Stochastic Gradient Descent

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$ (e.g. $x \in \mathbb{R}^n$ weights in NN)

Gradient Descent $x_{k+1} = x_k - \alpha \nabla f(x_k)$. $O(n)$

Stochastic Gradient Descent (SGD): $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$ $O(1)$

$i_k \in \{1, \dots, n\}$ chosen at random. ↑ uniformly

Goal: Convergence theory for SGD.

Define $\eta_k = \nabla f_{i_k}(x_k) - \nabla f(x_k)$.

Then SGD is

$$x_{k+1} = x_k - \alpha_k (\nabla f(x_k) + \eta_k)$$ ↓ noise

Conditional expectation

$$\mathbb{E}_k(\nabla f_{i_k}(x_k)) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) = \nabla f(x_k)$$

Since $\nu_k = \nabla f_{i_k}(x_k) - \nabla f(x_k)$,

$$\mathbb{E}_k(\nu_k) = \mathbb{E}_k\left(\nabla f_{i_k}(x_k) - \nabla f(x_k)\right) = 0$$

Assumption (*) $\mathbb{E}_k\left(\|\nu_k\|^2\right) \le \sigma^2, \quad \sigma \ge 0$

or noise variance

$$\mathbb{E}_k\left(\|\nu_k\|^2\right) = \mathbb{E}_k\left(\|\nabla f_{i_k}(x_k) - \nabla f(x_k)\|^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_k) - \nabla f(x_k)\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_k)\|^2 - \frac{2}{n}\sum_{i=1}^{n} \nabla f(x_k)^T \nabla f_i(x_k) + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x_k)\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_k)\|^2 - 2\nabla f(x_k)^T \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k)\right]}_{\nabla f(x_k)} + \|\nabla f(x_k)\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_k)\|^2 - \|\nabla f(x_k)\|^2$$

(*) is equivalent to $\forall x \in \mathbb{R}^n$

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x)\|^2 \le \|\nabla f(x)\|^2 + \sigma^2$$

Assume f is L-Lipschitz, so that

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

For SGD, $x_{k+1} - x_k = -\alpha_k \nabla f_{i_k}(x_k)$

$$f(x_{k+1}) \le f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L\alpha_k^2}{2}\|\nabla f_{i_k}(x_k)\|^2$$

Take conditional expectation on both sides

$$\mathbb{E}_k(f(x_{k+1})) \le f(x_k) - \alpha_k \nabla f(x_k)^T \underbrace{\mathbb{E}_k[\nabla f_{i_k}(x_k)]}_{\nabla f(x_k)} + \frac{L\alpha_k^2}{2}\underbrace{\mathbb{E}_k\left[\|\nabla f_{i_k}(x_k)\|^2\right]}$$

$$\hookrightarrow = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_k)\|^2 \le \|\nabla f(x_k)\|^2 + \sigma^2$$

by (*)

$$\mathbb{E}_k(f(x_{k+1})) \le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2\sigma^2}{2}$$

Assume $\alpha_k \le \frac{1}{L}$

then $\frac{L\alpha_k^2}{2} \le \frac{L}{2}\cdot\frac{1}{L}\cdot\alpha_k = \frac{\alpha_k}{2}$

Then

$$\mathbb{E}_k(f(x_{k+1})) \le f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2\sigma^2}{2}$$

↑ Noise in SGD

Rearrange:

$$\frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 \le f(x_k) - \underbrace{\mathbb{E}_k(f(x_{k+1}))}_{\text{only over } i_k} + \frac{L\alpha_k^2\sigma^2}{2}$$

Take expectation over everything $i_0, i_1, \dots, i_{k-1}$ and use iterated expectation

$$\mathbb{E}\,\mathbb{E}_k\left(f(x_{k+1})\right) = \mathbb{E}\left(f(x_{k+1})\right)$$

$$\frac{\alpha_k}{2}\,\mathbb{E}\,\|\nabla f(x_k)\|^2 \le \mathbb{E}(f(x_k)) - \mathbb{E}(f(x_{k+1})) + \frac{L\alpha_k^2\sigma^2}{2}$$

Sum both sides from $k=0,\dots,t-1$

$$\frac{1}{2}\sum_{k=0}^{t-1}\alpha_k\,\mathbb{E}\,\|\nabla f(x_k)\|^2 \le \sum_{k=0}^{t-1}\left[\mathbb{E}(f(x_k)) - \mathbb{E}(f(x_{k+1}))\right]$$

Telescoping

$$+\frac{L\sigma^2}{2}\sum_{k=0}^{t-1}\alpha_k^2$$

$$\frac{1}{2}\sum_{k=0}^{t-1}\alpha_k\,\mathbb{E}\,\|\nabla f(x_k)\|^2 \le f(x_0) - \mathbb{E}(f(x_t)) + \frac{L\sigma^2}{2}\sum_{k=0}^{t-1}\alpha_k^2$$

$$\le f(x_0) - f_* + \frac{L\sigma^2}{2}\sum_{k=0}^{t-1}\alpha_k^2$$

When $f_* = \min_{\mathbb{R}^n} f$. Now use that

$$\frac{\sum_{k=0}^{t-1}\alpha_k\,\mathbb{E}\,\|\nabla f(x_k)\|^2}{\sum_{k=0}^{t-1}\alpha_k} \ge \min_{0\le k\le t-1}\mathbb{E}\,\|\nabla f(x_k)\|^2$$

weighted average

Here

$$\sum_{k=0}^{t-1} \alpha_k \, \mathbb{E}\|\nabla f(x_k)\|^2 \geq \left(\min_{0\leq k\leq t-1} \mathbb{E}\|\nabla f(x_k)\|^2\right) \sum_{k=0}^{t-1} \alpha_k$$

Plug this in above to get

$$\left(\min_{0\leq k\leq t-1} \mathbb{E}\|\nabla f(x_k)\|^2\right) \sum_{k=0}^{t-1} \alpha_k \leq 2\left(f(x_0) - f_* + \frac{L\sigma^2}{2} \sum_{k=0}^{t-1} \alpha_k^2\right)$$

$$\min_{0\leq k\leq t-1} \mathbb{E}\|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0)-f_*)}{\sum_{k=0}^{t-1} \alpha_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} \alpha_k^2}{\sum_{k=0}^{t-1} \alpha_k}$$

SGD Noise

Remarks

① Constant step size $\alpha_k = \alpha$

$$\min_{0\leq k\leq t-1} \mathbb{E}\|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0)-f_*)}{\alpha t} + \frac{L\sigma^2}{2} \frac{\alpha^2 t}{\alpha t}$$

Noise does not vanish as $t \to \infty$. SGD with $\frac{L\sigma^2\alpha}{2}$

constant step size does not converge.

(2) $\alpha_k = \frac{\alpha}{k+1}$, $\sum_{k=0}^{t-1} \alpha_k = O(\alpha \log(t))$

Compare $\sum_{k=0}^{t-1} \frac{1}{k+1}$ to $\int_0^{t-1} \frac{1}{x+1}\,dx$

$$\sum_{k=0}^{t-1} \alpha_k^2 = \alpha^2 \sum_{k=0}^{t-1} \frac{1}{(k+1)^2} \leq \frac{\pi^2}{6}\alpha^2$$

Then error term is

$$\frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} \alpha_k^2}{\sum_{k=0}^{t-1} \alpha_k} = \frac{L\sigma^2}{2} O\left(\frac{1}{\log(t)}\right) \xrightarrow[\infty]{t} 0$$

(3) $\alpha_k = \frac{\alpha}{\sqrt{k+1}}$,

$$\sum_{k=0}^{t-1} \alpha_k^2 = \alpha^2 \sum_{k=0}^{t-1} \frac{1}{k+1} = \alpha O(\log(t))$$

$$\sum_{k=0}^{t-1} \alpha_k = \alpha \sum_{k=0}^{t-1} \frac{1}{\sqrt{k+1}} \sim \alpha \int_0^t \frac{1}{\sqrt{x+1}}\,dx \sim O(\alpha\sqrt{t})$$

Error term

$$\frac{L\sigma^2}{2}\frac{\sum_{k=0}^{t-1}\alpha_k^2}{\sum_{k=0}^{t}\alpha_k} = \frac{L\sigma^2}{2}O\left(\frac{\log(t)}{\sqrt{t}}\right)$$

much better

Note: Worse than $O(\frac{1}{t})$ when $\sigma = 0$.

Strong Convexity

Assume f is μ-strongly convex, so the PL-inequality

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_*)$$

Recall:

$$\mathbb{E}_k(f(x_{k+1})) \leq f(x_k) - \frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2\sigma^2}{2}$$

$$\text{PL} \leq f(x_k) - \alpha_k\mu(f(x_k) - f_*) + \frac{L\alpha_k^2\sigma^2}{2}$$

Subtract f_* from both sides

$$\mathbb{E}_k\left(f(x_{k+1}) - f_*\right) \le (1-\alpha_k\mu)\left(f(x_k) - f_*\right) + \frac{L\alpha_k^2\sigma^2}{2}$$

Take expectations on both sides over all $i_0, i_1, i_2, \dots, i_{k-1}$ and set

$$e_k = \mathbb{E}\left(f(x_k) - f_*\right)$$

$$e_{k+1} \le (1-\alpha_k\mu)\, e_k + \frac{L\alpha_k^2\sigma^2}{2}$$

Choose $\alpha_k = \dfrac{1}{\mu(k+2)}$, $\alpha_k\mu = \dfrac{1}{k+2}$

$$e_{k+1} \le \left(1 - \frac{1}{k+2}\right) e_k + \frac{L\sigma^2}{2(k+2)^2\mu^2}$$

Claim: $e_k = O\left(\frac{1}{k}\right)$

Let's consider $f_k = \dfrac{\mu^2}{\frac{1}{2}L\sigma^2}\, e_{k-2}$

$$\begin{cases} f_{k+1} \le \left(1-\frac{1}{k}\right) f_k + \frac{1}{k^2}, & k \ge 2 \\ f_2 = \frac{\mu^2}{\frac{1}{2}L_0^2} e_0 \end{cases}$$

Claim: $f_k \le \frac{c}{k}$.

Proof: $f_{k+1} \le \left(\frac{k-1}{k}\right) f_k + \frac{1}{k^2}$

$$\le \left(\frac{k-1}{k}\right)\left(\left(\frac{k-2}{k-1}\right) f_{k-1} + \frac{1}{(k-1)^2}\right) + \frac{1}{k^2}$$

$$= \left(\frac{k-2}{k}\right) f_{k-1} + \frac{1}{k}\left(\frac{1}{k-1} + \frac{1}{k}\right)$$

$$\le \left(\frac{k-2}{k}\right)\left(\left(\frac{k-3}{k-2}\right) f_{k-2} + \frac{1}{(k-2)^2}\right) + \frac{1}{k}\left(\frac{1}{k-1} + \frac{1}{k}\right)$$

$$= \left(\frac{k-3}{k}\right) f_{k-2} + \frac{1}{k}\left(\frac{1}{k-2} + \frac{1}{k-1} + \frac{1}{k}\right)$$

$$\leq \left(\frac{1}{k}\right) f_2 + \frac{1}{k}\left(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}\right)$$

$$\leq \frac{1}{k} \frac{\mu^2}{\frac{1}{2}L\sigma^2} e_0 + \frac{1}{k} \underbrace{\sum_{j=2}^{k} \frac{1}{j}}_{\log(k)}$$

$$= O\left(\frac{\log(k)}{k}\right).$$

Main Point: Compare to $O\left(\frac{1}{\sqrt{k}}\right)$ rate for general functions.

linear rate.

Recall rate is $O(\gamma^k)$ for gradient descent on μ-strongly convex functions $0 < \gamma < 1$.

SGD does not see strong convexity.