Maths of Deep Learning

Recitation #5

- Search and Descent phase in single-index models [Ben Arous et al. '21. JMLR]

$$\boxed{\text{almost all of the data is used simply in the initial search phase, except for the simplest tasks}}$$

- Simplest task: final convergence and overparametrization [Xu & Du, '23 COLT]

$$\boxed{\text{over-paramtrization may slow down final convergence exponentially}}$$

(All notations follow original papers)

- Online stochastic gradient desent on non-convex losses from high-dimension inference

Gerard Ben Arous, Rena Ghesissari, Aukosh Jagannath

- Setting: target to estimate: $\theta_N \in S^{N-1}$. N: dimension

data distribution: $\mathbb{P}_\theta = P_{\theta_N}$

data : $M = \alpha_N \cdot N$ i.i.d. samples $(Y^\ell)_{i=1}^M \subseteq \mathbb{R}^D$ from $\mathbb{P}_N$.

loss function: $L_N : S^{N-1} \times \mathbb{R}^D \to \mathbb{R}$

population loss:

$$\Phi_N(x) \triangleq \mathbb{E}_{Y \sim \mathbb{P}_\theta}[L_N(x; Y)] = \phi(m_N(x))$$

parameter $x \in S^{N-1}$

with $\phi : [-1,1] \to \mathbb{R}$. $m_N(x) = \langle x, \theta_N \rangle$

"correlation of x with θ_N"

– Weak recovery:

a sequence of estimators $\hat{\theta}_N \in S^{N-1}$ weakly recovers the parameter θ_N if for some $\eta > 0$.

$$\lim_{N \to \infty} P(m_N(\hat{\theta}_N) \geq \eta) = 1.$$

recall: if $\hat{\theta}_N$ is drawn uniformly at random, then $\langle \hat{\theta}_N, \theta_N \rangle \simeq N^{-1/2}$

– Strong recovery:

$\forall \eta > 0$,

$$\lim_{N \to \infty} P(m_N(\hat{\theta}_N) < 1 - \eta) = 0$$

– Algorithm: online SGD on sphere

$X_0 = x_0$ learning rate single sample "online SGD"

$$\tilde{X}_t = X_{t-1} - \frac{\delta}{N} \nabla L_N(X_{t-1}; Y^t)$$

$$X_t = \frac{\tilde{X}_t}{\|\tilde{X}_t\|}$$

spherical gradient

$$= \underbrace{\nabla_x L_N(X_{t-1}; Y^t)}_{\text{on } \mathbb{R}^N} - \langle \nabla_x L_N(X_{t-1}; Y^t), X_{t-1} \rangle X_{t-1}$$

remark: online SGD $\Rightarrow$ M i.i.d. samples from $\mathbb{P}_N$ stands for M steps

– Assumptions:

(A) ϕ is differentiable and ϕ' is strictlyly negative in $(0,1)$

— Information exponent:

Def: population loss $\Phi_N(x) = \phi(m_N(x))$ has information exponent k

if $\phi \in C^{k+1}([-1,+1])$ and there exists $C, c > 0$ s.t.

$$\begin{cases} \dfrac{d^\ell \phi}{dm^\ell}(0) = 0, & 1 \le \ell < k \\ \dfrac{d^k \phi}{dm^k}(0) \le -c < 0 & \longrightarrow \text{Assumption A: } \phi' < 0 \\ \left\| \dfrac{d^{k+1}\phi}{dm^{k+1}}(m) \right\|_\infty \le C \end{cases}$$

Def: recall $M = \alpha_N \cdot N$

Define $\alpha_c(N,k) = \begin{cases} 1, & k=1 \\ \log N, & k=2 \\ N^{k-2}, & k \ge 3 \end{cases}$

— Main result: Strong recovery in M steps

if $\begin{cases} (k=1) & \alpha_N = \frac{M}{N} \gg \alpha_c(N,1) \\ (k=2) & \alpha_N \gg \alpha_c(N,2) \cdot \log N \\ (k \ge 3) & \alpha_N \gg \alpha_c(N,k) \cdot (\log N)^2 \end{cases}$

Remark: Sample complexity of strong recovery is always at most polynomial

— Main result: weak recovery

if $\alpha_N \ll \alpha_c(N,k)$

then $\sup_{t \to M} |m_N(X_t)| \to 0$ in probability, and in L^p for any $p \ge 1$.

(lower bound of weak recovery)

Remark: lower bound of weak recovery

$\Rightarrow \alpha_c(N,k)$ is optimal up to $O((\log N)^2)$.

— Main result: search phase v.s. descent phase

Def: $\tau_\eta^+ = \inf\{t \mid m_N(X_t) > \eta\}$ end of search phase

$\tau_{1-\eta}^+ = \inf\{t \mid m_N(X_t) > 1-\eta\}$ end of descent phase

Theorem:

For $k \geq 2$, $\forall \eta > 0$, $\exists$ const $C = C(k,\eta) > 0$ s.t.

$$\tau_\eta^+ \gg \alpha_c(N,k)$$

$$|\tau_{1-\eta}^+ - \tau_\eta^+| \leq C \cdot N$$

with probability $1 - o(1)$.

Remark: this implies

$$\frac{\#\text{samples used in descent phase}}{\#\text{samples used in search phase}} = \frac{1}{\alpha_c(N,k)}$$ vanishes for $k \geq 2$.

Intuition behind $\alpha_c(N,k)$:

Consider GD for the population loss:

for small m_{t-1} and some $c > 0$

$$m_t = m_{t-1} - \frac{\delta}{N} \langle \phi'(m_{t-1}) \nabla m_{t-1}, \nabla m_{t-1} \rangle$$
$$= m_{t-1} - \frac{\delta}{N} \phi'(m_{t-1}) \cdot \|\nabla m_{t-1}\|^2$$
$$\approx m_{t-1} + \frac{\delta}{N} C \cdot m_{t-1}^{k-1}$$

assuming $\phi' < 0$

initialization $m_0 \simeq N^{-\frac{1}{2}}$ to achieve $m_T \geq \eta$

① $k=1$:

$$m_t \approx m_{t-1} + \frac{\delta}{N} \cdot C$$

$$\Rightarrow T \simeq \frac{N}{\delta}$$

② $k=2$:

$$m_t \approx m_{t-1} + \frac{\delta}{N}\cdot C \cdot m_{t-1} = \left(1+\frac{\delta}{N}C\right)\cdot m_{t-1}$$

$$\Rightarrow T \simeq \frac{\log(\eta / N^{-\frac{1}{2}})}{\log\left(1+\frac{\delta}{N}C\right)} \approx \frac{N}{\delta}\cdot \log N$$

③ $k \geq 3$:

$$m_t \approx m_{t-1} + \frac{\delta}{N} C \cdot m_{t-1}^{k-1}$$

ODE to estimate T:

$$\dot{m} = \frac{\delta}{N} C \cdot m^{k-1}$$

$$dm^{-k+2} = \frac{\delta}{N} C\, dt$$

$$T \simeq \frac{N}{\delta} N^{\frac{1}{2}(k-2)} \quad \text{as } m_0 \simeq N^{-\frac{1}{2}}$$

Over-parametrization Exponentially Slows Down Gradient Descent for Learning a Single Neuron

Weihang Xu, Simon S. Du

– Setting: target to estimate: $v \in \mathbb{R}^d$

model: $f(x; w) = \sum_{i=1}^{n} \mathrm{Relu}(\langle w_i, x\rangle)$

n: # neurons

student neurons

label: $g: \mathbb{R}^d \to \mathbb{R}$

$x \to \mathrm{relu}(\langle v, x\rangle)$

teacher neuron

loss: $L(w) = \mathbb{E}_{x\sim N(0,I)}\left[\frac{1}{2}\left(f(x;w) - g(x)\right)^2\right]$ population loss

Algorithm: GD on population loss

① exact parametrization [Yehudai and Shamir '20]

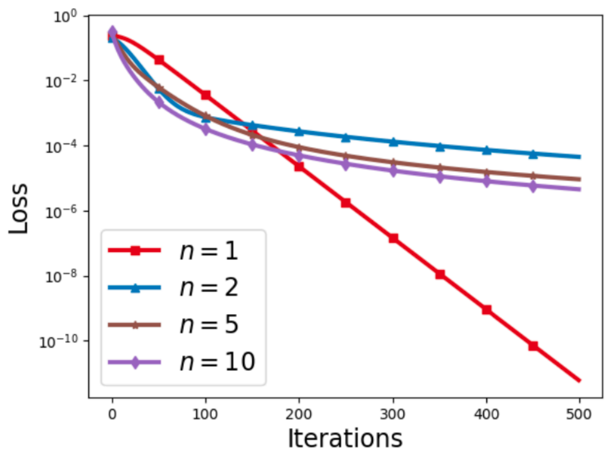
$(n=1)$

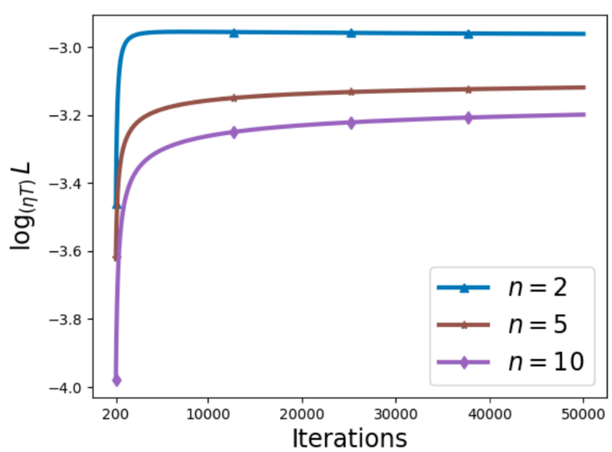
$$L(w(t)) \leqslant \exp(-\Omega(t))$$

② over-parametrization [Xu & Du]

$(n \geq 2)$

$$L(w(t)) = \Theta(T^{-3})$$





Upper bound: $L(w(t)) \leqslant O(T^{-3})$

denote θ_i = angle between w_i and v

Lemma: $L(w) \geq \frac{1}{30\pi} \|w_i\|^2 \cdot \theta_i^3$. $\forall i$

$L(w) = \Omega(\theta^3)$ implies

L is lower bounded by a cubic function of θ

when w is close to global minimizer of L

$\sum_{i=1}^{n} w_i \approx v$

$\exists\, i \in [n],\ \theta_i \neq 0$

Then, optimizing $\Omega(\theta^3)$ around $\theta \approx 0$ gives the upper bound

- Remark:

  θ^3 also implies a risk of slow convergence as

  - $|\theta|^3$ around $\theta=0$ is convex but not strongly convex

    so gradient flow does not have a guarantee

    of $L(w(t)) = \exp(-\Omega(t))$.

  - If the lower bound is stronger as

    $L(w) \gtrsim \Omega(\theta^2)$. maybe strong convexity gives

    $$L(w(t)) = \exp(-\Omega(t))$$

    (this does not hold for this problem)

Lower bound: $L(w(T)) \gtrsim \Omega(T^{-3})$

motivating examples:

① teacher direction is learnt:

$w_1 = \lambda_1 v_1,\ w_2 = \lambda_2 v_2,\ \cdots,\ w_n = \lambda_n v_n$.

then $\nabla_{w_i} L = \frac{1}{2}\left(\sum_j w_j - v\right) = \frac{1}{2}\left(\sum_j \lambda_j - 1\right) v$

$\Rightarrow \sum_j \lambda_j - 1 \to 0$ exponentially

$\Rightarrow L(w(t)) = \exp(-\Omega(t))$

② student neurons are aligned:

$w_1 = w_2 = \cdots = w_n$

simple computation will give

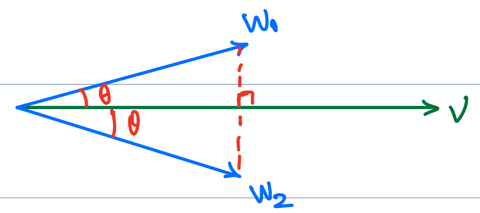
gradient of each neuron

$= n \times$ gradient of single-neuron case $(n=1)$

$\Rightarrow L(w(t)) \approx \exp(-\Omega(t))$

③ teacher direction is not learnt perfectly:

consider symmetric case with $n=2$:



$w_1(0) = \lambda_1(0)\, v + \lambda_2(0)\, v^\perp, \quad w_2(0) = \lambda_1(0)\, v - \lambda_2(0)\, v^\perp.$

- easy to see the symmetry always holds $\forall t \geq 0$

$$\begin{cases} w_1(t) = \lambda_1(t)\, v + \lambda_2(t)\, v^\perp \\ w_2(t) = \lambda_1(t)\, v - \lambda_2(t)\, v^\perp \end{cases}$$

GD gives

$$\begin{cases} \lambda_1(t+1) - \frac{1}{2} = \left(\lambda_1(t) - \frac{1}{2}\right)\left(1 - \eta\left(1 - \frac{\theta(t)}{\pi} + \frac{\sin 2\theta(t)}{\lambda_1(t)}\right)\right) & (*) \\ \lambda_2(t+1) = \lambda_2(t) \cdot \left(1 - \frac{\eta}{2\pi}\left(2\theta + \frac{\lambda_1 - \frac{1}{2}}{\lambda_1} \sin 2\theta\right)\right) & (**) \end{cases}$$

When $\theta = o(1)$,

$(*)$ implies $\lambda_1(t+1) - \frac{1}{2} \approx \left(\lambda_1(t) - \frac{1}{2}\right) \cdot (1-\eta)$

$\Rightarrow \lambda_1$ converges to $\frac{1}{2}$ exponentially

$(**)$ can be re-written as, with $\lambda_1 - \frac{1}{2} = o(1)$,

$$\lambda_2(t+1) \approx \lambda_2(t) \cdot \left(1 - \frac{2\eta}{\pi} \lambda_2(t)\right)$$

$\Rightarrow \lambda_2$ converges to 0 with rate $\lambda_2(t) \sim t^{-1}$

$\Rightarrow L(w(t)) \approx t^{-3}.$

This means the final convergence is $L(w(t)) \geq \Omega(t^{-3})$
due to the slow moving orthorgonal to v.