1 Representer theorem for SGD Solutions

Proof. We have

$$\nabla_\beta \hat{R}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i'(f_\beta(x_i), y_i)\, \Phi(x_i)^T \quad (\text{As } \nabla_\beta f(x,\beta) = \nabla_\beta [\beta^T \Phi(x)] = \Phi(x))$$

So SGD iterates as

$$\beta_0 = 0$$

$$\beta_{t+1} = \beta_t - \eta_t \frac{1}{B}\sum_{i=1}^{B} \ell_i'(f_{\beta_t}(x_{c_i}), y_{c_i})\, \Phi(x_{c_i}),$$

where $c_1 \sim c_B \overset{iid}{\sim} \text{Uniform}[n]$

Hence for $t \geq 0$ $\exists \alpha_t \in \mathbb{R}^n$, $\beta_t = \sum_{i=1}^{n} \alpha_{ti} \Phi(x_i)$

$$f(x, \beta_t) = \Phi^T(x)\beta_t = \Phi^T(x) \sum_{i=1}^{n} \alpha_{ti}\Phi(x_i)$$

$$= \sum_{i=1}^{n} \alpha_{ti}\, \Phi^T(x)\Phi(x_i)$$

$$= \sum_{i=1}^{n} \alpha_{ti}\, k(x, x_i)$$

#

2 SGD for training over-parameterized models

(a) $g_t = \left(f(x_{i_t};\theta_t) - y_{i_t}\right)\nabla f(x_{i_t};\theta_t)$

$\xi_t = g_t - \nabla L(\theta_t)$ we have $\sigma_t^2 := \mathbb{E}\|\xi_t\|^2 \le 2C_1^2 L(\theta_t)$

Proof.

$$\mathbb{E}\|\xi_t\|^2 = \mathbb{E}\left[(g_t^T - \nabla L(\theta_t)^T)(g_t - \nabla L(\theta_t))\right]$$

$$= \mathbb{E}\, g_t^T g_t - \|\nabla L(\theta_t)\|^2 \quad (\text{As } \mathbb{E}[g_t] = \nabla L(\theta_t))$$

$$= \mathbb{E}\,\nabla f(x_{i_t};\theta_t)^T \nabla f(x_{i_t};\theta_t)\left[f(x_{i_t};\theta_t) - y_{i_t}\right]^2 - \|\nabla L(\theta_t)\|^2$$

$$= \frac{1}{n}\sum_{i=1}^n \|\nabla f(x_i,\theta_t)\|^2 (f(x_i,\theta_t) - y_i)^2 - \left\|\frac{1}{n}\sum_{i=1}^n (f(x_i,\theta_t) - y_i)\nabla f(x_i;\theta_t)\right\|^2$$

$$\le \frac{1}{n}\sum_{i=1}^n C_1^2 (f(x_i,\theta_t) - y_i)^2$$

$$= 2C_1^2 L(\theta_t) \quad \#$$

(b) $\exists\, \theta^*$, $L(\theta^*) = 0$ Convex analysis. $L(\cdot)$ is convex.

(b.1) $\eta \le \frac{1}{C_2}$

Prove $\mathbb{E}[L(\theta_{t+1})] \le -\frac{1}{2\eta}\left(\mathbb{E}\|\theta_{t+1} - \theta^*\|^2 - \mathbb{E}\|\theta_t - \theta^*\|^2\right) + \frac{\eta + C_2\eta^2}{2}\sigma_t^2$

Proof. Following lecture note, (Theorem 2.5)

$$\mathbb{E}(f(\theta_{t+1})) \le \mathbb{E} f(\theta_t) - \frac{\eta}{2}\mathbb{E}\|\nabla f(\theta_t)\|^2 + \frac{\eta^2 C_2 \sigma_t}{2}$$

$$\mathbb{E}\left[f(\theta_{t+1}) - f(\theta^*)\right] \le \mathbb{E}\left[\langle \nabla f(\theta_t), \theta_t - \theta^*\rangle\right] - \frac{\eta}{2}\mathbb{E}\|\nabla f(\theta_t)\|^2 + \frac{\eta_t^2 C_2 \sigma^2}{2}$$

$$= -\frac{1}{2\eta}\left(\mathbb{E}\|\theta_t - \eta\nabla f(\theta_t) - \theta^*\|^2 - \|\theta_t - \theta^*\|^2\right) + \frac{\eta^2 C_2 \sigma_t^2}{2}$$

$$\le -\frac{1}{2\eta}\left(\mathbb{E}\left[\|\theta_{t+1} - \theta^*\|^2 - \|\theta_t - \theta^*\|^2\right]\right) + \frac{\eta\sigma_t^2}{2} + \frac{C_2\eta^2\sigma_t^2}{2}$$

#

(b.2) Prove $S_{t+1} \le \frac{\|\theta_0 - \theta^*\|^2}{2\eta} + L(\theta_0) + (\eta + C_2\eta^2) C_1^2 S_t$

We have $\mathbb{E} L(\theta_{t+1}) \le -\frac{1}{2\eta}(\mathbb{E}\|\theta_{t+1} - \theta^*\|^2 - \mathbb{E}\|\theta_t - \theta^*\|^2) + (\eta + C_2\eta^2) C_1^2 \mathbb{E} L(\theta_t)$

from (a), (b.1)

$\vdots$

$$\mathbb{E} L(\theta_1) \le -\frac{1}{2\eta}(\mathbb{E}\|\theta_1 - \theta^*\|^2 - \mathbb{E}\|\theta_0 - \theta^*\|^2) + (\eta + C_2\eta^2) C_1^2 \mathbb{E} L(\theta_0)$$

Summing over these t+1 terms yields

$$S_{t+1} - L(\theta_0) \le \frac{1}{2\eta}(\|\theta_0 - \theta^*\|^2 - \mathbb{E}\|\theta_{t+1} - \theta^*\|^2) + (\eta + C_2\eta^2) C_1^2 S_t$$

$$S_{t+1} \le \frac{1}{2\eta}\|\theta_0 - \theta^*\|^2 + L(\theta_0) + (\eta + C_2\eta^2) C_1^2 S_t$$

#

(b.3) Prove $\mathbb{E} L(\bar{\theta}_T) \le \frac{\|\theta_0 - \theta^*\|^2 + 2\eta L(\theta_0)}{2\eta T(1 - C_1^2\eta - C_1^2 C_2\eta^2)}$

Proof. Let $A = \eta C_1^2 + \eta^2 C_2 C_1^2$, $B = \frac{\|\theta_0 - \theta^*\|^2 + 2\eta L(\theta_0)}{2\eta}$, $0 < A < 1$

from (b.2), $S_{t+1} \le A S_t + B$

$$S_{t+1} - \frac{B}{1-A} \le A\left(S_t - \frac{B}{1-A}\right)$$

$$S_t - \frac{B}{1-A} \le A^t\left(S_0 - \frac{B}{1-A}\right)$$

$$S_0 - \frac{B}{1-A} = L(\theta_0) - \frac{1}{1 - \eta C_1^2 - \eta^2 C_2 C_1^2}\left(L(\theta_0) + \frac{\|\theta_0 - \theta^*\|^2}{2\eta}\right)$$

$$= \frac{-\eta C_1^2 - \eta^2 C_2 C_1^2}{1 - \eta C_1^2 - \eta^2 C_2 C_1^2} L(\theta_0) - \frac{\|\theta_0 - \theta^*\|^2}{2\eta(1 - \eta C_1^2 - \eta^2 C_2 C_1^2)} \le 0$$

$\Rightarrow S_t \le \frac{B}{1-A}$

$$\mathbb{E} L(\bar{\theta}_T) = \frac{1}{T}\sum_{i=0}^{T} \mathbb{E} L(\theta_i) = \frac{1}{T} S_T \le \frac{B}{T(1-A)} = \frac{\|\theta_0 - \theta^*\|^2 + 2\eta L(\theta_0)}{2\eta T(1 - C_1^2\eta - C_1^2 C_2\eta^2)}$$

#

(c) PL analysis $\|\nabla L(\theta)\|^2 \geq \mu L(\theta)$.

Prove $\mathbb{E} L(\theta_T) \leq (1-\frac{\mu\eta}{2}+C_1^2C_2\eta^2)^T L(\theta_0)$

Proof. Like in lecture note (Theorem 2.6)

$\mathbb{E} L(\theta_t) \leq \mathbb{E} L(\theta_{t-1}) - \frac{\eta}{2}\|\nabla L(\theta_{t-1})\|^2 + \frac{C_2\sigma_{t-1}^2\eta^2}{2}$

Noting the PL condition,

We have $\mathbb{E} L(\theta_t) \leq \mathbb{E} L(\theta_{t-1})(1-\frac{\mu\eta}{2}) + \frac{C_2\eta^2\sigma_{t-1}^2}{2}$

$\leq \mathbb{E} L(\theta_{t-1})(1-\frac{\mu\eta}{2}) + C_2C_1^2\eta^2\,\mathbb{E} L(\theta_{t-1})$ (since (a) holds)

$= \mathbb{E} L(\theta_{t-1})(1-\frac{\mu\eta}{2}+C_2C_1^2\eta^2)$

$\Rightarrow \mathbb{E} L(\theta_T) \leq (1-\frac{\mu\eta}{2}+C_1^2C_2\eta^2)^T L(\theta_0)$ #

3. Convergence of SGD under Robins- Monro condition

$f \in C^1(\mathbb{R}^d)$ $\inf_x f(x) = f(x^*) = 0$ f is L- smooth , $\|\nabla f(x)\|^2 \geq 2\mu f(x)$

$x_{t+1} = x_t - \gamma_t g_t$

- $0 < \gamma_t \leq \frac{1}{L}$ $\gamma_{t+1} \leq \gamma_t$, $\gamma_t \to 0 \ (t \to \infty)$
- g_t and x_t independent $\mathbb{E}[g_t] = \nabla f(x_t)$, $\mathrm{Var}[g_t] \leq \sigma^2$

(a) $$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \prod_{k=0}^{t}(1-\mu\gamma_k)\,\mathbb{E}[f(x_0) - f(x^*)] + \sigma^2 \sum_{k=0}^{t} \gamma_k^2 \prod_{l=k+1}^{t}(1-\mu\gamma_l) \quad (1)$$

Proof. By recursion, it sufficies to prove that

$$\mathbb{E} f(x_{t+1}) \leq (1-\mu\gamma_t)[\mathbb{E} f(x_t) + \sigma^2 \gamma_t^2], \ t \geq 0 \quad (2)$$

We have

$$\begin{aligned}
\mathbb{E} f(x_{t+1}) &\leq \mathbb{E} f(x_t) - \gamma_t \mathbb{E}\|\nabla f(x_t)\|^2 + \frac{\gamma_t^2 L \sigma^2}{2} + \frac{\gamma_t^2 L}{2} \mathbb{E}\|\nabla f(x_t)\|^2 \\
&\leq \mathbb{E} f(x_t) + \frac{\gamma_t \sigma^2}{2} - \mathbb{E}\|\nabla f(x_t)\|^2 \left(\gamma_t - \frac{\gamma_t^2 L}{2}\right) \\
&\leq \mathbb{E} f(x_t)(1-\mu\gamma_t) + \frac{\gamma_t \sigma^2}{2} \\
&\leq (1-\mu\gamma_t)(\mathbb{E} f(x_t) + \sigma^2\gamma_t^2) \quad \#
\end{aligned}$$

(b) Fix k.

Prove $\prod_{l=k}^{t}(1-\mu y_l) \to 0 \iff \sum_{k=0}^{t} y_k \to \infty$

Proof. $\prod_{l=k}^{\infty}(1-\mu y_l)$ converges $\iff \sum_{l=k}^{\infty} \ln(1-\mu y_l)$ converges (from analysis course)

If $\sum_{k=0}^{\infty} y_k = +\infty$ $\quad \sum_{l=k}^{\infty} \ln(1-\mu y_l) \le \sum_{l=k}^{\infty} -\mu y_l = -\infty$ we are done.

If $\sum_{l=k}^{\infty} \ln(1-\mu y_l) = -\infty$: if $y_l \not\to 0$

if $y_l \to 0$ $\exists L, \forall l \ge L, \alpha\mu y_l \le -\ln(1-\mu y_l) \le \beta\mu y_l$ ($\alpha, \beta > 0$ are constants)

(as $\lim_{x\to 0} \frac{-\ln(1-x)}{x} = 1$)

This yields $\beta \sum_{l=k}^{\infty} \mu y_l \ge \sum_{l=k}^{\infty} -\ln(1-\mu y_l) = +\infty$ #.

(c) $\sum_{k=0}^{\infty} y_k = \infty \quad \sum_{k=0}^{\infty} y_k^2 < \infty$

Prove $\lim_{t\to\infty} \sum_{k=0}^{t} y_k^2 \prod_{l=k+1}^{t}(1-\mu y_l) \to 0$

Proof. Denote $h_t(k) = \begin{cases} \prod_{l=k+1}^{t}(1-\mu y_l) & k \le t \\ 0 & k > t \end{cases}$, $|h_t(k)| \le 1$, which is integrable under measure mentioned in hint.

From (b), $\lim_{t\to\infty} h_t(k) = 0$

By dominated convergence, $\lim_{t\to\infty} \sum_{k=0}^{\infty} y_k^2 \prod_{l=k+1}^{t}(1-\mu y_l)$

$= \sum_{k=0}^{\infty} y_k^2 \lim_{t\to\infty} \prod_{l=k+1}^{t}(1-\mu y_l) = 0$

But $\lim_{t\to\infty} \sum_{k=0}^{\infty} y_k^2 \prod_{l=k+1}^{t}(1-\mu y_l) = \lim_{t\to\infty} \sum_{k=0}^{t} y_k^2 \prod_{l=k+1}^{t}(1-\mu y_t)$

We are done. #

(d) This directly follows from (a), (b), (c) by taking $t \to \infty$ in (1).

#

(e) Take $f(x) = x^2$ $(\mathbb{R} \to \mathbb{R})$ This is 2-smooth, 4-PL

$\inf_x f(x) = f(0) = 0$

Consider SGD $x_{t+1} = x_t - \eta_t g_t$

with $x_0 = 100$ $\eta_t = \frac{1}{(t+10)^2}$ $g_t = 2x_t + \xi_t$.

$$x_{t+1} = x_t - \frac{2x_t + \xi_t}{(t+10)^2}$$

$$\vdots$$

$$x_1 = x_0 - \frac{2x_0 + \xi_0}{10^2}$$

We have $\mathbb{E}(x_{t+1}) = \mathbb{E}(x_t) - \frac{2\mathbb{E}(x_t)}{(t+10)^2} = \mathbb{E}(x_t)\left(1 - \frac{2}{(t+10)^2}\right)$

$$\mathbb{E}(x_T) = A \prod_{t=0}^{T} \left(1 - \frac{2}{(t+10)^2}\right) \geq 100 \prod_{t=0}^{\infty} \left(1 - \frac{2}{(t+10)^2}\right) = C_0 > 0$$

SGD doesn't converge

$\{\eta_t\}$ satisfies $\sum \eta_t < \infty$ $\sum \eta_t^2 < \infty$

So $\sum \eta_t = \infty$ is necessary.

#