SGD on Lease Squares

Problem:

Let $\hat{L}_n(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle - y_i)^2$, $w \in \mathbb{R}^d$
where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ and assume $X^T X$
is invertible for $X = [x_1, \cdots, x_n]^T$

The ERM is $\hat{w}_n = (X^T X)^{-1} X^T y$

We run SGD:

$$w_{t+1} = w_t - \eta_t g_t,$$

$$g_t(w_t) = 2(\langle w_t, x_{i_t}\rangle - y_{i_t})\, x_{i_t}$$

where each i_t is chosen uniformly at random from $\{1, \cdots, n\}$ and the step sizes satisfy

$$\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty$$

Assume $\|x_i\| \le R \in \mathbb{R}$ $\forall i$

Show that $E\|w_t - \hat{w}_n\|^2 \to 0$ as $t \to \infty$

Solution:

The gradient:

$$\nabla \hat{L}_n(w) = \frac{2}{n} X^T (Xw - y) = Hw - \frac{2}{n} X^T y$$

Hessian:

$$H = \nabla^2 \hat{L}_n(w) = \frac{2}{n} \underbrace{X^T X}_{\succ 0}$$

$\Rightarrow$ H is positive definite

Let $\mu = \lambda_{min}(H) > 0$, $L = \lambda_{max}(H)$

Then $\hat{L}_n$ is μ-strongly convex and has L-Lipschitz gradient

---

$$\langle \nabla \hat{L}_n(w) - \nabla \hat{L}_n(\hat{w}_n), w - \hat{w}_n \rangle$$

$$= \langle H(w - \hat{w}_n), w - \hat{w}_n \rangle$$

$$= (w - \hat{w}_n)^T H (w - \hat{w}_n)$$

$$\geq \mu \| w - \hat{w}_n \|^2 \quad (1)$$

$$\| \nabla \hat{L}_n(w) - \underbrace{\nabla \hat{L}_n(\hat{w}_n)}_{=0} \| = \| H(w - \hat{w}_n) \|$$

$$\leq L \| w - \hat{w}_n \| \quad (2)$$

For any w

$E[g_t(w_t) \mid w_t = w]$

$= \frac{1}{n}\sum_{i=1}^{n} 2(\langle w, x_i\rangle - y_i)\, x_i = \nabla \hat{L}_n(w)$

$\Rightarrow g_t$ is an unbiased estimator of the true gradient

Let $r_i = \langle \hat{w}_n, x_i\rangle - y_i$ be the residuals at the ERM

$\| \underbrace{g_t(w) - g_t(\hat{w}_n)} \| = \| 2\langle w - \hat{w}_n ; x_{i_t}\rangle x_{i_t} \|$

$= 2 \| x_{i_t}^T (w - \hat{w}_n) x_{i_t} \|$

$\leq 2 \underbrace{\|x_{i_t}\|} \, \|w - \hat{w}_n\| \, \underbrace{\|x_{i_t}\|}$

$\leq 2R^2 \|w - \hat{w}_n\|$

Using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$

$\underline{E\|g_t(w)\|^2} = E\| \underbrace{g_t(w) - g_t(\hat{w}_n)}_{a} + \underbrace{g_t(\hat{w}_n)}_{b} \|$

$\leq 2E\underbrace{\|g_t(w) - g_t(\hat{w}_n)\|^2}_{4R^4\|w - \hat{w}_n\|^2} + 2E\|g_t(\hat{w}_n)\|^2$

$C :=$

$$\leq \underbrace{8R^4 \|w - \hat{w}_n\|^2 + \boxed{\frac{2}{n}\sum_{i=1}^{n} 4 r_i^2 R^2}}_{< \infty}$$

$\leftarrow$ bound on $\|x_i\|^2$

2^2 residual def

$E\langle A, B\rangle$

$E\langle E[A], E[B]\rangle$

Expand one SGD step

$$\|w_{t+1} - \hat{w}_n\|^2 = \|w_t - \hat{w}_n - \eta_t g_t\|^2$$

$$= \|w_t - \hat{w}_n\|^2 \cancel{- 2\eta_t g_t \|w_t}$$

$$- 2\eta_t \langle w_t - \hat{w}_n, g_t\rangle + \eta_t^2 \|g_t\|^2$$

$$\Rightarrow E\|w_{t+1} - \hat{w}_n\|^2 = E\|w_t - \hat{w}_n\|^2$$

$$- 2\eta_t \underbrace{E\langle w_t - \hat{w}_n, g_t\rangle}_{} + \eta_t^2 E\|g_t\|^2$$

$$\|$$

$$E\langle w_t - \hat{w}_n, \underbrace{E[g_t]}_{\nabla \hat{L}_n(w_t) - \nabla \hat{L}_n(w_n)}\rangle$$

By (1), $E\langle w_t - \hat{w}_n, g_t(w_t)\rangle$

$$\geq \mu \|w - \hat{w}_n\|^2$$

$$E\|w_{t+1} - \hat{w}_n\|^2 = (\underbrace{1 - 2\mu\eta_t + \eta_t^2 \cdot 8R^4}) E\|w_t - \hat{w}_n\|^2 + C\eta_t^2$$

We know $\eta_t \to 0$ as $t \to \infty$

$\underline{\exists N}$ s.t. $\forall t \geq N$

$$\underline{2\mu\eta_t - 8R^4\eta_t^2 \geq \mu\eta_t} \quad \left(\eta_N \leq \frac{\mu}{8R^4}\right)$$

$$\Rightarrow E\|w_{t+1} - \hat{w}_n\|^2 \le (1-\mu\eta_t)\underbrace{E\|w_t - \hat{w}_n\|^2}_{a_t} + C\eta_t^2$$

$$\forall t \ge N \quad a_{t+1} \le (1-\mu\eta_t) a_t + C\eta_t^2$$

$$a_t \le (1-\mu\eta_{t-1}) a_{t-1} + C\eta_{t-1}^2$$

$$a_{t+1} \le (1-\mu\eta_t)(1-\mu\eta_{t-1})\underline{a_{t-1}} + \underbrace{(1-\mu\eta_t)}_{\Delta} C\,\underline{\eta_{t-1}^2} + C\eta_t^2$$

$$\le \cdots$$

$$\le \boxed{a_N} \prod_{k=N}^{t}(1-\mu\eta_k) + C\sum_{k=N}^{t}\eta_k^2 \prod_{j=k+1}^{t}(1-\mu\eta_j)$$

where $\prod_{j=t+1}^{t}(1-\mu\eta_j) = 1$

$$M = \max_{i=1,\cdots,N} a_i \quad \cancel{\{a_1, a_2, \cdots\}}, a_N$$

$$a_{t+1} \le M\underbrace{\boxed{\prod_{k=N}^{t}(1-\mu\eta_k)}}_{\to 0 \text{ as } t\to\infty} + \cancel{C\sum_{k=N}^{t}} + \underbrace{C\sum_{k=N}^{t}\eta_k^2 \underbrace{\prod_{j=(k+1)}^{t}(1-\mu_j\eta_j)}_{\to 0}}_{\to 0}$$

$$\log(1-x) \le -x$$

$$\log\prod(1-\mu\eta_k) = \sum\log(1-\mu\eta_k) \le \sum -\mu\eta_k \le -\infty$$

$$\Rightarrow E\|w_{t+1} - \hat{w}_n\|^2 \to 0 \text{ as } t\to\infty$$