Prev: dedicated algo.

Now: Generic. [DLS 22']

$\{(x_i, y_i)\}_{i=1..n}$ samples. $f^*(x) = g(\underset{\sim}{U}x)$. assume $\deg(f^*) = p$.

$\mathbb{R}^d \to \mathbb{R}^r$

$(d \gg r)$.

$y_i = f^*(x_i) + \varepsilon_i$.

$\varepsilon_i \sim \{-\zeta, \zeta\}$.

$f_\theta(x)$: Shallow NN. $f_\theta(x) = a^T \sigma(Wx+b) = \sum_{j=1}^{m} a_j \sigma(w_j \cdot x + b)$.

squared loss: $L_D(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (f_\theta(x) - f^*(x))^2$. $a_j = \{-1, 1\}$.

$w_j \sim N(0, \frac{1}{d} I_d)$.

Kernel regime: $n \asymp d^p$. [GMMM 19']

This paper: $n \asymp d^2 r + d r^p$,

(Feature learning).

Idea: First grad. step (learn feature W).

$n \gtrsim O(d^2)$,

r directions $d^2 r$.

Remaining: learn a. $W^{(1)}$ depends on (x, y).

a and $W^{(1)}$ not independent.

$\underset{\smile}{d} \underset{\sim}{r}^p$, d can be removed by resampling. (transfer learn setting.

subspace

(similar to exhaustiveness of PhD estimator).

Assumptions: ① Non-degeneracy of $H = \mathbb{E}[\nabla^2 f^*(x)]$ $(l^* \leq 2)$

$\mathrm{rank}(H) = r$. $\mathrm{span}(H) = S^*$, denote $k = \frac{\|H^+\|}{\sqrt{r}}$.

condition #.

CSQ L.B. argument shows necessary.

② Symmetric.

$a_j = -a_{m-j}$. $w_j = w_{m-j}$. $b_j = b_{m-j}$.

s.t. $f_{\theta_0}(x) = 0$.

Study gradient:

$$\nabla_{w_j} L_D(\theta) = \mathbb{E}_{x\sim D}\left[2(\underbrace{f_\theta(x)}_{=0 \text{ by symmetric init.}} - f^*(x))\, \nabla_{w_j} f_\theta(x) \right]$$

$$= -2\, \mathbb{E}_{x\sim D}\left[f^*(x)\, \nabla_{w_j} f_\theta(x) \right].$$ Recall $\sum_{j=1}^{m} a_j \sigma(w_j \cdot x + b)$.

$$= -2a_j\, \mathbb{E}_{x\sim D}\left[f^*(x)\, x\, \sigma'(w_j \cdot x) \right].$$

$$= \mathbb{E}_{x\sim D}\left[\underbrace{\nabla f^*(x)}_{\text{Stein's Lemma}}\, \sigma'(w\cdot x) + w f^*(x)\, \sigma''(w\cdot x) \right].$$

Take Hermite expansion over w.

$$f^*(x) = \sum_{k=0}^{p} \frac{\langle C_k, He_k(x)\rangle}{k!}$$ define $C_k = \mathbb{E}_x[\nabla^k f^*(x)]$. $C_2 = H$.

$$\sigma'(x) = \sum_{k\geq 0} \frac{c_k}{k!} He_k(x).$$ c_k: Hermite coeff. of $\sigma'(x) = \mathbb{1}_{x\geq 0}$.

$c_1 = \frac{1}{2}$, $c_2 = \frac{1}{\sqrt{2\pi}}$.

$$= \sum_{k=0}^{p-1} \frac{c_{k+1}\, \mathbb{E}_x[\nabla_x^{k+1} f^*(x)]\,(w^{\otimes k})}{k!} + w \sum_{k=0}^{p} \frac{c_{k+2}\, \mathbb{E}_x[\nabla_x^k f^*(x)]\,(w^{\otimes k})}{k!}$$

$$= \sum_{k=0}^{p-1} \frac{1}{k!}\left[c_{k+1} C_{k+1}(w^{\otimes k}) + c_{k+2}\, w\, (w^{\otimes k}) \right].$$

$$= \underbrace{\frac{Hw}{\sqrt{2\pi}}}_{O(d^{-1/2})} + \frac{1}{2}\underbrace{\left[c_3 C_3(w,w) + c_4 w\, C_2(w,w) \right]}_{O(d^{-1})} + \frac{1}{6}\underbrace{[\cdots]}_{O(d^{-3/2})} + \underbrace{\cdots}_{\text{higher-order.}}$$

Note: $\| C_{k+1} w^{\otimes k} \|.\ \| C_k w^{\otimes k} \| = O(d^{-k/2})$.

$= O(d^{-1/2})$.

$$\| \nabla_{w_j} \hat{L}_D(\theta) - \nabla_{w_j} L_D(\theta) | \lesssim \sqrt{\frac{d}{n}}.$$

$$d^{-1/2} \gtrsim \sqrt{\frac{d}{n}} \Rightarrow n \gtrsim d^2.$$

Algorithm 1: Gradient-based training

Input : Learning rates η_t, weight decay λ_t, number of steps T

preprocess data

$\alpha \leftarrow \frac{1}{n}\sum_{i=1}^{n} y_i$, $\beta \leftarrow \frac{1}{n}\sum_{i=1}^{n} y_i x_i$ Normalizing y.

$y_i \leftarrow y_i - \alpha - \beta \cdot x_i$ for $i = 1, \ldots, n$

end

$W^{(1)} \leftarrow W^{(0)} - \eta_1[\nabla_W \mathcal{L}(\theta) + \lambda_1 W]$

re-initialize $b_j \sim N(0,1)$ $\lambda_1 = \frac{1}{\eta_1}$.

for $t = 2$ **to** T **do**

$a^{(t)} \leftarrow a^{(t-1)} - \eta_t[\nabla_a \mathcal{L}(\theta^{(t-1)}) + \lambda_t a^{(t-1)}]$

end

return Prediction function $x \rightarrow \alpha + \beta \cdot x + a^T \sigma(Wx + b)$

$W^{(1)} = -\eta_1 \nabla_{W_0} \mathcal{L}(\theta)$.

dominated by first order $\frac{Hw}{\sqrt{2\pi}}$.

dominated by Hw. ($w \in S^{d-1}$).

$\subset S^*$. ($\because$ span$(H) = S^*$).

In S^* where rank$(S^*) = r$, choose parameter such as k),

repeat linearized regime $n \gtrsim r^p$, width $m \gtrsim r^p$.

Remaining steps (learn a). build ReLU to approx. x^k $\Rightarrow$ approx. high-dim. make use of $w^{(1)}$.

1. Construct a^* s.t. $\mathcal{L}(a^*, w^{(1)}, b) \ll 1$. $\|a^*\| = \tilde{O}\left(\frac{r^p k^{2p}}{\sqrt{m}}\right)$.

2. $\exists \lambda > 0$ s.t. $\mathcal{L}(a^{(T)}, w^{(1)}, b) \ll 1$, $\|a^{(T)}\| \leq \|a^*\|$ where $T = \tilde{\theta}(\eta^{-1}\lambda^{-1})$

3. Rademacher Gen. Bound. (standard). $\tilde{O}\left(\sqrt{\frac{d r^p k^{2p}}{n}} + \sqrt{\frac{r^p k^{2p}}{m}} + \frac{1}{n^{1/4}}\right)$.

$n \gtrsim d r^p$ $m \gtrsim r^p$. uniform conv.

$n \gtrsim \underbrace{d^2}_{\text{first grad. step}} \overset{r \text{ directions.}}{r} + \boxed{d} r^p$. $\leftarrow$ after 1 step.

$n \gtrsim d^2 r$. (first iteration).

$m \gtrsim r^p$.

Lower bound: (Convexity of span(H) ∋ f^*).

Construct $\mathcal{F}_p$ of poly. with deg p. each function depends on single relevant direction.

not satisfying assumption 2. q queries, $(r=1)$.

tolerence $\tau \leq \frac{\log^{p/4}(qd)}{d^{p/4}}$.

to output $f \in \mathcal{F}_p$ with loss ≤ 1.

$$\boxed{\tau \asymp \frac{1}{\sqrt{n}} \Rightarrow n \gtrsim d^{p/2},\quad q}$$

Pf: Recall SQ $q(x,y) \to$ output $\hat{q}$ with $|\hat{q} - \mathbb{E}_{x,y}\, q(x,y)| \leq \tau$.

CSQ $q(x,y) = \underbrace{y\,h(x)}_{\text{correlation}}$.

$$\nabla_{w_j} L_D(\theta) = \mathbb{E}_{x,y}[y\,h(x)] \text{ where } h(x) = -2a_j\, x\, \sigma'(w_j \cdot x).$$

Idea: Construct function class with small pairwise correlations.

Lemma: $\mathbb{E}_{x\sim D}[f(x)^2] = 1$. $\left|\mathbb{E}_{x\sim D}[f(x)g(x)]\right| \leq \varepsilon$. $\forall f,g \in \mathcal{F}$. $f \neq g$.

$\Rightarrow$ Require $q \geq \frac{|\mathcal{F}|(\tau^2-\varepsilon)}{2}$ queries to have loss $\leq 2-2\varepsilon$.

How to find such $\mathcal{F}$?

Fact: $\exists\, e^{c\varepsilon^2 d}$ vectors on $\mathbb{R}^d$ s.t. their inner product $\lesssim \varepsilon$.

$$f_u(x) = \frac{He^{p}(u\cdot x)}{\sqrt{p!}}.$$ Construct

$$\mathbb{E}_{x\sim D}[f_u(x)\cdot f_v(x)] = (u\cdot v).$$

$\therefore |u\cdot v| \lesssim \varepsilon \quad \Rightarrow \left|\mathbb{E}[f_u(x)\cdot f_v(x)]\right| \lesssim \varepsilon^p$.

By Lemma, $q \geq \frac{e^{c\varepsilon^2 d}(\tau^2 - \varepsilon^p)}{2}$.

$$e^{c\varepsilon^2 d} \leq \frac{2q}{\tau^2 - \varepsilon^p},$$

take $\varepsilon = \sqrt{\dfrac{\log(2q(cd)^{p/2})}{cd}}$,

$$\tau^2 \lesssim \frac{\log^{p/2}(qd)}{d^{p/2}}.$$

Last time, Kernel $n \asymp d^p$. $\quad$ ($f^*(x) = g(Lx)$ $\quad L: \mathbb{R}^d \to \mathbb{R}^r$ ($d \gg r$) . $\deg(f^*) = p$.

[DWS 22] One grad. step + learn second layer.

$n \asymp d^2 r + d r^p$.

Today [BES+22]. how one grad. step improves rep. [Based on observation "non-kernel" behavior often occurs in early phase, especially in large lr [JGF+20, FDP+20]

High-dim Regime. (Follow not. from [BES+22].

Asymptotic: data size n, input dimension d. NN width $N \to \infty$,

$n/d \to \psi_1$. $N/d \to \psi_2$. $\quad \psi_1 \uparrow \to$ sample size $\uparrow$ $\quad$ $\psi_2 \uparrow \to$ NN width $\uparrow$. $\quad$ n. N comparable. so $N \not\gg n$. in NTK.

$$f(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \, \sigma(\langle x, w_i \rangle),$$

$$= \frac{1}{\sqrt{N}} \underset{1\times N}{\underset{\sim}{a}}^T \sigma(\underset{N\times d}{\underset{\sim}{W}}^T \underset{d\times 1}{\underset{\sim}{x}}) \text{ . with squared loss.}$$

$x \in \mathbb{R}^d$,

Limitations of prev methods:

Build kernel on $x \mapsto \sigma(W_0^T x)$: conjugate kernel.

RF Method (CK, NTK). Rotation invariant kernel.

[EK10. HC20. MZ20. BMR21].

$$\inf_{\lambda > 0} R_{RF}(\lambda) \gtrsim \underbrace{\| P_{>1} f^* \|_{L^2}^2}_{} + o_d(1),$$

projection to deg > 1 poly,

cannot perform better than linear method.

This paper: Small lr $\eta = \Theta(1)$. better than RF. still linear regime.

Large lr $\eta = \Theta(\sqrt{N})$. risk $< \| P_{>1} f^* \|_{L^2}^2$.

Assumption: Recall $f(x) = \frac{1}{\sqrt{N}} a^T \sigma(W^T x)$.

① Init: $\sqrt{d} \cdot [W_0]_{ij} \overset{iid}{\sim} N(0,1)$,

$\sqrt{N} \cdot [a]_j \overset{iid}{\sim} N(0,1)$.

$\Rightarrow \frac{1}{n}$ Mean-field scaling.

② $y_i = f^*(x) + \varepsilon_i$. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

f^*: Lipschitz. $\|f^*\|_{L^2} = \Theta_d(1)$.

In particular, we study single-index.

$$f^*(x) = \sigma^*(\langle x, \beta_* \rangle), \ \beta_* \in \mathbb{R}^d,$$

$$(\text{where } \mathbb{E}[z\sigma^*(z)] \neq 0 \Rightarrow \ell^* = 1)$$

③ σ: $\mathbb{E}[\sigma(z)] = 0$. $\mathbb{E}[z\sigma(z)] \neq 0$.

first three deriv. bounded a.s.

This paper:

First step: learn W. (Goal: Gradient matrix is ≈ rank-1, which contains info. of labels y).

$$W_{t+1} = W_t - \eta \frac{\partial L}{\partial W_t} = W_t + \eta \sqrt{N} G_t,$$

$$\text{where } \underset{d \times N}{G_t} = \frac{1}{n} \underset{d \times n}{X^T} \Big[\Big(\frac{1}{\sqrt{N}} \underbrace{\big(y - \tfrac{1}{\sqrt{N}} \sigma(XW_t a)\big)}_{n \times 1} \underbrace{a^T}_{1 \times N} \Big) \odot \underbrace{\sigma'(XW_t)}_{n \times N} \Big]$$

(neglect derivation).

Orthogonal decomposition of σ:

$\sigma(z) = \mu_1 z + \sigma_\perp(z)$, where $\mu_1 = \mathbb{E}[z\sigma(z)] \neq 0$,

$\mathbb{E}[\sigma_\perp(z)] = \mathbb{E}[z\sigma_\perp(z)] = 0$.

$\mathbb{E}[\sigma_\perp(z)^2] = \mu_2^2$. where $\mu_2 = \sqrt{\mathbb{E}[\sigma(z)^2] - \mu_1^2}$, Denote $G_1 := \frac{1}{\eta\sqrt{N}}(W_1 - W_0)$. rank-1 matrix $A := \frac{\mu_1}{n\sqrt{N}} X^T \underline{y} a^T$.

w.h.p. $\| G_0 - A \| \leq \frac{C\log^2 n}{\sqrt{n}} \| G_0 \|$.

(Depends on the decomposition above. App. B.1.1).

Expect: learnt y. W_1 should have alignment with f^*.

$f^*(x) = \mu_0^* + \mu_1^* \langle x, \beta^* \rangle + P_{>1} f^*(x)$

denote $\| P_{>1} f^* \|_{L^2} \to \mu_2^*$ as $d \to \infty$.

Learning rate regime:

$\sqrt{d} \cdot [W_0]_{ij} \overset{iid}{\sim} N(0,1)$.

$\|W_0\| = \Theta_d(1)$. $\|W_0\|_F = \Theta(\sqrt{d})$.

Small lr: $\eta = \Theta(1)$ $\Rightarrow \|W_1 - W_0\| \asymp \|W_0\|$

Large lr: $\eta = \Theta(\underline{\sqrt{N}})$ $\Rightarrow \|W_1 - W_0\|_F \asymp \|W_0\|_F$.

(Adhere to μP-scaling).

N is the width.

Thm 3: Can derive asymptotic limit of leading singular value $s_1(W_1)$ & $\underset{lr}{\underline{\eta}}$.

(when $\eta = \Theta(1)$)

corr. left sing. vector u_1

$|\langle u_1, \beta_* \rangle|^2$,

$\forall i > 1$, $| s_i(W_1) - s_i(W_0) | = o_d(1)$.

spiked model, (Figure 3. of paper). bulk doesn't change.

Book: RMT4ML. Liao et al. Ch 2.5, 2.6.

- After first step, obtain feature map $x \mapsto \sigma(W_1^T x)$.

train conjugate kernel on top.

similar to [BES22].

Remaining step on a:

Do ridge regression on a with fresh samples $(\tilde{X}, \tilde{y})$.

$\because W_1$ correlates with (X, y).

$$\hat{a} = \arg\min_a \left\{ \frac{1}{n} \| \tilde{y} - \Phi a \|^2 + \frac{\lambda}{N} \|a\|^2 \right\}.$$

where $\Phi = \frac{1}{\sqrt{N}} \sigma(\tilde{X} W_1)$.

$n \times N$.

[HL20, MS22]

- Gaussian Equivalence property. $(\eta = \Theta(1))$ precise asymptotics.

$R_{CK}(\lambda) \simeq R_{GE}(\lambda)$

for $\phi_{GE}(x) = \frac{1}{\sqrt{N}} (\mu_1 W^T x + \mu_2 \underset{\sim}{z})$.

$z \sim N(0,1)$,

nonlinear kernel behaves like noisy linear method.

Can be computed explicitly.

Conclusion: Can improve over initial CK.

but $R_{GE}(\lambda) \geq \| P_{>1} f^* \|_{L^2}^2$.

$\eta = \Theta(\sqrt{N})$. (Cannot derive asymptotic behavior).

Given W_1. construct second layer $\tilde{a}$ s.t. (Exist good sol. $\tilde{a}$).

$\tilde{f}(x) = \frac{1}{\sqrt{N}} \tilde{a}^T \sigma(W_1^T x)$ has risk $\simeq K^*$.

$$\text{where } K^* \triangleq \inf_{K \in \mathbb{R}} \mathbb{E}\Big[\big(\underbrace{\sigma^*(\xi_1)}_{\text{teacher}} - \underbrace{\mathbb{E}_{\xi_2} \sigma(K\xi_1 + \xi_2)}_{\text{student}}\big)^2\Big]$$

(?)

$\xi_1, \xi_2 \sim N(0,1)$.

For suff. large $\psi_1 \triangleq \frac{n}{d}$. $R_1(\lambda) \le 10k^* + C(\sqrt{k^*} \cdot \sqrt{\frac{d}{n}} + \frac{d}{n})$.

ridge estimator

w.p. 1 when $n, d, N \to \infty$.

ridge penalty $n^{\varepsilon - 1} < N^{-1}\lambda < n^{-\varepsilon}$

for some $\varepsilon > 0$.

If $\|P_{>1} f^*\|_{L^2} > 10k^*$, then it outperforms linear method.

e.g.: $\sigma = \sigma^* = \tanh$.