6. Approximation:
GT as an infinite-width net

Recap

- Last few lectures, we have covered basic universal approximation results

- **Key idea.** Neural nets can express the basis of other functions
  - Pulses
  - Fourier basis

- Sometimes, we managed to prove explicit bounds on the #neurons needed
  - Unfortunately, when invoking Stone-Weierstrass, no explicit bound

Today

- Play with a powerful tool: sampling!
    - Widely used in the analysis of algorithms

- **Rough sketch**
    - Ground truth $g(\cdot)$: An infinite-width neural network
    - Neural net $f(\cdot)$: A neural net constructed by sampling the GT neurons

    - As the number of samples (i.e., neurons) increase, we have

$$f(\cdot) \to g(\cdot), \qquad \text{at some rate}$$

        - Analyze this to get finite-width guarantees

Today

- **Key Questions**
  - **Q1.** How do we express $g(\cdot)$ as an infinite-width neural net?
  - **Q2.** How do we analyze the convergence rate of $f(\cdot) \to g(\cdot)$?

- Today, we'll cover Q1, and do warm-up for Q2

Formalization

- First, we'll formalize the concept of (uncountably) infinite-width two-layer net
  - Unfortunately, we'll stick to threshold nets only

- We will show that:

$$g(\mathbf{x}) = \int \pi(\mathbf{w}, b) \cdot a(\mathbf{w}, b) \cdot \mathbf{1}\{\mathbf{w}^\top \mathbf{x} \geq b\} \, \mathrm{d}\mathbf{w} \, \mathrm{d}b$$

  - Here, we have:
    - $(\mathbf{w}, b)$ specifies each neuron — unique 1st layer parameters
    - $a(\mathbf{w}, b)$ is the corresponding second layer weight
    - $\pi(\mathbf{w}, b)$ is the probability density over the neurons

  - **Remark.** This is an exact equality, not an approximation

Formalization

- From this distribution of neurons, we will sample the neurons to build a finite-width net
    - **Step 1.** Draw the neurons:

$$(\mathbf{w}_i, b_i) \sim \pi(\mathbf{w}, b)$$

    - **Step 2.** Build

$$f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} a(\mathbf{w}_i, b_i) \cdot \mathbf{1}\{\mathbf{w}_i^\top \mathbf{x} \geq b_i\}$$

- If $m \to \infty$, we have certain convergence
    - Later, we'll study good tools to quantify the convergence

GT as an infinite-width net

Univariate case

- First, let's convince ourselves that any GT $g(\cdot)$ is an infinite-width two-layer threshold net
  - Let us first consider the easy case: univariate

Proposition 3.1.

Suppose that we have a univariate function over a compact domain, $g : [0,1] \to \mathbb{R}$.
Suppose further that $g(0) = 0$. Then, for $x \in [0,1]$, we have

$$g(x) = \int_0^1 g'(b) \cdot \mathbf{1}[x \geq b] \, \mathrm{d}b$$

- Any proof ideas?

Proof idea

- Recall the "fundamental theorem of calculus"

First part [edit]

This part is sometimes referred to as the *first fundamental theorem of calculus*.[6]

Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined, for all x in $[a, b]$, by

$$F(x) = \int_a^x f(t)\,dt.$$

Then F is uniformly continuous on $[a, b]$ and differentiable on the open interval (a, b), and

$$F'(x) = f(x)$$

for all x in (a, b) so F is an antiderivative of f.

Univariate case

- Let's take another look at what we proved:

$$g(x) = \int_0^1 g'(b) \cdot \mathbf{1}[x \geq b] \, \mathrm{d}b$$

- This is an infinite-width two-layer threshold network, with
  - 1st layer weights $w = 1$
  - biases b (the only parameter)
  - 2nd layer weights $a(b) = g'(b)$
  - probability density $\pi(b) = \mathrm{Unif}([0,1])$

Flashback

$$g(x) = \int_0^1 g'(b) \cdot \mathbf{1}[x \geq b] \, \mathrm{d}b$$

- Recall that, several lecture ago, we considered a neural net construction

$$f(x) = \sum_{i=1}^{m} \left(g(b_i) - g(b_{i-1})\right) \cdot \mathbf{1}[x \geq b_i]$$

  - This can also be viewed as a version of sampling:
    - Using a uniform grid — instead of uniform distribution
    - Using differentials — instead of derivatives

- In this sense, what we are working on today is extending this idea further for a general technique

Multivariate case

- How do we extend this to a multivariate input case?
  - Ultimately, we want to prove something like:

Claim (informal)

Under *some conditions*, we have

$$g(x) = \int\int q(\mathbf{w}, b) \cdot \mathbf{1}\{\mathbf{w}^\top \mathbf{x} \geq b\} \, \mathrm{d}\mathbf{w} \, \mathrm{d}b$$

- Here, for simplicity, we are using a merged form

$$q(\mathbf{w}, b) = \pi(\mathbf{w}, b) \cdot a(\mathbf{w}, b)$$

  - Given some q, we can always come up with (π, a) where π is a valid probability density

Multivariate case

- Unfortunately, this is not very easy…
  - Can you think of a good multivariate analogue of FTC?
    (there is one for the line integral, which is meh)
  - Handling various "directions" is the key challenge

- **Tool.** Fourier transform and complex numbers
  - Will follow the exposition of "new" MJT notes

Preliminaries:
Fourier Transform

Notations and assumptions

- **Notation.** For a complex number, the absolute value $|\cdot|$ denotes the ℓ_2 norm, i.e.,

$$|b + ci| = \sqrt{b^2 + c^2}$$

Definition (Integrable)

A function $g : \mathbb{R}^d \to \mathbb{R}$ is called integrable whenever it satisfies

$$\int_{\mathbb{R}^d} |g(\mathbf{x})| \mathrm{d}\mathbf{x} < \infty$$

- We will write $g \in L^1$
- Will be our running assumption

Fourier Transform

Definition (Fourier Transform)

The Fourier transform $\tilde{g} : \mathbb{R}^d \to \mathbb{C}$ of an integrable function $g : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\tilde{g}(\mathbf{w}) = \int_{\mathbb{R}^d} \exp(-2\pi i \mathbf{w}^\top \mathbf{x}) \cdot g(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

- If you are not familiar with this form, recall that (one of) the Euler's formula says:

$$\exp(ix) = \cos(x) + i \cdot \sin(x)$$

- That is, the Fourier transform is simply extracting the frequency components of $g(\mathbf{x})$
  - Two sinusoids with different frequencies are orthogonal
  - In multivariate case, the frequencies will have "directions" in addition to magnitudes

Properties

- Here are some well-known properties of the Fourier transform:
  - **Inversion.** If $\tilde{g} \in L^1$, then
$$g(\mathbf{x}) = \int \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w}) \, \mathrm{d}\mathbf{w}$$
  - **Derivatives.** Given some $\mathbf{w} \in \mathbb{R}^d$, we have
$$2\pi \|\mathbf{w}\| \cdot |\tilde{g}(\mathbf{w})| = \|\widetilde{\nabla g}\|$$
  - **Real parts.** Let $\Re[b + ic] = b$ denote the real part of a complex number. Then, for an integrable complex function $h : \mathbb{R}^d \to \mathbb{C}$, we have:
$$\Re\left[\int_{\mathbb{R}^d} h(\mathbf{x}) \, \mathrm{d}\mathbf{x}\right] = \int_{\mathbb{R}^d} \Re[h(\mathbf{x})] \, \mathrm{d}x$$

</preliminaries>

Inverse Fourier Transforms

- Notice that the inverse Fourier transform can be readily viewed as an infinite-width net

$$g(x) = \int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\, \mathrm{d}\mathbf{w}$$

  - Indeed, this is the case where
    - $\tilde{g}(\mathbf{w})$ is the neuron density (multiplied by 2nd layer weights)
    - $t = \exp(2\pi i t)$ is the activation function
    - b there is no bias!

Inverse Fourier Transforms

$$g(x) = \int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w}) \, \mathrm{d}\mathbf{w}$$

- Our goal is to re-write this, using threshold activations

$$g(x) = \int_{\mathbb{R}^d} \mathbf{1}[\mathbf{u}(\mathbf{w})^\top x \geq b(\mathbf{w})] \cdot a(\mathbf{w}) \, \mathrm{d}\mathbf{w}$$

  - Note that we are using a slightly different notation now
    - First-layer weights $\mathbf{u}$
    - Biases b

- This is done in two steps:
  - **Step 1.** Turn IFT into cosine nets
  - **Step 2.** Turn cosine nets into threshold nets

Step 1. IFT -> Cosine nets

$g(\mathbf{x}) = \Re[g(\mathbf{x})]$

Step 1. IFT -> Cosine nets

$$g(\mathbf{x}) = \Re[g(\mathbf{x})]$$

$$= \Re\left[\int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\, \mathrm{d}\mathbf{w}\right]$$

IFT

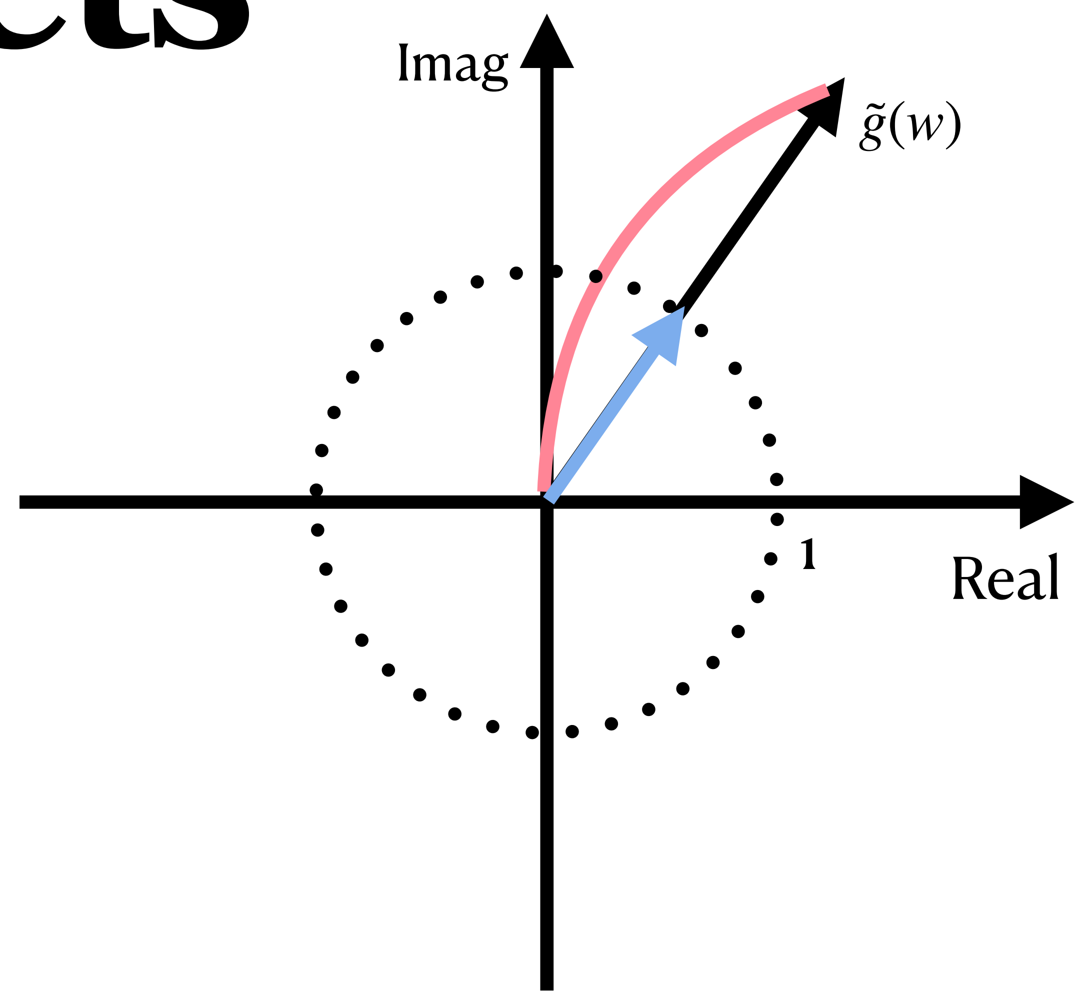
Step 1. IFT -> Cosine nets

$$
\begin{aligned}
g(\mathbf{x}) &= \Re[g(\mathbf{x})] \\
&= \Re\left[\int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\, \mathrm{d}\mathbf{w}\right] \\
&= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\right] \mathrm{d}\mathbf{w}
\end{aligned}
$$

"Real Parts" property

Step 1. IFT -> Cosine nets

$$g(\mathbf{x}) = \Re[g(\mathbf{x})]$$
$$= \Re\left[\int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\, \mathrm{d}\mathbf{w}\right]$$
$$= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\right] \mathrm{d}\mathbf{w}$$
$$= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \exp(2\pi i \theta_{\tilde{g}}(\mathbf{w})) \cdot |\tilde{g}(\mathbf{w})|\right] \mathrm{d}\mathbf{w}$$



Polar decomposition

Step 1. IFT -> Cosine nets

$$
\begin{aligned}
g(\mathbf{x}) &= \Re[g(\mathbf{x})] \\
&= \Re\left[\int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\, \mathrm{d}\mathbf{w}\right] \\
&= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\right] \mathrm{d}\mathbf{w} \\
&= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \exp(2\pi i \theta_{\tilde{g}}(\mathbf{w})) \cdot |\tilde{g}(\mathbf{w})|\right] \mathrm{d}\mathbf{w} \\
&= \int_{\mathbb{R}^d} \Re\left[\exp\Big(2\pi i \big(\mathbf{w}^\top \mathbf{x} + \theta_{\tilde{g}}(\mathbf{w})\big)\Big)\right] \cdot |\tilde{g}(\mathbf{w})|\, \mathrm{d}\mathbf{w}
\end{aligned}
$$

Magnitude is real

Step 1. IFT -> Cosine nets

$$
\begin{aligned}
g(\mathbf{x}) &= \Re[g(\mathbf{x})] \\
&= \Re\left[\int_{\mathbb{R}^d} \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\, \mathrm{d}\mathbf{w}\right] \\
&= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \tilde{g}(\mathbf{w})\right] \mathrm{d}\mathbf{w} \\
&= \int_{\mathbb{R}^d} \Re\left[\exp(2\pi i \mathbf{w}^\top \mathbf{x}) \cdot \exp(2\pi i \theta_{\tilde{g}}(\mathbf{w})) \cdot |\tilde{g}(\mathbf{w})|\right] \mathrm{d}\mathbf{w} \\
&= \int_{\mathbb{R}^d} \Re\left[\exp\Big(2\pi i\big(\mathbf{w}^\top \mathbf{x} + \theta_{\tilde{g}}(\mathbf{w})\big)\Big)\right] \cdot |\tilde{g}(\mathbf{w})|\, \mathrm{d}\mathbf{w} \\
&= \int_{\mathbb{R}^d} \cos\big(2\pi\big(\mathbf{w}^\top \mathbf{x} + \theta_{\tilde{g}}(\mathbf{w})\big)\big) \cdot |\tilde{g}(\mathbf{w})|\, \mathrm{d}\mathbf{w}
\end{aligned}
$$

Euler's formula

Step 1. IFT -> Cosine nets

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} \cos\Big(2\pi\Big(\mathbf{w}^\top \mathbf{x} + \theta_{\tilde{g}}(\mathbf{w})\Big)\Big) \cdot |\tilde{g}(\mathbf{w})| \, \mathrm{d}\mathbf{w}$$

- That is, $g(\cdot)$ is an infinite-width two-layer cosine network

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} \widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) \cdot |\tilde{g}(\mathbf{w})| \, \mathrm{d}\mathbf{w}$$

  - Here, we use the shorthand notations
    - $\widetilde{\cos}(x) := \cos(2\pi x)$
    - $\theta_{\mathbf{w}} = \theta_{\tilde{g}}(\mathbf{w})$

  - Density $|\tilde{g}(\mathbf{w})|$
  - 1st layer weight $\mathbf{w}$
  - bias $\theta_{\mathbf{w}}$

Step 2. Cosine nets -> Threshold nets

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} \widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) \cdot |\,\tilde{g}(\mathbf{w})\,| \,\mathrm{d}\mathbf{w}$$

- Now we want to turn this into a threshold network!
    - Need to do something that is not very straightforward…

Step 2. Cosine nets -> Threshold nets

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} \widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) \cdot |\tilde{g}(\mathbf{w})| \, \mathrm{d}\mathbf{w}$$

- Now we want to turn this into a threshold network!
    - Need to do something that is not very straightforward...

$$\widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) - \widetilde{\cos}\left(\theta_{\mathbf{w}}\right)$$

$$= -2\pi \int_0^{\mathbf{w}^\top \mathbf{x}} \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \, \mathrm{d}b$$

Difference as an integration

Step 2. Cosine nets -> Threshold nets

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} \widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) \cdot |\tilde{g}(\mathbf{w})| \, \mathrm{d}\mathbf{w}$$

- Now we want to turn this into a threshold network!
  - Need to do something that is not very straightforward...

$$\widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) - \widetilde{\cos}\left(\theta_{\mathbf{w}}\right)$$

$$= -2\pi \int_0^{\mathbf{w}^\top \mathbf{x}} \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \mathrm{d}b$$

$$= -2\pi \int_0^{\|\mathbf{w}\|} \mathbf{1}[\mathbf{w}^\top \mathbf{x} \geq b] \cdot \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \mathrm{d}b + 2\pi \int_{-\|\mathbf{w}\|}^{0} \mathbf{1}[\mathbf{w}^\top \mathbf{x} \leq b] \cdot \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \mathrm{d}b$$

Generate thresholds, by dividing it into cases

Step 2. Cosine nets -> Threshold nets

$$g(\mathbf{x}) = \int_{\mathbb{R}^d} \widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) \cdot |\tilde{g}(\mathbf{w})| \, \mathrm{d}\mathbf{w}$$

- Now we want to turn this into a threshold network!
  - Need to do something that is not very straightforward...

$$\widetilde{\cos}\left(\mathbf{w}^\top \mathbf{x} + \theta_{\mathbf{w}}\right) - \widetilde{\cos}\left(\theta_{\mathbf{w}}\right)$$

$$= -2\pi \int_0^{\mathbf{w}^\top \mathbf{x}} \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \mathrm{d}b$$

$$= -2\pi \int_0^{\|\mathbf{w}\|} \mathbf{1}[\mathbf{w}^\top \mathbf{x} \geq b] \cdot \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \mathrm{d}b + 2\pi \int_{-\|\mathbf{w}\|}^{0} \mathbf{1}[\mathbf{w}^\top \mathbf{x} \leq b] \cdot \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right) \mathrm{d}b$$

$$= 2\pi \int_0^{\|\mathbf{w}\|} \left[\widetilde{\sin}\left(-b + \theta_{-\mathbf{w}}\right) - \widetilde{\sin}\left(b + \theta_{\mathbf{w}}\right)\right] \cdot \mathbf{1}[\mathbf{w}^\top \mathbf{x} \geq b] \, \mathrm{d}b$$

Reparametrize and combine

Theorem

- Plugging into the $g(\mathbf{x})$, we get the following theorem:

Theorem.

Let $g, \tilde{g} \in L^1$ and $g(0) = 0$. Then, we have

$$g(\mathbf{x}) = \int\int q(\mathbf{w}, b) \cdot \mathbf{1}[\mathbf{w}^\top \mathbf{x} \geq b] \, \mathrm{d}\mathbf{w} \, \mathrm{d}b$$

where $q(\mathbf{w}, b)$ is the parameter density

$$q(w, b) = 2\pi |\tilde{g}(\mathbf{w})| \left(\widetilde{\sin}(-b + \theta_{-\mathbf{w}}) - \widetilde{\sin}(b + \theta_{\mathbf{w}}) \right) \cdot \mathbf{1}[0 \leq b \leq \|\mathbf{w}\|]$$

Moreover, we have

$$\int\int |q(\mathbf{w}, b)| \, \mathrm{d}\mathbf{w} \, \mathrm{d}b \leq 2 \int \|\widetilde{\nabla g}\| \mathrm{d}\mathbf{w}$$

- Note. Where did $\widetilde{\cos}(\theta_{\mathbf{w}})$ go?

Next up

- Sampling from the infinite-width nets
    - Analysis