CS292F Lecture 8 stochastic subgradient method

Alg: for $k=1,\dots,K$

$x^{(k+1)} = x^{(k)} - t_k \hat{g}_k$

Assumption ① $E[\hat{g}_k \mid x_k] = g_k$ when g_k is a subgradient of f at x_k

② $E[\|\hat{g}_k - g_k\|^2 \mid x_k] \le \sigma^2$ (σ^2: hidden dimension)

③ f is G-Lipschiz. $\|g_k\|_2 \le G$

$x \to x^+$

$\|x^+ - x^*\|^2 = \|x - t\hat{g} - x^*\|^2$

$= \|x - x^*\|^2 - 2t\hat{g}^T(x - x^*) + t^2\|\hat{g}\|^2$

Take $E[\cdot \mid x]$

$E[\|x^+ - x^*\|^2 \mid x] = \|x - x^*\|^2 - 2t\,E[\hat{g} \mid x]^T(x - x^*) + t^2 E[\|\hat{g}\|^2 \mid x]$

Apply ① $= \|x - x^*\|^2 - 2t\,g^T(x - x^*) + t^2\left(\|g\|^2 + E[\|\hat{g} - g\|^2 \mid x]\right)$

② $\le \|x - x^*\|^2 - 2t\,g^T(x - x^*) + t^2(\|g\|^2 + \sigma^2)$

$\le \|x - x^*\|^2 - 2t\,(f(x) - f^*) + t^2(G^2 + \sigma^2)$

$E[g^T(\hat{g} - g) \mid x] = 0$

Take full $E[\cdot]$

$E[\|x^+ - x^*\|^2] \le E[\|x - x^*\|^2] - 2t\,(E[f(x)] - f^*) + t^2(G^2 + \sigma^2)$

Telescope $k=1,2,\dots,K$

$$E[\|x^{(2)}-x^*\|^2] \le E[\|x^{(1)}-x^*\|^2] - 2t_1[E[f(x^{(1)})]-f^*] + t_1^2(G^2+\sigma^2)$$

$$E[\|x^{(3)}-x^*\|^2] \le E[\|x^{(2)}-x^*\|^2] - 2t_2[E[f(x^{(2)})]-f^*] + t_2^2(G^2+\sigma^2)$$

$$\vdots$$

$$E[\|x^{(K+1)}-x^*\|^2] \le E[\|x^{(K)}-x^*\|^2] - 2t_K[E[f(x^{(K)})]-f^*] + t_K^2(G^2+\sigma^2)$$

Add up

$$2\sum_{i=1}^{K} t_i\left(E[f(x^{(i)})]-f^*\right) \le E[\|x^{(1)}-x^*\|^2] - E[\|x^{(K+1)}-x^*\|^2] + \sum_{i=1}^{K} t_i^2\cdot(G^2+\sigma^2)$$

$\min_i a_i \sum_i b_i \le \sum_i a_i b_i$ or $\min_i b_i \sum_i a_i \le \sum_i a_i b_i$ when $a_i, b_i \ge 0$

Strategy 1.

$$2\left(\sum_{i=1}^{K} t_i\right)\cdot\min_i\left(Ef(x^{(i)})-f^*\right) \le \frac{E[\|x^{(1)}-x^*\|^2] + \sum_{i=1}^{K} t_i^2(G^2+\sigma^2)}{2\sum_{i=1}^{K} t_i}$$

$\sum_{i=1}^{K} t_i^2 < \infty$

$\sum_{i=1}^{K} t_i \to \infty$

Strategy 2:

$$2\min_i t_i \cdot \sum_{i=1}^{k} \left(\mathbb{E}[f(x^{(i-1)})] - f^* \right) \leq \|x^{(1)} - x^*\|^2 + \sum_{i=1}^{k} t_i^2 (G^2 + \sigma^2)$$

only difference from non-stochastic subgradient

Divide both sides by $2K \cdot \min_i t_i$

$$\frac{1}{K} \sum_{i=1}^{k} \left(\mathbb{E}[f(x^{(i-1)})] - f^* \right) \leq \frac{\|x^{(1)} - x^*\|^2 + \sum_{i=1}^{k} t_i^2 (G^2 + \sigma^2)}{2 \cdot K \cdot \min_i t_i}$$

average suboptimality through out the updates

$$t_i = \frac{1}{\sqrt{K}}$$

$$= \frac{\|x^{(1)} - x^*\|^2 + G^2 + \sigma^2}{2\sqrt{K}}$$

Nonconvex case: assumption. f is L-smooth

$$\mathbb{E}[\hat{g}|x] = \nabla f(x), \quad \mathbb{E}[\|\hat{g} - \nabla f(x)\|^2 | x] \le G^2$$

By L-smoothness

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

$$\mathbb{E}[f(x^+)|x] \le f(x) - t\langle \mathbb{E}[\hat{g}|x], \nabla f(x)\rangle + \frac{L}{2}t^2 \mathbb{E}[\|\hat{g}\|^2|x]$$

$$= f(x) - t\|\nabla f(x)\|^2 + \frac{L}{2}t^2[\|\nabla f(x)\|^2 + G^2]$$

$$= f(x) - (t - \frac{Lt^2}{2})\|\nabla f(x)\|^2 + \frac{L}{2}t^2G^2$$

$t \le \frac{1}{L}$

$$\le f(x) - \underbrace{(t - \frac{1}{2}t)}_{\frac{t}{2}}\|\nabla f(x)\|^2 + \frac{L}{2}t^2G^2$$

new for SGD.

Take $\mathbb{E}[\cdot]$

$$\mathbb{E}[f(x^+)] \le \mathbb{E}[f(x)] - \frac{t}{2}\mathbb{E}[\|\nabla f(x)\|^2] + \frac{L}{2}t^2G^2$$

Telescope:

$$\sum_{i=1}^{k} \frac{t_i}{2} \mathbb{E}\left[\|\nabla f(x^{(i)})\|^2\right] \leq \mathbb{E}[f(x^{(1)})] - \mathbb{E}[f(x^{(k+1)})] + \frac{LG^2}{2} \sum_{i=1}^{k} t_i^2$$

$$t_i = \frac{1}{\sqrt{k}}$$

$$\frac{1}{2\sqrt{k}} \sum_{i=1}^{k} \mathbb{E}\left[\|\nabla f(x^{(i)})\|^2\right] \leq \mathbb{E}[f(x^{(1)})] - f^* + \frac{LG^2}{2} \cdot 1$$

Divide both sides by $\frac{\sqrt{k}}{2}$

$$\frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\left[\|\nabla f(x^{(i)})\|^2\right] \leq \frac{2\left(\mathbb{E}[f(x^{(1)})] - f^* + \frac{LG^2}{2}\right)}{\sqrt{k}}$$

□