Learning sparse functions in the mean-field regime

Theodor Misiakiewicz (Stanford)

June 30th, 2022

Joint work with Emmanuel Abbe (EPFL) and Enric Boix-Adsera (MIT).

Learning: Optimization and Stochastics (Summer Research Institute 2022)

* *The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks*, Abbe, Boix-Adsera, Misiakiewicz, COLT 2022.

Online SGD on 2-layer Neural Network

2-layer neural network: M hidden units and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_j)_{j\in[M]} = (a_j, \boldsymbol{w}_j)_{j\in[M]} \in \mathbb{R}^{M(d+1)}$,

$$\boldsymbol{x} \in \mathbb{R}^d , \qquad \hat{f}_{\mathsf{NN}}(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{1}{M}\sum_{j\in[M]} \sigma_*(\boldsymbol{x};\boldsymbol{\theta}_j) = \frac{1}{M}\sum_{j\in[M]} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle) .$$

Goal: fit a target function f_* by minimizing

$$\min_{\boldsymbol{\Theta}} R(f_*, \boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\left(f_*(\boldsymbol{x}) - \hat{f}_{\mathsf{NN}}(\boldsymbol{x};\boldsymbol{\Theta})\right)^2\right] .$$

Online SGD:

- *Initialization:* $(\boldsymbol{\theta}_j)_{j\in[M]} \sim_{iid} \rho_0$.
- *Update:* at each step k, fresh sample $(\boldsymbol{x}_k, y_k)$ with $y_k = f_*(\boldsymbol{x}_k) + \varepsilon_k$,

$$\boldsymbol{\theta}_j^{k+1} = \boldsymbol{\theta}_j^k + \eta\left(y_k - \hat{f}_{\mathsf{NN}}(\boldsymbol{x}_k;\boldsymbol{\Theta}^k)\right) \cdot \nabla_{\boldsymbol{\theta}_j}\sigma_*(\boldsymbol{x}_k;\boldsymbol{\theta}_j^k) .$$

Mean-field approximation of the dynamics

[Mei,Montanari,Nguyen,'18], [Chizat,Bach,'18], [Rotskoff,Vanden-Eijnden,'18], [Sirignano,Spiliopoulos,'18]

- $M \to \infty$ limit: $(\boldsymbol{\theta}_j)_{j\in[M]}$ replaced by $\rho \in \mathcal{P}(\mathbb{R}^{d+1})$

$$\hat{f}_{\mathsf{NN}}(\boldsymbol{x};\boldsymbol{\Theta}) = \frac{1}{M}\sum_{j\in[M]} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)\,, \qquad \longrightarrow \qquad \hat{f}_{\mathsf{NN}}(\boldsymbol{x};\rho) = \int a\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\rho(\mathrm{d}\boldsymbol{\theta})\,.$$

- $\eta \to 0$ limit: gradient flow on the population loss, $(\rho_t)_{t\geq 0}$ solution of PDE with:

$$\boldsymbol{\theta}^t \sim \rho_t\,, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}^t = \mathbb{E}_{\boldsymbol{x}}\Big[\big(f_*(\boldsymbol{x}) - \hat{f}_{\mathsf{NN}}(\boldsymbol{x};\rho_t)\big)\nabla_{\boldsymbol{\theta}}\sigma_*(\boldsymbol{x};\boldsymbol{\theta}^t)\Big]\,.$$

Mean-field dynamics = gradient flow on population loss with $M = \infty$.

- [Mei,**M.**,Montanari,'19] with probability at least $1 - 1/M$:

$$\sup_{k\in[0,T/\eta]\cap\mathbb{N}} \big\|\hat{f}_{\mathsf{NN}}(\cdot;\boldsymbol{\Theta}^k) - \hat{f}_{\mathsf{NN}}(\cdot;\rho_{k\eta})\big\|_{L^2} \leq K e^{KT^3}\Bigg[\underbrace{\sqrt{\frac{\log(M)}{M}}}_{M\to\infty} + \underbrace{\sqrt{d\eta}}_{\eta\to 0}\Bigg]\,.$$

- Consider $\boldsymbol{x} \sim \mathrm{Unif}(\{+1,-1\}^d)$ and $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{r})$, $\boldsymbol{z} \in \mathbb{R}^P$, $\boldsymbol{r} \in \mathbb{R}^{d-P}$,

$$f_*(\boldsymbol{x}) = h_*(\boldsymbol{z}), \qquad \boldsymbol{z} \in \{+1,-1\}^P \text{ latent (unknown) support } (P \ll d).$$

- $a^0 \sim \mu_a$, $\boldsymbol{w}^0 \sim \mathrm{N}(0, \kappa^2 \boldsymbol{I}_d/d)$, and $\boldsymbol{w}^t = (\boldsymbol{u}^t, \boldsymbol{v}^t)$, $\boldsymbol{u}^t \in \mathbb{R}^P$ and $\boldsymbol{v}^t \in \mathbb{R}^{d-P}$.

$$\begin{aligned}\hat{f}_{\mathrm{NN}}(\boldsymbol{x}; \rho_t) &= \int a^t \sigma(\langle \boldsymbol{u}^t, \boldsymbol{z} \rangle + \langle \boldsymbol{v}^t, \boldsymbol{r} \rangle) \rho_t(\mathrm{d}\boldsymbol{\theta}^t) \\ &= \int a^t \mathbb{E}_{\boldsymbol{r}}[\sigma(\langle \boldsymbol{u}^t, \boldsymbol{z} \rangle + \langle \boldsymbol{v}^t, \boldsymbol{r} \rangle)] \rho_t(\mathrm{d}\boldsymbol{\theta}^t) =: \hat{f}_{\mathrm{NN}}(\boldsymbol{z}; \rho_t)\end{aligned}$$

- As $d \to \infty$ (P fixed),

$$\mathbb{E}_{\boldsymbol{r}}[\sigma(\langle \boldsymbol{u}^t, \boldsymbol{z} \rangle + \langle \boldsymbol{v}^t, \boldsymbol{r} \rangle)] \to \mathbb{E}_G[\sigma(\langle \boldsymbol{u}^t, \boldsymbol{z} \rangle + \|\boldsymbol{v}^t\|_2 G)] =: \sigma_{\|\boldsymbol{v}^t\|_2}(\langle \boldsymbol{u}^t, \boldsymbol{z} \rangle),$$

and $\boldsymbol{u}^0 \to 0$, $\|\boldsymbol{v}^0\| \to \kappa$.

Dimension-free dynamics

- As $d \to \infty$, $(a^t, \boldsymbol{u}^t, \boldsymbol{v}^t) \sim \rho_t$ approximated by $(\overline{a}^t, \overline{\boldsymbol{u}}^t, \overline{s}^t) \sim \overline{\rho}_t \in \mathcal{P}(\mathbb{R}^{P+2})$ where $\overline{\rho}_t$ follows the **DF-PDE dynamics:** learning $h_*(\boldsymbol{z})$ with gradient flow on the square loss with effective NN:

$$\hat{f}_{\text{NN}}(\boldsymbol{z}; \overline{\rho}_t) = \int \overline{a}^t \mathbb{E}_G[\sigma(\langle \overline{\boldsymbol{u}}^t, \boldsymbol{z} \rangle + \overline{s}^t G)] \overline{\rho}_t(\overline{\boldsymbol{\theta}}^t),$$

  from initialization $\overline{a}^0 \sim \mu_a$, $\overline{\boldsymbol{u}}^0 = 0$ and $\overline{s}^0 = \kappa$.

DF dynamics = MF dynamics when $d \to \infty$!

- With probability at least $1 - 1/M$:

$$\sup_{k \in [0, T/\eta] \cap \mathbb{N}} \left\| \hat{f}_{\text{NN}}(\cdot; \boldsymbol{\Theta}^k) - \hat{f}_{\text{NN}}(\cdot; \overline{\rho}_{k\eta}) \right\|_{L^2} \leq K e^{KT^7} \Bigg[\underbrace{\sqrt{\frac{P}{d}}}_{d \to \infty} + \underbrace{\sqrt{\frac{\log(M)}{M}}}_{M \to \infty} + \underbrace{\sqrt{d\eta}}_{\eta \to 0} \Bigg]$$

- If DF-PDE achieves $O(\varepsilon)$-test error in $T_* = T(h_*, \varepsilon)$, so does SGD w.h.p. when

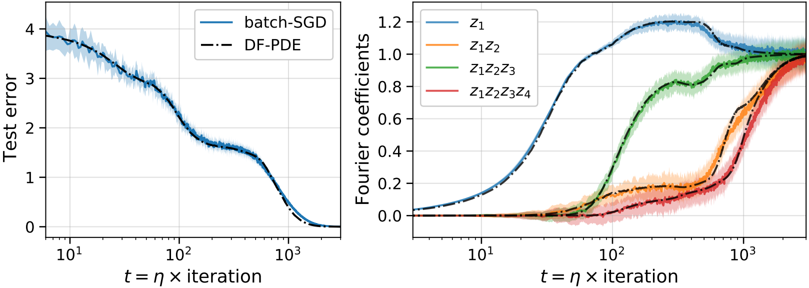
$$d \gtrsim C(T_*) P/\varepsilon, \qquad M \gtrsim C(T_*)/\varepsilon, \qquad \eta \lesssim d^{-1} \varepsilon / C(T_*),$$

  Number of online SGD iterations (# samples) $\approx C(T_*) d/\varepsilon = O_d(d)$.

Numerical illustration

$d = 100,\ M = 100$:

$$h_*(\boldsymbol{z}) = z_1 + z_1 z_2 + z_1 z_2 z_3 + z_1 z_2 z_3 z_4$$



Application: merged staircase functions

Fourier coefficient for $S \subseteq [P]$: $\hat{h}_*(S) = \mathbb{E}_{\boldsymbol{z}}\left[h_*(\boldsymbol{z})\chi_S(\boldsymbol{z})\right]$ where $\chi_S(\boldsymbol{z}) = \prod_{i \in S} z_i$.

$$h_*(\boldsymbol{z}) = \sum_{S \in \mathcal{Q}} \hat{h}_*(S)\chi_S(\boldsymbol{z}),$$

where $\mathcal{Q}$ contains all non-zero Fourier coefficients $\hat{h}_*(S) \neq 0$.

Merged-Staircase property (MSP)

$h_* : \{-1, +1\}^P \to \mathbb{R}$ has the *merged-staircase property* (MSP) if we can write elements of $\mathcal{Q}$ in order $(S_1, \ldots, S_r)$ such that for any $j \in [r]$, we have $|S_j \setminus (S_1 \cup \ldots \cup S_{j-1})| \leq 1$.

Examples of MSP functions:

$$h_*(\boldsymbol{z}) = z_1 + z_1 z_2 + z_1 z_2 z_3 + z_1 z_2 z_3 z_4,$$
$$h_*(\boldsymbol{z}) = z_1 + z_1 z_2 + z_2 z_3 + z_3 z_4 + z_3 z_4 z_5.$$

Examples of non-MSP functions:

$$h_*(\boldsymbol{z}) = z_1 + z_1 z_2 z_3 + z_1 z_2 z_3 z_4,$$
$$h_*(\boldsymbol{z}) = z_1 + z_1 z_2 + z_3 z_4 + z_3 z_4 z_5.$$

Tight characterization of learnability in this regime

Theorem [Abbe,Boix-Adsera,Misiakiewicz]

MSP is necessary and nearly sufficient* for DF-PDE to converge to zero test error**.

*Excludes a set of MSP fcts $h_*(\boldsymbol{z}) = \sum_{S \in \mathcal{Q}} h_*(S)\chi_S(\boldsymbol{z})$ with $\{h_*(S)\}_{S \in \mathcal{Q}}$ of measure 0.
(This is unavoidable: DF-PDE does not converge for some degenerate MSP)

**For sufficiency, train first then second layer (hard to directly analyse cv of PDEs)

$$\underbrace{h_*(\boldsymbol{z}) = z_1 + z_1 z_2}_{\substack{k = O_d(d) \text{ online SGD iterations is enough} \\ \text{In particular, } n = O_d(d) \text{ samples is enough}}} \qquad , \qquad \underbrace{h_*(\boldsymbol{z}) = z_1 z_2}_{\text{needs } k \gg d \text{ iterations}^{***}} \qquad .$$

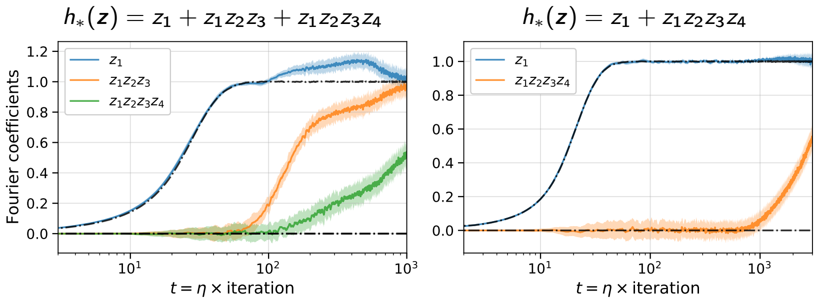
***Conjecture: $k = O_d(d\log(d))$ and more generally $k = \tilde{O}_d(d^{\ell-1})$ for leap-ℓ MSP.

Proposition [Abbe,Boix-Adsera,Misiakiewicz]

Any linear method require $n = \Omega_d(d^P)$ samples to learn $f_*(\boldsymbol{x}) = h_*(\boldsymbol{z})$ that contains the degree-P monomial.

Thank you!

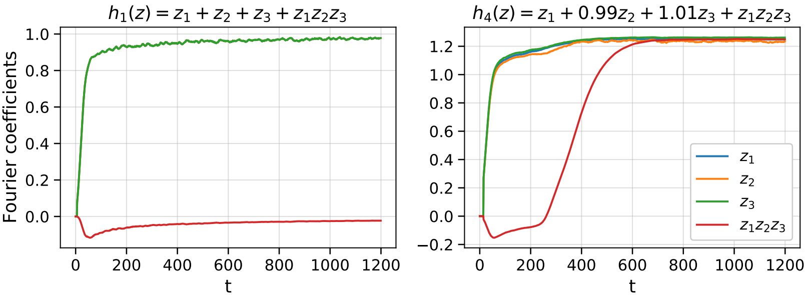
$d = 100, \ M = 100$:



DF-PDE approximation only valid for $T = O(1)$ (i.e., $n = O(d)$). For $T = \omega_d(1)$, online SGD escapes the saddle. This is an interesting regime for future work.

Degenerate MSP

$d = 100,\ M = 100$:



$h_*(\boldsymbol{z}) = z_1 + z_2 + z_3 + z_1z_2z_3$: we have $u_1^t = u_2^t = u_3^t$ during the dynamics.