DLT Homework 2

20200130 Yujin Kim

#2-1

f: L-smooth, convex $\quad x_{t+1} = x_t - \eta_t \nabla f(x_t) = x_t - \frac{1}{L}\nabla f(x_t)$.

(Step 1)

$$\|x_{t+1} - x^*\|^2 = \|x_t - \tfrac{1}{L}\nabla f(x_t) - x^*\|^2$$
$$= \|x_t - x^*\|^2 - \tfrac{2}{L}\nabla f(x_t)^T(x_t - x^*) + \tfrac{1}{L^2}\|\nabla f(x_t)\|^2 \quad \cdots ①$$

(Step 2)

$$f(x^*) \le f(x - \tfrac{1}{L}\nabla f(x))$$
$$\le f(x) + \nabla f(x)^T(-\tfrac{1}{L}\nabla f(x)) + \tfrac{L}{2}\|-\tfrac{1}{L}\nabla f(x)\|^2 \quad (\text{By smoothness}).$$
$$= f(x) - \tfrac{1}{2L}\|\nabla f(x)\|^2. \quad \cdots ②$$

Descent Lemma

In particular, we have $f(x_{t+1}) \le f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2 \le f(x_t)$

and so $f(x_T) \le f(x_t)\ \forall t \le T \quad \cdots ③$

(Step 3)

Lemmas for convex & smooth fns

$g(y) = f(y) - \nabla f(x)^T y$ is L-smooth convex and has minimum at $y = x$.

By ②, we have $g(x) \le g(y) - \frac{1}{2L}\|\nabla g(y)\|^2$

$$\Rightarrow g(y) - g(x) = f(y) - f(x) - \nabla f(x)^T(y - x) \ge \tfrac{1}{2L}\|\nabla g(y)\|^2$$
$$= \tfrac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$$

Hence, $\nabla f(x)^T(y - x) \le f(y) - f(x) - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$

In particular, $\nabla f(x)(x^* - x) \le f(x^*) - f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \quad \cdots ④$

(Step 4)

Using ④ in ①, $\|x_{t+1} - x^*\| \le \|x_t - x^*\|^2 + \frac{2}{L}\left(f(x^*) - f(x_t) - \frac{1}{2L}\|\nabla f(x)\|^2\right)$
$$+ \tfrac{1}{L^2}\|\nabla f(x_t)\|^2$$
$$= \|x_t - x^*\|^2 + \tfrac{2}{L}(f(x^*) - f(x_t))$$
$$\le \|x_t - x^*\|^2 + \tfrac{2}{L}(f(x^*) - f(x_T)) \quad (\text{By } ③)$$

Convergence Rate

Sum this over $t = 0, \cdots, T-1$ to obtain $\|x_T - x^*\| \le \|x_0 - x^*\|^2 + \frac{2T}{L}(f(x^*) - f(x_T))$.

Thus, $f(x^T) - f(x^*) \le \frac{L}{2T}(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2)$
$$\le \tfrac{L}{2T}\|x_0 - x^*\|^2.$$

#2.2

(a). By definition of Θ, for $\theta \in \Theta$, $(x_i, f(x_i;\theta))$ is interpolated by a linear model, i.e. $\exists w_\theta, b_\theta$ s.t. $f(x_i;\theta) = \langle w_\theta, x_i \rangle + b_\theta \ \forall i$.

↳ (Because it is not affected by the nonlinearity of ReLU).

$\theta^* \in \Theta$. For any $\theta \in \Theta$,

$$\hat{R}(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\theta^*) - y_i)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(\langle w^*, x_i\rangle + b - y_i)^2$$
$$\le \frac{1}{n}\sum_{i=1}^{n}(\langle w_\theta, x_i\rangle + b_\theta - y_i)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\theta) - y_i)^2 = \hat{R}(\theta) \quad \cdots ①$$

As Θ is finite intersection of open sets $\underbrace{\{\theta \mid W_\ell a_{\ell-1}(x_i) + b_\ell > 0\}}$,

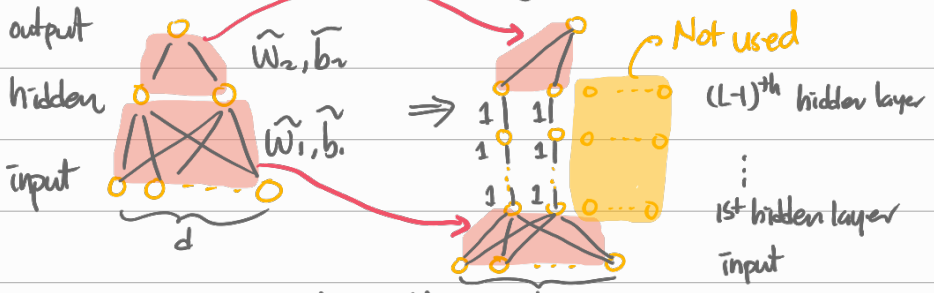
open for being inverse image of open set under a continuous map.

Θ is open. $\cdots ②$

By ①, ②, θ^* is a local minima of $\hat{R}$.

(b) Let $(\tilde{W}_1, \tilde{b}_1, \tilde{W}_2, \tilde{b}_2)$ be given as the problem.

Pictorial Construction



By the assumption $d_\ell \ge 2 \ \forall \ell$, We can let

Abstract Construction

① $(W_1)_{ij} = \begin{cases} (\tilde{W}_1)_{ij} & \text{for } i = 1,2 \\ 0 & \text{otherwise} \end{cases}$, $(b_1)_i = \begin{cases} (\tilde{b}_1)_i & \text{for } i = 1,2 \\ 0 & \text{otherwise} \end{cases}$

② $(W_\ell)_{ij} = \begin{cases} 1 & \text{if } i = j-1 \text{ or } i = j-2 \\ 0 & \text{otherwise.} \end{cases}$, $(b_\ell)_i = 0 \ \forall i$

③ $(W_L)_{ij} = \begin{cases} (\tilde{W}_2)_{ij} & \text{if } (i,j) = (1,1) \text{ or } (1,2) \\ 0 & \text{otherwise} \end{cases}$, $b_L = \tilde{b}_2$

#2.3

(a) $f_W(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j \sigma(w_j^T x)$

$f_{W_0+W\Sigma}(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j \sigma((w_j^0 + \Sigma_{jj} w_j)^T x)$

where $W_0 = [w_1^0 \cdots w_m^0]$, $W = [w_1 \cdots w_m]$

$\sigma((w_j^0 + \Sigma_{jj} w_j)^T x) \approx \sigma(w_j^{0T} x) + \sigma'(w_j^{0T} x)(\Sigma_{jj} w_j^T x) + \frac{\sigma''(w_j^{0T} x)}{2}(\Sigma_{jj} w_j^T x)^2$

Thus, $f_{W\Sigma}^L(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} s_j \sigma'(w_j^{0T} x)(\Sigma_{jj} w_j^T x)$

$= \frac{1}{\sqrt{m}} \langle s\sigma'(W_0^T x), \Sigma W^T x \rangle$

$\sigma'\left(\begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}\right) = \begin{bmatrix} \sigma'(v_1) \\ \vdots \\ \sigma'(v_m) \end{bmatrix}$

$\text{diag}\{s_1, \cdots, s_m\}$

and $f_{W\Sigma}^Q(x) = \frac{1}{2\sqrt{m}} \sum_{j=1}^{m} s_j \sigma''(w_j^{0T} x)(\Sigma_{jj} w_j^T x)^2$

$= \frac{1}{2\sqrt{m}} (\Sigma W^T x)^T H_0 (\Sigma W^T x)$

, for $H_0 = \text{diag}\{s_1 \sigma''(w_1^{0T} x), \cdots, s_m \sigma''(w_m^{0T} x)\}$

$= \text{diag}\{s\ \sigma''(W_0^T x)\}$

$\text{diag}\{s_1, \cdots, s_m\}$. $\sigma''\left(\begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}\right) = \begin{bmatrix} \sigma''(v_1) \\ \vdots \\ \sigma''(v_m) \end{bmatrix}$

Note that $f_{W\Sigma}^L(x)$ is linear to each $(W\Sigma)_{ij}$

and $f_{W\Sigma}^Q(x)$ is quadratic to each $(W\Sigma)_{ij}$.

(b) $\Sigma_{jj} \sim \text{Unif}\{\pm 1\}$. Then, $\mathbb{E}[\Sigma_{jj}] = 0$. $\Rightarrow \mathbb{E}[f_{W\Sigma}^L(x)] = 0$.

$\Sigma_{jj}^2 \equiv 1 \Rightarrow \Sigma = I \Rightarrow f_{W\Sigma}^Q(x) \equiv f_W^Q(x)$.

(c) $\left(\mathbb{E}_\Sigma[|f_{W\Sigma}^L(x)|]\right)^2 \le \mathbb{E}_\Sigma[(f_{W\Sigma}^L(x))^2]$

$= \mathbb{E}_\Sigma\left[\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} s_i s_j \sigma'(w_i^{0T} x) \sigma'(w_j^{0T} x) \Sigma_{ii} \Sigma_{jj} (w_i^T x)(w_j^T x)\right]$

$= \mathbb{E}_\Sigma\left[\frac{1}{m} \sum_{j=1}^{m} s_j^2 \sigma(w_j^{0T} x)^2 \Sigma_{jj}^2 (w_j^T x)^2\right] = \Theta\left(\frac{1}{m} \times m \times m^{-1/2}\right)$

$\sigma(1)$

$\because \mathbb{E}\Sigma_{ii}\Sigma_{jj} = 0$ for $i \neq j$

Thus, $\mathbb{E}_\Sigma[|f_{W\Sigma}^L(x)|] = \Theta(m^{-1/4})$

(d) In comparison,

$$\mathbb{E}\left[f^{Q}_{W_2^*}(x)\right] = f^{Q}_{W^*}(x) = \frac{1}{2\sqrt{m}}\sum_{j=1}^{m} s_j \sigma''(w_j^{*T}x)(w_j^{*T}x)^2$$

$$= O\left(\frac{1}{\sqrt{m}} \times m \times \left(m^{-(1/4)}\right)^2\right) = O(1).$$

Thus, f^Q dominates over f^L, and hence NTK regime may not be useful this case.

#2.4

(Step 1) Expectation Recursion

$$\mathbb{E}[\theta_t] = \mathbb{E}[\theta_{t-1}] - \eta_t \mathbb{E}\left[(\langle x_{\tau(t)}, \theta_{t-1}\rangle - y_{\tau(t)}) x_{\tau(t)}\right]$$

$$\mathbb{E}\left[(\langle x_{\tau(t)}, \theta_{t-1}\rangle - y_{\tau(t)}) x_{\tau(t)} \mid \theta_{t-1}\right] = \frac{1}{n}\sum_{i=1}^{n}(\langle x_i, \theta_{t-1}\rangle - y_i) x_i$$

$$= \frac{1}{n} X^T (X\theta_{t-1} - y)$$

Thus, $\mathbb{E}\left[(\langle x_{\tau(t)}, \theta_{t-1}\rangle - y_{\tau(t)}) x_{\tau(t)}\right] = \mathbb{E}_{\theta_{t-1}}\left[\frac{1}{n} X^T (X\theta_{t-1} - y)\right]$

$$= \frac{1}{n} X^T (X\mathbb{E}[\theta_{t-1}] - y).$$

Hence, $\mathbb{E}[\theta_t] = \mathbb{E}[\theta_{t-1}] - \frac{1}{\|X\|^2} X^T (X\mathbb{E}[\theta_{t-1}] - y)$

$$= \left(I - \frac{1}{\|X\|^2} X^T X\right)\mathbb{E}[\theta_{t-1}] + \frac{1}{\|X\|^2} X^T y \quad \cdots ①$$

(Step 2) Convergence

By letting $\rho_t = \mathbb{E}[\theta_t]$, $A = \left(I - \frac{1}{\|X\|^2} X^T X\right)$, $b = \frac{1}{\|X\|^2} X^T y$,

we have $\rho_0 = 0$, $\rho_{t+1} = A\rho_t + b$, for $0 \le A < 1$. $\cdots ②$

$$\Rightarrow \rho_t = A^t \rho_0 + \sum_{i=0}^{t-1} A^i b = \sum_{i=0}^{t-1} A^i b = \left(\sum_{i=0}^{t-1} A^i\right) \frac{1}{\|X\|^2} X^T y \quad \cdots ③$$

→ convergence in elementwise sense, by ②.

$$(I-A)\left(\sum_{i=0}^{t-1} A^i\right) = I - A^t \to I \text{ as } t \to \infty. \text{ Hence, } \left(\sum_{i=0}^{t-1} A_i\right)\frac{X^T X}{\|X\|^2} \to I \text{ as } t \to \infty$$

Denote $B_t = \sum_{i=0}^{t-1} A_i$. Then, $B_t X^T X \to \|X\|^2 I$ as $t \to \infty$.

$$\Rightarrow B_t X^T X X^T \to \|X\|^2 X^T \text{ as } t \to \infty$$

$$\Rightarrow B_t X^T \to \|X\|^2 X^T (XX^T)^{-1} \text{ as } t \to \infty$$

$$\Rightarrow B_t b = B_t \frac{X^T}{\|X\|^2} y = \frac{1}{\|X\|^2}(B_t X^T) y$$

$$\to X^T (XX^T)^{-1} \text{ as } t \to \infty \quad \cdots ④$$

Thus, ④ with ③ gives

$$\mathbb{E}[\theta_t] = \rho_t = B_t b \to X^T (XX^T)^{-1} y \text{ as } t \to \infty.$$