# Depth Separation with Multilayer Mean-Field Networks

Yunwei Ren
CMU

Mo Zhou
Duke University

Rong Ge
Duke University

# Background: Depth Separation

- Why deeper networks are more powerful than shallow ones?
- Eldan & Shamir, 2016; Telgarsky, 2016; Daniely, 2017 ...
  - There exists some functions that are **approximable** by deep networks by not by shallow ones.

**Theorem (Theorem 5 of (Safran et al., 2019))**

*There exists a* **spherically symmetric** *input distribution* $\mathcal{D}$ *s.t.* **no** *2-layer networks with width* $\text{poly}(d/\varepsilon)$ *can* **approximate** *the target function* $f_*(\boldsymbol{x}) = \mathrm{ReLU}(1 - \|\boldsymbol{x}\|)$, $\boldsymbol{x} \in \mathbb{R}^d$ *to MSE* $\leq \varepsilon$, *which can be easily achieved by a 3-layer network.*

- **Question:** Is this separation algorithmic?
  - Can GD + a 3-layer network learn this function?

# Our Results

**Theorem (Informal version of Theorem 2.1)**

*Same setting as in (Safran et al., 2019). There exists a 3-layer network s.t. for any input dim $d$ and target $\varepsilon$, we can choose layer widths $m_1 = \mathrm{poly}(d/\varepsilon)$, $m_2 = \Theta(1)$ so that, with probability $\geq 1 - 1/\mathrm{poly}(d/\varepsilon)$ over random initialization, running a simple variant of GF will reduce the MSE to $\varepsilon$ within $\mathrm{poly}(d/\varepsilon)$ time.*

- Main techniques/proof strategy:
  - A simple framework for multilayer mean-field networks,
    - where we can reason about multilayer networks with potentially infinitely many neurons.
  - Characterizing the infinite-width mean-field dynamics.
  - $\mathrm{poly}(d/\varepsilon)$-width discretization under symmetry.
    - In general, tracking the mean-field trajectory requires $\exp(d)$ neurons because of the compounding error.
- Comparison with (Safran & Lee, 2022): Different target/learner/techniques.

# Outline

1. 2-layer mean-field networks and our extension.
2. The infinite-width dynamics.
3. poly($d$)-width discretization under symmetry.

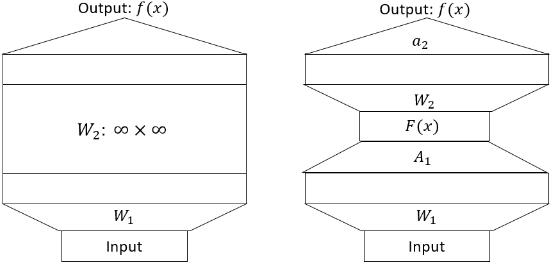
# Background: Mean-Field Networks

- Let $\mu$ be the empirical distribution of $\{\boldsymbol{w}_k\}_{k=1}^m$.

$$f(\boldsymbol{x}; \mathrm{W}) = \frac{1}{m} \sum_{k=1}^{m} \phi(\boldsymbol{x}; \boldsymbol{w}_k) = \int \phi(\boldsymbol{x}; \boldsymbol{w}) \, \mathrm{d}\mu(\boldsymbol{w}) =: f(\boldsymbol{x}; \mu).$$

- Allowing $\mu$ to be any (nice) distribution; Let $m \to \infty$.
  - $\Rightarrow$ 2-layer mean-field networks.
- Question: How to generalize this to $\geq 3$ layers (without introducing too much math)?

# Our Results: Multilayer Mean-Field Networks




- Main challenge: as width $\to \infty$, $\boldsymbol{W}_2$ becomes an $\infty \times \infty$ matrix.
  - Existing workarounds: distribution over functions, introducing an indexing set, ...
- Our solution: project the intermediate representations to $D$-dimensional vectors $\boldsymbol{F}(\boldsymbol{x})$ ($D < \infty$).
  - Reminiscent of the bottleneck structure from ResNets;
  - Much easier to use;
  - Retain the permutation invariant property of the neurons.

# The Infinite-Width Dynamics

- **Learner network:** (We choose the bottleneck dimension $D$ to be 1.)

$$\text{First layer:} \qquad F(\boldsymbol{x};\mu_1) = \mathop{\mathbb{E}}_{\boldsymbol{w}_1 \sim \mu_1} \|\boldsymbol{w}_1\| \operatorname{ReLU}(\boldsymbol{w}_1 \cdot \boldsymbol{x}),$$

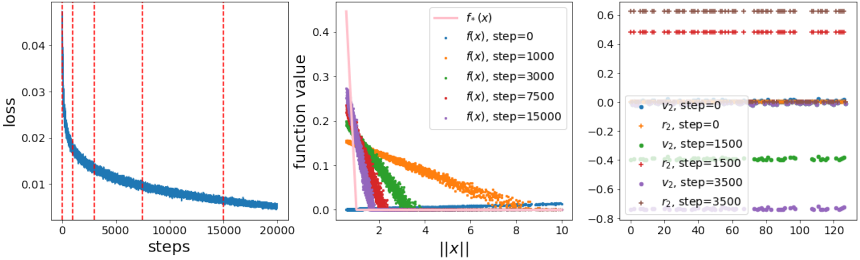
$$\text{Second layer:} \qquad f(\boldsymbol{x};\mu_2,\mu_1) = \mathop{\mathbb{E}}_{(w_2,b_2)\sim\mu_2} \operatorname{ReLU}(w_2 F(\boldsymbol{x};\mu_1) + b_2).$$

- **Lemma:** $\mu_{1,t}$ spherically symmetric $\Rightarrow F(\boldsymbol{x};\mu_{1,t}) = \alpha_t \|\boldsymbol{x}\|$, $\alpha_t \in \mathbb{R}_+$, where $\alpha_t$ depends only on $\mathbb{E}_{\boldsymbol{w}_1} \|\boldsymbol{w}_1\|^2$.

# The Infinite-Width Dynamics

- **Lemma:** $\mu_{1,t}$ spherically symmetric $\Rightarrow F(\boldsymbol{x}; \mu_{1,t}) = \alpha_t \|\boldsymbol{x}\|$, $\alpha_t \in \mathbb{R}_+$.
- **Facts/Claims:**
  1. The (infinite-width) initial distribution $\mu_{1,0}$, input distribution $\mathcal{D}$, and target function $f_*$ are all spherically symmetric.
  2. By symmetry, $\mu_{1,t}$ remains spherically symmetric for all $t \geq 0$.
  3. The dynamics of $\alpha_t$ depend on $\mu_{1,t}$ only through $\alpha_t$.
- $\Rightarrow$ The infinite-width dynamics of the first layer are simple! Only need to look at a single real number $\alpha_t$.
- **Claim:** GF + the infinite-width network will fit the target function.

# Finite-Width Simulation




- $f$ is always approximately spherically symmetric.
- $f$ eventually fits the target function $f_*(\boldsymbol{x}) = \mathrm{ReLU}(1 - \|\boldsymbol{x}\|)$.
- (The second layer behaves like a single neuron $(\bar{w}_2, \bar{b}_2)$.)
- **Observation:** The finite-width network closely tracks the infinite-width one (at least empirical).

# poly($d$)-width discretization

- **Main challenge:** Errors can compound; the discretization error can potentially grow exponentially fast; need $\exp(d)$ neurons to make the initial error exponentially small.
- **Observation:** The infinite-width network is a symmetrization of the finite-width network.
  - Any $\mu_{1,t}$ (not necessarily spherically symmetric),

$$\tilde{F}(\boldsymbol{x}; \mu_{1,t}) := \mathop{\mathbb{E}}_{\boldsymbol{x}' \in \|\boldsymbol{x}\| \mathbb{S}^{d-1}} F(\boldsymbol{x}') = \alpha_t \|\boldsymbol{x}\|.$$

- **Decomposition of the MSE loss:**

$$\mathcal{L} = \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} (f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))^2 + \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} (f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))^2 - \mathop{\mathbb{E}}_{\boldsymbol{x}} (f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))(f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))$$

  - Error of the infinite-width network
  - Discretization error: $\approx (\bar{w}_2^2/2) \, \mathbb{E}_{\boldsymbol{x}} (F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x}))^2$
  - $= 0$ as a result of symmetrization

# poly($d$)-width discretization (continued)

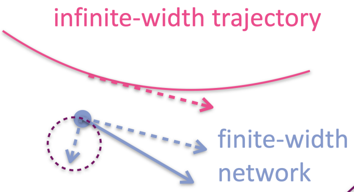
- Decomposition of the MSE loss:

$$\mathcal{L} \approx \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} (f_*(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x}))^2 + \frac{\bar{w}_2^2}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}} (F(\boldsymbol{x}) - \tilde{F}(\boldsymbol{x}))^2.$$

- **Claim:** The gradients of these two terms do not interfere with each other.
  - $\Rightarrow$ The second term ensures the discretization error does not grow. (No compounding errors!)
  - $\Rightarrow$ Only need the initial error to be inversely polynomially small. (Can be achieved using poly($d$) neurons.)

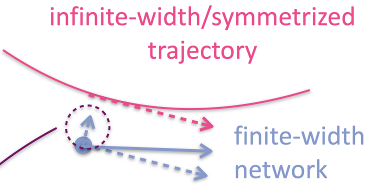
# poly($d$)-width discretization (the operational aspect)

**The General Case**




Compounding errors!

**Under Symmetry**




Pointing towards the infinite-width trajectory.
No compounding errors!

Q: How to show this?
A: Taylor expand the dynamics around the infinite-width trajectory. Look at the first-order terms.

# Conclusion

- Poster: #134, 4:30 PM - 6:30 PM (today)
- Takeaways:
  - The 3-layer vs 2-layer separation is algorithmic.
  - With symmetry, the infinite-width dynamics can be much simpler than the finite-width ones.
  - With symmetry, poly($d$)-width discretization is possible.
- Future directions:
  - General second-layer function.
  - Subspace version: $f_*(\boldsymbol{x}) = \mathrm{ReLU}(1 - \|\boldsymbol{A}^\top \boldsymbol{x}\|)$ where $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ is column orthogonal.
  - More generic poly($d$)-width discretization.