§5. Depth Separation

Q: Why deep neural networks when 2-layer NNs are already universal approximators?

A1: $\exists$ input distr. D, target function f_* s.t. w.r.t. D,

no $\mathrm{poly}(d/\varepsilon)$-width 2-layer networks can approximate f_* (ES16, SES19)

while GF + some $\mathrm{poly}(d/\varepsilon)$-width 3-layer network can efficiently learn f_* (RZG23)

§5.1 Negative results.

1. Fourier transform.

(Assume all integrals exist and are finite)

Fourier transform. $\mathcal{F}: L^2 \to L^2$

$$f \mapsto (\mathcal{F}f)(\xi) := \int_{\mathbb{R}^d} f(x) \exp(-2\pi i \langle x, \xi \rangle)\, dx.$$

Notation: $\hat{f} := \mathcal{F} f$.

Properties: $\mathcal{F}$ linear, invertible, isometric,

$$\widehat{fg} = \hat{f} * \hat{g}.$$

2. $R_d > 0$ s.t. $\mathrm{Vol}(R_d B_d) = 1$

$\hookrightarrow$ unit ball in $\mathbb{R}^d$.

φ - inverse Fourier transform of $\mathbb{1}_{R_d B_d}$.

$\tilde{\mathcal{F}}$ · isometric $\Rightarrow$ φ^2 is a density.

$\hookrightarrow$ let this be the input distr.

g - target

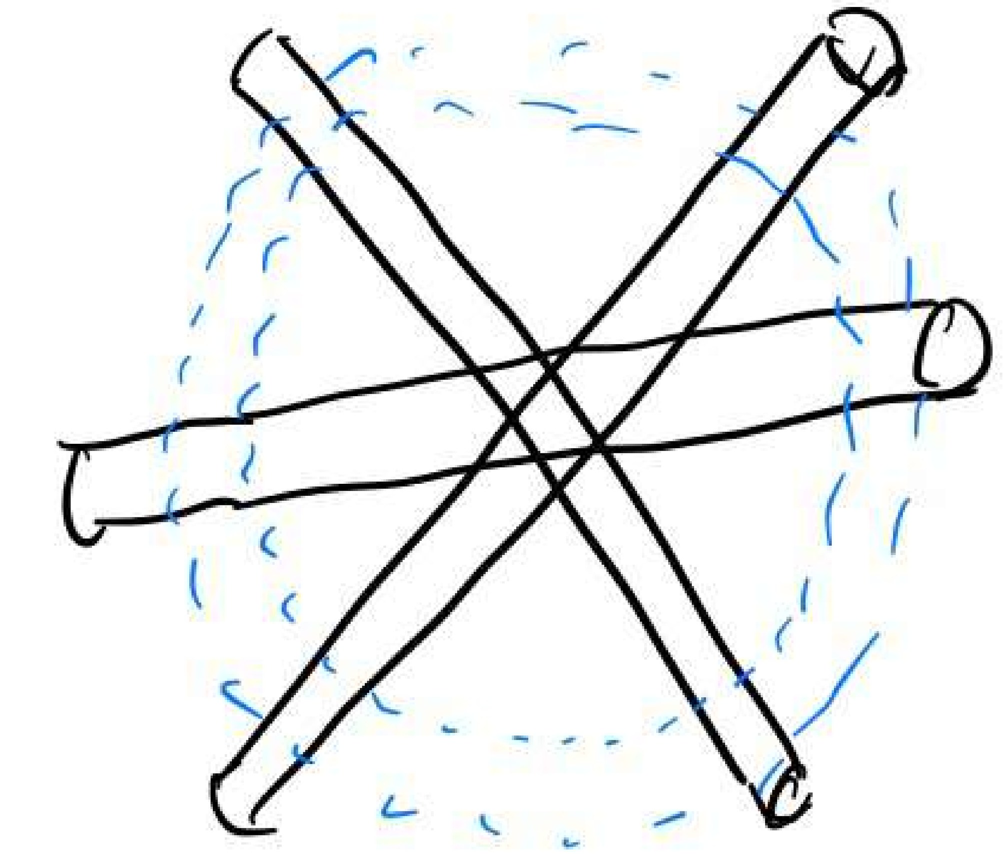
$f(x) = \sum_{k=1}^{m} f_i(v_k \cdot x)$. - two-layer network

$(v_k)_{k=1}^{m}$ weights.

$$\mathrm{Loss} = \|f - g\|^2_{L^2(\varphi^2)} = \|f\varphi - g\varphi\|^2_{L^2} = \|\widehat{f\varphi} - \widehat{g\varphi}\|^2_{L^2}.$$

3. Lemma. $\widehat{fg}$ is supported within some tubes.

$$\mathrm{supp}(\widehat{fg}) \subset \bigcup_{k=1}^{m} (\mathrm{span}(v_k) + R_d B_d) =: T$$



sparse when d is large.

$\widehat{g}$. radial. has some mass away from the origin.

Pf. (Informal). $\widehat{fg} = \sum_{k=1}^{m} \widehat{f_k g} = \sum_{k=1}^{m} \widehat{f}_k * \widehat{g} = \sum_{k=1}^{m} \widehat{f}_k * 1_{R_d B_d}$

$$\Rightarrow \mathrm{supp}(\widehat{fg}) \subset \bigcup_{k=1}^{m} (\mathrm{supp}(\widehat{f}_k) + R_d B_d).$$

Claim. $\mathrm{supp}\, \widehat{f}_k \subseteq \mathrm{span}\, v_k$.

$\xi \in \mathbb{R}^d$. $\xi = a v_k + u$. $a \in \mathbb{R}$. $u \perp v_k$.

$$\widehat{f}_k(\xi) = \int f_k(v_k \cdot x) \exp(-2\pi i \langle x, \xi \rangle) dx$$

$$= \int f_k(v_k \cdot x) \exp(-2\pi i\, a \langle x, v_k \rangle) \exp(-2\pi i \langle x, u \rangle)\, dx =: \widehat{f}_k(a, u).$$

Symmetry $\Rightarrow$ $\widehat{f}_k(a,u) = \widehat{f}_k(a,-u)$

$$\Rightarrow \widehat{f}_k(\xi) = \frac{1}{2}\left(\widehat{f}_k(a,u) + \widehat{f}_k(a,-u)\right) = \int f_k(v_k \cdot x) \exp(-2\pi i\, a \langle x, v_k \rangle) \cos(-2\pi \langle x, u \rangle)\, dx$$

$\Rightarrow = 0$ if $u \neq 0$.

$\boxtimes$

4. Lemma. q, w. $\|q\|_{L^2} = \|w\|_{L^2} = 1$.

$\operatorname{supp} q \subseteq T$.

w - radial, $\int_{(2R_d B_d)^c} w^2 \geq \delta$. $\delta \in [0,1]$

↳ has some mass away from the origin.

$\Rightarrow \langle q, w \rangle_{L^2} \leq 1 - \frac{\delta}{2} + \underbrace{m \exp(-cd)}_{\to 0 \text{ as } d \to \infty}$

Pf. Define $A = (2R_d B_d)^c$ and

$$h(r) = \frac{\operatorname{Vol}(rS^{d-1} \cap T)}{\operatorname{Vol}(rS^{d-1})}.$$

claim (without pf). $h \downarrow$. $h(r) \leq m \exp(-cd)$.

Write $\langle q, w \rangle_{L^2} = \int_{A^c} qw + \int_A qw$

$\leq \|q\|_{L^2} \|w \mathbb{1}_A\|_{L^2} + \int_A qw$

$\leq \sqrt{1-\delta} + \int_A qw$.

claim. $\int_A qw \leq m \exp(-cd/2)$.

Pf. $\int_A qw = \int_A \tilde{q} w \leq \overset{\leq 1}{\cancel{\|w\|_{L^2}}} \|\tilde{q} \mathbb{1}_A\|_{L^2}$

$$= \sqrt{\int_{2R_d}^{\infty} \tilde{q}^2(r) \operatorname{Vol}(rS^{d-1}) \, dr}$$

Note. $\tilde{q}(r) = \frac{\int_{T \cap rS^{d-1}} q}{\operatorname{Vol}(rS^{d-1})} = h(r) \frac{\int_{T \cap rS^{d-1}} q}{\operatorname{Vol}(T \cap rS^{d-1})}$

$\leq h(r) \sqrt{\frac{\int_{T \cap rS^{d-1}} q^2}{\operatorname{Vol}(T \cap rS^{d-1})}}$

$$\int_A g w \le \sqrt{\int_{2R_d B_d}^{\infty} h^2(r) \frac{\int_{rS^{d-1}\cap T} g^2}{\mathrm{Vol}(T\cap rS^{d-1})} \mathrm{Vol}(rS^{d-1})}$$

$$= \sqrt{\int_{2R_d B_d}^{\infty} h(r) \int_{rS^{d-1}} g^2}$$

$$\le \sqrt{h(2R_d B_d) \underbrace{\int_{2R_d B_d}^{\infty} \int_{T\cap rS^{d-1}} g^2}_{\le \|g\|_2^2}}$$

$$\le \sqrt{h(2R_d B_d)} \le m\exp(-cd/2). \quad \square$$

5. __Lemma.__ $\exists\ \tilde{g}(x) = \sum_{i=1}^{N} \varepsilon_i g_i(\|x\|)$.

where $\varepsilon_i \in \{\pm 1\}$. g_i is the indicator function of some interval Δ_i. s.t.

$w = \widehat{\tilde{g} g} / \|\tilde{g} g\|_2$, satisfies the condition of the previous lemma. for some universal constant δ.

__Corollary.__ $\|f - \tilde{g}\|_{L^2(g^2)} \ge \Omega(\delta)$.

[SS17]: 2-layer NN cannot efficiently approximate Ball indicators

[SES19]: $\sim$ $\mathrm{ReLU}(1-\|x\|)$.

General message: high-frequency $\Rightarrow$ harder to approximate / learn.

Hierachical structure?

[Abbe, Boix-Adsera, Misiakiewicz, 22]. Merged-staircase property

[ABM, 23] follow up.

[Allen-Zhu, Li. 21]. Backward feature correction.

§5.2 Positive results.

Target $\quad f_*(x) = \sigma(1 - \|x\|)$. $\quad \sigma$ - ReLU.

learner
$$\begin{cases} F(x;\mu_1) = \mathbb{E}_{v_1\sim\mu_1} \|w\| \, \sigma(v \cdot x). & \text{first layer} \\ f(x;\mu_2,\mu_1) = \mathbb{E}_{\substack{(w_2,b_2)\\ \sim\mu_2}} \sigma\left(w_2 F(x;\mu_1) + b_2\right). & \text{second layer} \end{cases}$$

Loss. $\quad \mathcal{L}(\mu_1,\mu_2) = \text{MSE} = \frac{1}{2} \mathbb{E}_x \left\{ \left(f_*(x) - f(x;\mu_2,\mu_1) \right)^2 \right\}$.

GF. $\quad v_1 \in \mu_1$. $\quad \dot{v}_1 = \mathbb{E}_x \left\{ S_2(x) \left(\bar{v}_1 \sigma(v_1 \cdot x) + \|v_1\| \sigma'(v_1 \cdot x)\, x \right) \right\}$.

where $\quad S_2(x) = (f_*(x) - f(x)) \, \mathbb{E}_{w_2,b_2} \left\{ \sigma'(w_2 F(x) + b_2)\, w_2 \right\}$.

(The dynamics of 2nd layer neurons are not very important).

1. The infinite-width dynamics are much simpler than the finite-width ones.

Lemma. μ spherically symmetric. $\Rightarrow$ $\mathbb{E}_{w\sim\mu} \|w\| \sigma(w\cdot x) = C_\sigma \frac{\mathbb{E}_\mu \|w\|^2}{\sqrt{d}} \|x\|$. $C_\sigma = \Theta(1)$.

Pf. $\mathbb{E}_w \|w\| \sigma(w\cdot x) = \mathbb{E}_w \|w\|^2 \sigma(\bar{w}\cdot x) = \left(\mathbb{E}_{\|w\|} \|w\|^2\right)\left(\mathbb{E}_{\bar{w}} \sigma(\bar{w}\cdot x)\right)$ $\square$

$\downarrow$ 1-homogeneity $\quad\downarrow$ spherical symmetry $\quad\hookrightarrow \frac{C_\sigma}{\sqrt{d}}\|x\|$

· Define $\alpha = \frac{C_\sigma}{\sqrt{d}} \mathbb{E}\|w\|^2$. ($\leftarrow$ well-defined for discrete μ too).

· In the infinite-width limit $F(x;\mu_t) = \alpha \|x\|$. (1)

· Lemma. spherically symmetric g. $\Rightarrow$ $\mathbb{E}_x g(x)\sigma(v\cdot x) = \frac{C_\sigma}{\sqrt{d}} \mathbb{E}_x\{g(x)\|x\|\} \|v\|$, $\mathbb{E}_x g(x)\sigma'(v\cdot x) x = \frac{C_\sigma}{\sqrt{d}} \mathbb{E}_x\{g(x)\|x\|\} \bar{v}$.

$\Rightarrow \frac{d}{dt}\alpha = \mathbb{E}_{w_t} \frac{\partial\alpha}{\partial w_t}\frac{\partial w_t}{\partial t} = \cdots = \frac{4C_\sigma^2}{d} \mathbb{E}_x\{S(x)\|x\|\} \mathbb{E}_{w_t}\|w_t\|^2 = \frac{4C_\sigma}{\sqrt{d}} \mathbb{E}_x\{S(x)\|x\|\}\alpha$. (2)

(1), (2) $\Rightarrow$ $\alpha_t \in \mathbb{R}$ characterize the first layer (in the infinite-width limit).

Next goal: poly(d)-width discretization.

2. Symmetrization.

Given $g : \mathbb{R}^d \to \mathbb{R}$ define its symmetrization $\tilde{g}(x) := \mathbb{E}_{x' \in \|x\| S^{d-1}} g(x')$

<u>Observation</u>. $\tilde{F}(x) = \mathbb{E}_{w \sim \mu_1} \left[\|w\|^2 \underbrace{\mathbb{E}_{x' \sim \|x\| S^{d-1}} \sigma(\bar{w} \cdot x')}_{\to \frac{C}{\sqrt{d}} \|x\|} \right] = \frac{C}{\sqrt{d}} \mathbb{E} \|w\|^2 \|x\| = \alpha \|x\|$

not necessarily spherically symmetric

i.e. the infinite-width network $\to$ the symmetrization of the finite-width network.

<u>Observation</u>. $\mathbb{E}_x \left[(f_*(x) - f(x))^2 \right] = \mathbb{E}_x \left[(f_*(x) - \tilde{f}(x))^2 \right] + \mathbb{E}_x \left[(\tilde{f}(x) - f(x))^2 \right]$ (3)

$$- 2 \underbrace{\mathbb{E}_x \left[(f_*(x) - \tilde{f}(x)) (\tilde{f}(x) - f(x)) \right]}_{= 0}$$

$$\approx \mathbb{E}_x \left[(f_*(x) - \tilde{f}(x))^2 \right] + \frac{\overline{w}_2^2}{2} \mathbb{E}_x \left[(F(x) - \tilde{F}(x))^2 \right]$$

3. poly(d)-width discretization under symmetry. (In general, making $W_2(\mu, \hat{\mu}) \le \varepsilon$ requires exponentially many samples.)

Q: How to characterize the discretization error?

A1: Sample m neurons $\hat{\mu}_0$ from μ_0. move them according to the infinite-/finite-width dynamics. characterize the deviation.

A2: Characterize the deviation of the relevant function. (In our case, $\bar{F} = F/\alpha$ vs. $\|\cdot\|_2$)

<u>Lemma</u>. $$\frac{d}{dt} F(x) \approx -\frac{\overline{W}_2^2}{2} \mathop{\mathbb{E}}_{w_i} \left\{ \left\langle \nabla_{w_i} \bar{F}(x), \nabla_{w_i} \mathbb{E}\left(\tilde{F}(x) - F(x')\right)^2 \right\rangle \right\}.$$

Pf sketch. $$\frac{d}{dt}\bar{F}(x) = \frac{\frac{d}{dt}F(x)}{\alpha} - F(x)\frac{\dot{\alpha}}{\alpha} = -\frac{1}{\alpha}\mathop{\mathbb{E}}_{w_i}\langle \nabla_{w_i} F(x), \nabla_{w_i} L\rangle + \bar{F}(x)\frac{1}{\alpha}\mathop{\mathbb{E}}_{w_i}\langle \nabla_{w_i}\alpha, \nabla_{w_i} L\rangle$$

Recall (3). $L = L_1 + L_2$. Claim. ∇L_1 contributes little.

By symmetry, $\nabla_{v_i} L_1 = \beta v_i$ for some $\beta \in \mathbb{R}$.

$$\langle \nabla_{v_i} F(x), v_i \rangle = \langle \nabla_{v_i}(\|v_i\|^2 \sigma(\bar{v}_i \cdot x)), v_i \rangle = \sigma(\bar{v}_i\cdot x)\langle \nabla_{v_i}\|v_i\|^2, v_i\rangle = 2\|v_i\|^2\sigma(\bar{v}_i\cdot x)$$

$$\langle \nabla_{v_i}\alpha, v_i\rangle = \frac{C_\sigma}{\sqrt{d}}\langle \nabla_{v_i}\|v_i\|^2, v_i\rangle = 2C_\sigma/\sqrt{d}\,\|v_i\|^2$$

$$\Rightarrow \left.\frac{d}{dt}\bar{F}(x)\right|_{L_1} = -\frac{\beta}{\alpha}\, 2\mathop{\mathbb{E}}_{w_i}\|w_i\|^2\sigma(\bar{w}_i\cdot x) + \bar{F}(x)\frac{\beta}{\alpha}\frac{2C_\sigma}{\sqrt{d}}\mathbb{E}\|w_i\|^2 = -2\beta\bar{F}(x) + 2\beta\bar{F}(x) = 0$$

Corollary. $\frac{d}{dt}\| F - \|\cdot\|_2 \|_{L^2}^2 \lesssim 0$. In words, the discretization error barely grows.

Takeaways.

1. The infinite-width dynamics can be much simpler than the finite-width ones.

2. poly(d)-width discretization is sometimes possible.

   * Essentially, we are Taylor expanding the dynamics around the infinite-width trajectory and showing that the first-order error terms are good. (no compounding errors).