Mean-Field Analysis for Learning Subspace-Sparse Polynomials with Gaussian Input

Ziang Chen[1]

Joint work with Rong Ge

[1]Department of Mathematics, Massachusetts Institute of Technology

38th Conference on Neural Information Processing Systems
Vancouver, Canada

Problem Setup

- Subspace-sparse polynomial:
$$f^* : \mathbb{R}^d \to \mathbb{R}, \quad f^*(x) = h^*(x_V).$$
- $h^* : V \to \mathbb{R}$, where V is a subspace of $\mathbb{R}^d$ with $\dim(V) = p \ll d$.
- x_V is the orthogonal projection of x onto the subspace V.
- Two-layer neural networks:
$$f_{\mathrm{NN}}(x;\Theta) := \frac{1}{N}\sum_{i=1}^{N} \tau(x;\theta_i) = \frac{1}{N}\sum_{i=1}^{N} a_i\sigma(w_i^\top x).$$
- Loss function:
$$\min_{\Theta} \ \mathcal{E}_N(\Theta) := \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[|f^*(x) - f_{\mathrm{NN}}(x;\Theta)|^2\right].$$
- Stochastic gradient descent (SGD):
$$\theta_i^{(k+1)} = \theta_i^{(k)} + \gamma^{(k)}\left(f^*(x_k) - f_{\mathrm{NN}}(x_k;\Theta^{(k)})\right)\nabla_\theta\tau(x_k;\theta_i^{(k)}),$$
where $x_k,\ k = 1, 2, \ldots$ are the i.i.d. samples drawn from $\mathcal{N}(0, I_d)$.

Problem Setup

- Infinite-width two-layer neural network:

$$f_{\mathrm{NN}}(x;\rho) := \int \tau(x;\theta)\rho(d\theta) = \int a\sigma(w^\top x)\rho(da, dw).$$

- Loss functional:

$$\mathcal{E}(\rho) := \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[\left|f^*(x) - f_{\mathrm{NN}}(x;\rho)\right|^2\right].$$

- Mean-field dynamics of SGD:

$$\begin{cases} \partial_t \rho_t = \nabla_\theta \cdot \left(\rho_t \xi(t) \nabla_\theta \Phi(\theta;\rho_t)\right), \\ \rho_t\big|_{t=0} = \rho_0, \end{cases}$$

- **Question:** Can the mean-field dynamics of SGD learn a subspace-space polynomial within a finite time?

Necessary Condition

- Abbe et al. (2022): Merged-staircase property for polynomials on hypercubes, $h^*(z) = z_1 + z_1 z_2 + \cdots + z_1 z_2 \cdots z_p$.
- Related works in this direction: Abbe et al. (2023), Bietti et al. (2023), Dandi et al. (2023), Dandi et al. (2024), etc.
- **Reflective property:** For some subspace $S \subset V$,

$$\mathbb{E}_{z \sim \mathcal{N}(0, I_V)} \left[h^*(z) \sigma' \left(u + v^\top z_S^\perp \right) z_S \right] = 0, \quad \forall\, u \in \mathbb{R},\ v \in V.$$

- Characterize both the property of the target function and the expressiveness of the activation function.

Necessary Condition

- **Main theorem:** If the reflective property is satisfied for nontrivial subspace $S \subset V$, then for fixed T, we have for sufficiently large d that

$$\inf_{0 \le t \le T} \mathcal{E}(\rho_t) \ge C > 0,$$

  where C is a dimension-free constant.
- The dynamics cannot learn any information about f^* on S:

$$(\mathcal{P}_S)_\# \rho_0 = \delta_S \implies (\mathcal{P}_S)_\# \rho_t = \delta_S, \quad \forall\ t \ge 0.$$

- The flow ρ_t is always supported in $S^\perp$.

Sufficient Condition

- **Assumption:** The Taylor's expansion up to s-th order of the following flow $\hat{w}_V(t)$ at $t = 0$ is not contained in any proper subspace of V:

$$\begin{cases} \frac{d}{dt}\hat{w}_V(t) = \mathbb{E}_z\left[zh^*(z)\sigma'(\hat{w}_V(t)^\top z)\right], \\ \hat{w}_V(0) = 0. \end{cases}$$

- The assumption is true if $h^*(z) = c_1 z_1 + c_2 z_1 z_2 + \cdots + c_p z_1 z_2 \cdots z_p$ with nonzero $c_1, c_2, \ldots, c_p$.
- Training strategy: train the first layer for p times and take the average; then train the second layer.
- **Main theorem:** For some dimension-free constant $C_1, C_2 > 0$,

$$\mathcal{E}(\rho_t) \leq C_1 \exp(-C_2 t), \quad \forall\ t \geq 0.$$

The End

Thanks for your listening!