given that show neural networks can fit continuous functions. the neural networks will have the following form.

(*Picture drawn in class.* Networks will be organized into layers. For convenience, layer 0 consists of the inputs themselves. In subsequent layers, each node will take the vector of outputs from the previous layer, y, form an affine combination $a^\top y + b$, and then apply a nonlinearity. For today, the convention is that all layers *except* 0 and the final layer will use a single nonlinearity σ; the last layer will have no nonlinearity, it will just form a linear combination.)

With this convention, networks may be written as a single expression as follows. The first layer (ignoring the input) first computes $A_1 x + b$, where $A_1 \in \mathbb{R}^{k_1 \times d}$ is a matrix and k_1 is the number of nodes in the first (non-input) layer. Letting $\vec{\sigma}$ denote a function which applies σ to each coordinate, the output of the first layer is $\vec{\sigma}(A_1 x + b)$. Continuing in this way, a network with 3 (non-input) layers computes

$$x \mapsto A_3 \vec{\sigma}(A_2 \vec{\sigma}(A_1 x + b_1) + b_2) + b_3.$$

In the subsequent theorems, let $\mathcal{N}_l$ denote the set of networks with l (non-input) layers, a single $\sigma : \mathbb{R} \to \mathbb{R}$ across the whole set, and any choices for the number of nodes per layer, as well as any choices for the *weights*, meaning the matrices $(A_1, \dots, A_l)$ and offset vectors $(b_1, \dots, b_l)$.

Remarks.

1. It may seem annoying that the output node does not apply σ. A homework problem will point out that this does not really change things.
2. It may seem useless to prove two theorems today, since the result for $\mathcal{N}_2$ implies the result for $\mathcal{N}_3$. Each proof will carry its own insight, however.

3-layer networks

Theorem. Fix nonlinearity $\sigma(z) := \max\{0, z\}$, the ReLU. Then

$$\sup_{g \text{ continuous}} \inf_{f \in \mathcal{N}_3} \|f - g\|_1 = 0.$$

Proof. *(Proof in class had many pictures.)* By the lemma discussed in the preceeding two lectures, it suffices to prove the following: given $\tau > 0$ (suppose $\tau \leq 1$ WLOG) and any rectangle $R := [a_i, b_i] \times \cdots \times [a_d, b_d]$, there exists $f_R \in \mathcal{N}_3$ with $\|f_R - \mathbf{1}_R\| \leq \tau$.

The idea of the proof is the following. It is easy to build a function in $\mathcal{N}_2$ which approximates 1_R in the univariate case. If we try combining these for each dimension, we don't get quite what we want, and we need to do some cleanup with another layer.

In more detail, fix a dimension $i \in \{1, \ldots, d\}$, and define

$$f_i(x) := \sigma\left(\frac{x_i - a_i}{\delta} + 1\right) - \sigma\left(\frac{x_i - a_i}{\delta}\right) - \sigma\left(\frac{x_i - b_i}{\delta}\right),$$

where $\delta := 20d\tau$ (chosen to make the end of the proof go through). Restricting attention to dimension i, f_i approximates the indicator on the interval $[a_i, b_i]$, except for the region beyond b_i which will not matter *(picture drawn in class where f_i is build in stages around $\mathbf{1}_{[a_i,b_i]}$)*. Next define

$$f_R := \sigma\left(\sum_i 2f_i - (2d - 1)\right).$$

(Picture drawn in $\mathbb{R}^2$ in class: without the outer σ, the function $\sum_i f_i$ is correct on R, but lots of slop elsewhere; adding a threshold and a ReLU cleans this up.) Note that

$$f_R(x) \begin{cases} = 1 & x \in R, \\ \in [0, 1) & x \notin R, \inf_{y \in R} \|x - y\|_\infty \leq \delta, \\ \in 0 & \text{otherwise}. \end{cases}$$

To calculate $\|f_R - \mathbf{1}_R\|_1$, it suffices to upper bound the volume of the region along the boundary of R which f_R maps within $[0, 1)$ as above. But this excess is no more than the excess given by the δ-neighborhood of $[0, 1]^d$, meaning

$$\|f_R - \mathbf{1}_R\|_1 \leq \text{vol}([-\delta, 1 + \delta]^d) - \text{vol}([0, 1]^d) = (1 + 2\delta)^d - 1 \leq \exp(2d\delta) - 1 = \exp(\tau/10) - 1 \leq \tau,$$

which used the tangent inequality $1 + x \leq e^x$, which holds for $x \in \mathbb{R}$, and the secant inequality $e^x \leq 1 + 10x$, which holds along $[0, 1]$. □

Remark (on inequalities)**.** The inequalities with $\exp(\cdot)$ may seem very loose near the end. A rule of thumb for inequalities is to make sure they are tight in the regime that matters the most; for instance $\exp(z)$ and $1 + z$ become arbitrarily close as $z \downarrow 0$.

2-layer networks

As such, it's very intuitive that 3 layers suffice. On the other hand, each layer seemed crucial in the preceeding construction, thus something interesting must happen to show that 2 layers suffice.

The following theorem will have three improvements over the earlier one.

1. It will work with $\mathcal{N}_2$, not $\mathcal{N}_3$.

2. σ will be more general than the ReLU.

3. The notion of function distance will be more restrictive; namely, it will be the “uniform norm”

$$\|f - g\|_{\mathrm{u}} := \sup_{x \in [0,1]} |f(x) - g(x)|.$$

Theorem (Hornik, Stinchcombe, and White 1989)**.** Let any continuous nondecreasing $\sigma : \mathbb{R} \to \mathbb{R}$ be given with

$$\lim_{z \to \infty} \sigma(sz) = \begin{cases} 1 & \text{when } s = +1, \\ 0 & \text{when } s = -1. \end{cases}$$

For any function $\phi : \mathbb{R} \to \mathbb{R}$, define $\mathcal{H}_\phi := \{x \mapsto \phi(a^\top x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}$. Then

$$\sup_{g \text{ continuous}} \inf_{f \in \operatorname{span}(\mathcal{H}_\sigma)} \|f - g\|_{\mathrm{u}} = 0.$$

Remark. The result also works for ReLU: for any $\epsilon > 0$, apply the theorem on $\sigma(z) := \max\{0, z\} - \max\{0, z - 1\}$ to get an approximating $f \in \operatorname{span}(\mathcal{H}_\sigma)$, and then rewrite each σ as two ReLUs and collect terms.

Returning to the proof, it will have the following scheme:

1. Let $\operatorname{prod}(\mathcal{F})$ denote products of functions in $\mathcal{F}$, meaning

   $$\operatorname{prod}(\mathcal{H}) := \{x \mapsto f_1(x) \cdots f_r(x) : r \in \mathbb{Z}_+, (f_1, \ldots, f_r) \in \mathcal{H}^r\}.$$

   The first step is to show that $\operatorname{span}(\operatorname{prod}(\mathcal{H}_{\cos}))$ can fit continuous functions. It makes sense that $\operatorname{prod}(\cdot)$ is useful here; in the proof of the theorem for $\mathcal{N}_3$, rather than adding together indicators on intervals and applying a ReLU, it would have been convenient to simply product indicators of intervals together. On the other hand, the appearance of $\cos(\cdot)$ makes no sense for now; the choice is made for the next step.

2. The next step shows that $\operatorname{span}(\operatorname{prod}(\mathcal{H}_{\cos})) = \operatorname{span}(\mathcal{H}_{\cos})$. This step is a consequence of properties of $\cos(\cdot)$; in my opinion, it was brilliant of Hornik, Stinchcombe, and White (1989) to see that $\cos(\cdot)$ could be useful in proving representation properties of neural networks.

3. The last step is to show that elements of $\mathcal{H}_{\cos}$ can be approximated with elements of $\operatorname{span}(\mathcal{H}_\sigma)$.

Lemma 1. For convenience, define $\mathcal{F}_1 := \operatorname{span}(\operatorname{prod}(\mathcal{H}_{\cos}))$.

$$\sup_{g \text{ continuous}} \inf_{f \in \mathcal{F}_1} \|f - g\|_{\mathrm{u}} = 0.$$

Proof. In lecture 2, we mentioned Weierstrass’s theorem, which says that polynomials can fit continuous functions; to see the possibility of relevance, note that polynomials have the form

$$\text{polynomials over } \mathbb{R}^d = \operatorname{span}(\operatorname{prod}(\{x \mapsto x_i : i \in [d]\})).$$

There is a stronger theorem, the Stone-Weierstrass theorem, which says that even things just resembling polynomials in a few ways can fit continuous functions. The proof of the present lemma is thus a direct consequence of Stone-Weierstrass, which can be applied after checking the following properties (Folland 1999, Corollary 4.50).

1. $\mathcal{F}_1$ must be an *algebra*, meaning it is closed under scalar multiplication, addition, and products. The first two are direct from $\operatorname{span}(\cdot)$, and the second since multiplication distributes over addition.

2. $\mathcal{F}_1$ must be continuous. This holds since σ is continuous, thus $\mathcal{H}_\sigma$ is continuous (it’s the composition of σ with affine functions), and lastly $\operatorname{span}(\operatorname{prod}(\mathcal{H}_\sigma))$ is continuous (products and linear combinations preserve continuity).

3. $\mathcal{F}_1$ *separates points*, meaning for any $x \neq y$, there exists $f \in \mathcal{F}_1$ with $f(x) \neq f(y)$. It suffices to define

$$f(z) := \cos\left((z-x)^\top (y-x)/\|y-x\|^2\right),$$

   whereby $f \in \mathcal{F}_1$ and $f(x) = 0 \neq 1 = f(y)$.

4. For every $x \in \mathbb{R}^d$, there exists $f \in \mathcal{F}_1$ such that $f(x) \neq 0$. This one is easy: ignore x, and just define $f(x) := \cos(0^\top x + 1) = \cos(1) \neq 0$.

□

Lemma 2. Define $\mathcal{F}_2 := \operatorname{span}(\mathcal{H}_{\cos})$. Then $\mathcal{F}_1 = \operatorname{span}(\operatorname{prod}(\mathcal{H}_{\cos})) = \operatorname{span}(\mathcal{H}_{\cos}) = \mathcal{F}_2$.

Proof. Each $f \in \mathcal{F}_1$ can be written as a linear combination of *terms*, where each term is a product of elements of $\mathcal{H}_{\cos}$; let the *size* of a term be the number of elements of $\mathcal{H}_{\cos}$ used to form the product.

Let $\mathcal{G}_i$ denote the elements of $\mathcal{F}_1$ where the largest term has size at most i. Consequently, $\mathcal{F}_1 = \cup_{i\geq 1}\mathcal{G}_i$. Note that $\mathcal{G}_i$ is defined *syntactically*, meaning the definition depends on the way elements of $\mathcal{F}_1$ are written down; the proof will show $\mathcal{G}_i = \mathcal{F}_2 = \mathcal{F}_1$, thus every element of any $\mathcal{G}_i$ can certainly be written with products of size 1.

The proof now establishes, by induction, that $\mathcal{G}_i = \mathcal{F}_2$. The base case $i = 1$ holds since $\mathcal{G}_1 = \operatorname{span}(\mathcal{H}_{\cos}) = \mathcal{F}_2$. Now consider the inductive step, some $\mathcal{G}_i$ with $i \geq 2$. Consider any element of $f \in \mathcal{G}_i$ whose largest term has size i (if no such element exists, there is nothing to show). Consider any product term of size i within f. By the trigonometric identity

$$\cos(a^\top x + b)\cos(r\top x + s) = \frac{1}{2}\left(\cos\left((a+r)^\top x + b + s\right) + \cos\left((a+r)^\top x - b + s\right)\right),$$

this product of i elements of $\mathcal{H}_{\cos}$ can be written as sum of a product of $i-1$ elements of $\mathcal{H}_{\cos}$. Proceeding in this way for each term of size i in f, then f can be rewritten to lie within $\mathcal{G}_{i-1}$. Since $f \in \mathcal{G}_{i-1}$ was arbitrary, then $\mathcal{G}_i \subseteq \mathcal{G}_{i-1}$. But $\mathcal{G}_i \supseteq \mathcal{G}_{i-1}$ by definition, thus $\mathcal{G}_i = \mathcal{G}_{i-1}$, which combined with the inductive hypothesis gives $\mathcal{G}_i = \mathcal{G}_{i-1} = \mathcal{F}_2$. □

Lemma 3. For any $f \in \mathcal{H}_{\cos}$ and any $\epsilon > 0$, there exists $g \in \operatorname{span}(\mathcal{H}_\sigma)$ with $\|f - g\|_{\mathrm{u}} \leq \epsilon$.

Proof. *(In class, a picture was drawn; the full proof will be an optional homework exercise.)* □

Proof (of main theorem)**.** It suffices to show that for any continuous function g and any $\epsilon > 0$, there exists $f \in \mathcal{N}_2$ with $\|f - g\|_{\mathrm{u}} \leq \epsilon$. By the first two lemmas, there exists $h \in \mathcal{F}_2 = \operatorname{span}(\mathcal{H}_{\cos})$ with $\|h - g\|_{\mathrm{u}} \leq \epsilon/2$. Now let N denote the number of terms in h, and moreover suppose it has the form

$$h = \sum_i a_i h_i,$$

where $h_i \in \mathcal{H}_{\cos}$, and without loss of generality each a_i is nonzero. By Lemma 3, for each h_i there exists $f_i \in \operatorname{span}(\mathcal{H}_\sigma)$ with $\|f_i - g_i\|_{\mathrm{u}} \leq \epsilon/(2|a_i|N)$. Defining $f := \sum_i a_i f_i \in \operatorname{span}(\mathcal{H}_\sigma) = \mathcal{N}_2$, the triangle inequality grants

$$\|g - f\|_{\mathrm{u}} \leq \|g - h\|_{\mathrm{u}} + \|h - f\|_{\mathrm{u}} \leq \frac{\epsilon}{2} + \left\|\sum_i (a_i h_i - a_i g_i)\right\|_{\mathrm{u}} \leq \frac{\epsilon}{2} + \sum_i |a_i|\|h_i - g_i\|_{\mathrm{u}} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$

□

Remark.

1. Notice that both constructions today, just like those from earlier lectures, have some sort of exponential blowup; for instance, in the proof of Lemma 2, reducing from products of i things to products of $i-1$ things will double the number of terms.

2. Performing multiplication was useful in the proof for $\mathcal{N}_2$, but it was also shown to be unnecessary. It is an **open problem** whether multiplication nodes really help neural networks in some fundamental sense. For instance, there is extensive work on "sum-product networks" (Poon and Domingos 2011), but they do not seem to be used in many of the practical successes of neural networks (which in turn led to their recent popularity).

References

Folland, Gerald B. 1999. *Real Analysis: Modern Techniques and Their Applications.* 2nd ed. Wiley Interscience.

Hornik, K., M. Stinchcombe, and H. White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2 (5). Elsevier Science Ltd.: 359–66.

Poon, Hoifung, and Pedro M. Domingos. 2011. "Sum-Product Networks: A New Deep Architecture." In *UAI 2011*, 337–46.