GD & Convergence

2020年9月23日 8:24

Gradient Descent:

$$f(w') = \underbrace{f(w) + \langle \nabla f(w), w'-w \rangle}_{\text{Estimation}} + \underbrace{\frac{g(w')}{2}\|w'-w\|^2}_{\text{tail}}$$

- Smoothness assumption:

$\exists L,\ |g(w')| \le L$ for all w'. → which means gradient is L-lipschitz.

$$f(y) = f(x) + \int_0^1 \langle f'(x+\tau(y-x)), y-x \rangle \, d\tau$$

$$| f(y) - f(x) - \langle f'(x), y-x \rangle \le \int_0^1 \| f'(x, \tau(y-x)) - f''(x) \| \cdot \|y-x\| \, d\tau$$

$$= \frac{L}{2}\|y-x\|^2$$

$$\Rightarrow f(w') \le f(w) + \langle \nabla f(w), w'-w \rangle + \frac{L}{2}\|w'-w\|^2$$

If $w' = w - \eta \nabla f(w)$

$$\Rightarrow f(w') - f(w) \le \langle \nabla f(w), -\eta \nabla f(w) \rangle + \frac{L\eta^2}{2}\|\nabla f(w)\|^2$$

$$= -\eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(w)\|^2$$

Set $\eta < \frac{2}{L}$. make sure $f(w)$ non-increasing

$$\overset{\text{convex}}{0 \le} f(y) - f(x) - f'(x)(y-x) \overset{\text{Lipschitz}}{\le} \frac{L}{2}\|y-x\|^2$$

Let $z = y - \frac{1}{L}(f'(y) - f'(x))$

$$\Rightarrow 0 \le f(z) - f(y) - f'(y)(z-y) \le \frac{1}{2L}\|f'(y) - f'(x)\|^2$$

$$0 \le f(z) - f(x) - f'(x)(z-x)$$

$$\Rightarrow f(y) - f(x) + f'(y)(z-y) - f'(x)(z-x) \ge \frac{-1}{2L}\|f'(y) - f'(x)\|^2$$

$$f(y) - f(x) - f'(x)(y-x) \ge \frac{1}{2L}\|f'(y) - f'(x)\|^2$$

Convergence Analysis

Suppose f is L-smooth & convex.

Goal $\Rightarrow$ $\boxed{\eta \le \frac{1}{L} \Rightarrow f(w_t) \le f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta t}}$

1. By L-smooth. $f(w_{i+1}) \le f(w_i) - \frac{\eta}{2}\|\nabla f(w_i)\|_2^2$ (proved above)
2. By convexity. $f(w_i) \le f(w^*) + \langle \nabla f(w_i), w_i - w^* \rangle$

$$f(w_{i+1}) \le f(w^*) + \langle \nabla f(w_i), w_i - w^* \rangle - \frac{\eta}{2}\|\nabla f(w_i)\|_2^2$$

$$= f(w^*) - \frac{1}{\eta}\langle w_{i+1} - w_i, w_i - w^* \rangle - \frac{1}{2\eta}\|w_i - w_{i+1}\|^2$$

$$\le f(w^*) + \frac{1}{2\eta}\|w_i - w^*\|_2^2 - {}^{2}\frac{1}{2\eta}\|w_{i+1} - w^*\|_2^2$$

$$\sum_{i=0}^{t-1}\left(\underbrace{f(w_{i+1})}_{} - f(w^*)\right) \le \frac{1}{2\eta}\left(\|w_0 - w^*\|_2^2 - \|w_t - w^*\|_2^2\right) \le \frac{\|w_0 - w^*\|_2^2}{2\eta}$$

Since $f(w_i)$ non-increasing,

$$f(w_t) - f(w^*) \le \frac{\|w_0 - w^*\|_2^2}{2\eta t}$$

GD has two limitations.

- Computing full gradient is slow for big data
- (Theoretical) get stuck at $\nabla f = 0$ points

↓ ...

— (Theoretical) get stuck at $\nabla f = 0$ points

↓

SGD

GD: $\nabla L(w, X, Y) = \frac{1}{N}\sum_i l(w, x_i, y_i)$.

SGD: $G_t = \frac{1}{|S|}\sum_{i\in S} l(w, x_i, y_i)$.

S: mini-batch

Analysis of SGD:

f: L-smooth convex function, $Var(G_t) \le \sigma^2$

$$\Rightarrow \mathbb{E}[f(\overline{w}_t)] \le f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2\eta t} + \eta\sigma^2$$

$$\left(\overline{w}_t = \frac{\sum_{i=1}^t w_i}{t}\right)$$

$$\mathbb{E}[f(w_{i+1})] \le f(w_i) + \mathbb{E}\langle \nabla f(w_i), w_{i+1} - w_i\rangle + \mathbb{E}\left(\frac{L}{2}\|w_{i+1} - w_i\|_2^2\right) \quad \text{(smoothness)}$$

$$= f(w_i) - \eta\langle \nabla f(w_i), \nabla f(w_i)\rangle + \frac{L\eta^2}{2}\mathbb{E}(\|G_i\|^2).$$

$$= f(w_i) - \eta\|\nabla f(w_i)\|_2^2 + \frac{L\eta^2}{2}(\|\nabla f(w_i)\|_2^2 + Var(G_i)) \qquad Var(S) = \overline{(S - \overline{S})^2}$$

$$\le f(w_i) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(w_i)\|_2^2 + \frac{L\eta^2}{2}\sigma^2$$

By convexity $f(w_i) \le f(w^*) + \langle \nabla f(w_i), w_i - w^*\rangle$

$$\mathbb{E}[f(w_{i+1})] \le f(w^*) + \langle \nabla f(w_i), w_i - w^*\rangle - \frac{\eta}{2}\|\nabla f(w_i)\|_2^2 + \frac{\eta}{2}\sigma^2$$

$$\le f(w^*) + \mathbb{E}(\underbrace{\langle G_i, w_i - w^*\rangle - \frac{\eta}{2}\|G_i\|_2^2}_{= \frac{1}{2\eta}(\|w_i - w^*\|_2^2 - \|w_{i+1} - w^*\|_2^2)}) + \eta\sigma^2$$

(Bregman divergence)

Hence. $\mathbb{E}(f(w_{i+1})) \le f(w^*) + \frac{1}{2\eta}\mathbb{E}[\|w_i - w^*\|_2^2 - \|w_{i+1} - w^*\|_2^2] + \eta\sigma^2$

$$\sum_{i=0}^{t}[\mathbb{E}(f(w_{i+1})) - f(w^*)] \le \frac{1}{2\eta}(\|w_0 - w^*\|_2^2 - \mathbb{E}\|w_t - w^*\|_2^2)^2 + t\eta\sigma^2$$

$$\le \frac{1}{2\eta}\|w_0 - w^*\|_2^2 + t\eta\sigma^2$$

$$f(\overline{w}_t) \le \frac{\sum_{i=1}^t f(w_i)}{t}$$ (we do not have $\mathbb{E}f(w_i)$ non-incresing).

$$\Rightarrow \mathbb{E}f(\overline{w}_t) \le f(w^*) + \frac{\|w_0 - w^*\|_2^2}{2t\eta} + \eta\sigma^2$$

If f is strongly convex and smooth,

GD $\Rightarrow$ $(1-p)^t$ linear convergence

SGD $\Rightarrow$ $\frac{1}{T}$ convergence $\Rightarrow$ noise helps escaping local optimum

SVRG (Stochastic variance reduction gradient)

Compute full gradient for every m steps

SVRG (Stochastic variance reduction gradient)

Compute full gradient for every m steps

$X_0 \Rightarrow$ full gradient g_1

$\searrow$ SG (with same data of X_t) : g_2

X_t : SG g_3

$\Rightarrow$ unbiased $g_1 + g_3 - g_2$

$\mathbb{E}(g_1+g_3-g_2) = \mathbb{E}(g_3)$ an estimation for GD (X_t)

but reduce variance

SVRG algorithm

For $s = 1, 2, \cdots$

$\tilde{w} = \tilde{w}_{s-1}$ $\Leftarrow$ place we need to compute full gradient.

$\tilde{u} = \frac{1}{N} \sum_{i=1}^{N} \nabla l_i(\tilde{w}) = \nabla f(\tilde{w})$ (full Gradient)

For $t = 1, 2, \cdots, m$ # SG.

- Randomly pick mini-batch S
- $w_t = w_{t-1} - \eta \underbrace{\left(\tilde{u} + \sum_{i \in S} \nabla l_i(w_{t-1}) - \sum_{i \in S} \nabla l_i(\tilde{w})\right)}_{= v_t}$

Option 1: $\tilde{w}_s = w_m$

Option 2: $\tilde{w}_s = w_i$ for i randomly chosen from $[m]$.

Analysis. (Assume L-smooth & μ-strongly convex).

$$\mathbb{E} \| \tilde{u} + \sum_{i \in S} \nabla l_i(w_{t-1}) - \sum_{i \in S} \nabla l_i(\tilde{w}) \|$$

$$\leq 2\mathbb{E} \| \sum \nabla l_i(w_{t-1}) - \sum \nabla l_i(w^*) \|^2 + 2\mathbb{E} \| \sum_{i \in S} \nabla l_i(\tilde{w}) - \sum \nabla l_i(w^*) - \nabla f(\tilde{w}) \|^2$$

$$= 2\mathbb{E} \| \sum \nabla l_i(w_{t-1}) - \sum \nabla l_i(w^*) \|^2 + 2\mathbb{E} \| \sum_{i \in S} \nabla l_i(\tilde{w}) - \sum \nabla l_i(w^*) - \mathbb{E}\left[\sum \nabla l_i(\tilde{w}) - \sum \nabla l_i(w^*)\right] \|^2$$

$$\leq 2\mathbb{E} \| \sum \nabla l_i(w_{t-1}) - \sum \nabla l_i(w^*) \|^2 + 2\mathbb{E} \| \sum \nabla l_i(\tilde{w}) - \sum \nabla l_i(w^*) \|^2$$

$$\leq 4L \left(f(w_{t-1}) - f(w^*) + f(\tilde{w}) - f(w^*)\right)$$

$$\mathbb{E} \| w_t - w^* \|^2$$

$$= \mathbb{E} \| w_{t-1} - w^* \|^2 - 2\eta \langle w_{t-1} - w^*, \mathbb{E}(v_t) \rangle + \eta^2 \mathbb{E}(\| v_t \|^2)$$

$$\leq \mathbb{E} \| w_{t-1} - w^* \|^2 - 2\eta \langle w_{t-1} - w^*, \nabla f(w_{t-1}) \rangle + 4L\eta^2 \left(f(w_{t-1}) - f(w^*) + f(\tilde{w}) - f(w^*)\right)$$

$$\leq \mathbb{E} \| w_{t-1} - w^* \|^2 - 2\eta \left(f(w_{t-1}) - f(w^*)\right) + 4L\eta^2 \left(f(w_{t-1}) - f(w^*) + f(\tilde{w}) - f(w^*)\right)$$

$$= \mathbb{E} \| w_{t-1} - w^* \|^2 - 2\eta(1 - 2L\eta)\left(f(w_{t-1}) - f(w^*)\right) + 4L\eta^2 \left(f(\tilde{w}) - f(w^*)\right)$$

$$\Rightarrow \mathbb{E} \| w_m - w^* \|^2 \leq \mathbb{E} \| w_0 - w^* \|^2 - 2\eta(1-2L\eta)\mathbb{E}\left(\sum_{t=1}^{m} f(w_{t-1}) - f(w^*) m\right) + 4Lm\eta^2 \, \mathbb{E}\left(f(\tilde{w}) - f(w^*)\right)$$

$$= \mathbb{E} \| \tilde{w} - w^* \|^2 - 2\eta(1-2L\eta) m \, \mathbb{E}\left(f(\tilde{w}_s) - f(w^*)\right) + 4Lm\eta^2 \, \mathbb{E}\left(f(\tilde{w}) - f(w^*)\right)$$

$$\leq \frac{2}{\mu} \mathbb{E}\left(f(\tilde{w}) - f(w^*)\right) - 2\eta(1-2L\eta) m \, \mathbb{E}\left(f(\tilde{w}_s) - f(w^*)\right) + 4Lm\eta^2 \, \mathbb{E}\left(f(\tilde{w}) - f(w^*)\right)$$

$$\Rightarrow \mathbb{E}[f(\tilde{w}_s)-f(w^*)] \le \left[\frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} \right] \mathbb{E}[f(\tilde{w}_{s-1}) - f(w^*)]$$

$\Rightarrow$ Linear convergence rate