Chapter 8: Optimization for Data Science
Stochastic gradient descent

TANN Chantara

Institute of Technology of Cambodia

October 8, 2022

Table of Contents

Stochastic gradient descent algorithm

$$\min_{x \in \mathbb{R}^n} f(x)$$

Consider a sum of structured objective functions

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

We choose an arbitrary $x_0 \in \mathbb{R}^n$ and for $t > 0$ define

Stochastic gradient descent:

1) sample $i \in [n]$ uniformly at random
2) $x_{t+1} = x_t - \gamma_t \nabla f_i(x_t)$

- The vector $\nabla f_i(x_t)$ is called stochastic gradient
- Computing $\nabla f_i(x_t)$ is n-times cheaper than $\nabla f(x_t)$

Unbiasedness

- Stochastic gradient $\nabla f_i(x_t)$ is potentially far form the true gradient $\nabla f(x_t)$
- On expectation they are the same, i.e.,

$$\mathbb{E}[\nabla f_i(x_t)|x_t = x] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x) = \nabla f(x), \quad x \in \mathbb{R}^n$$

- Using the fact that $\{x_t = x\}$ can occur only for x in some finite set X (one element for every choice of indices throughout all iterations),

$$\begin{aligned}\mathbb{E}[\nabla f_i(x_t)^\top (x_t - x^\star)] &= \sum_{x \in X} \mathbb{E}[\nabla f_i(x_t)^\top (x - x^\star)|x_t = x]\mathbb{P}(x_t = x) \\ &= \sum_{x \in X} \nabla f(x)^\top (x - x^\star)\mathbb{P}(x_t = x) \\ &= \mathbb{E}[\nabla f(x_t)^\top (x_t - x^\star)]\end{aligned}$$

- Hence,

$$\mathbb{E}[\nabla f_i(x_t)^\top (x_t - x^\star)] = \mathbb{E}[\nabla f(x_t)^\top (x_t - x^\star)] \geq \mathbb{E}[f(x_t) - f(x^\star)] \quad (1)$$

Bounded stochastic gradients

Theorem (Stochastic gradient descent for Lipschitz functions): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable with global minimum $x^\star$. Suppose that $\|x_0 - x^\star\| \leq R$ and $\mathbb{E}[\|\nabla f_i(x_t)\|^2] \leq B^2$ $\forall t > 0$. Choosing the step size $\gamma = \frac{R}{B\sqrt{T}}$, the stochastic gradient descent for $x_0 \in \mathbb{R}^n$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[f(x_t)] - f(x^\star) \leq \frac{RB}{\sqrt{T}}$$

Proof: Taking the expectation of both sides of the analysis ($\bigstar$) from the standard gradient descent and using the linearity of the expectations gives

$$\sum_{t=0}^{T-1} \mathbb{E}[\nabla f_i(x_t)^\top (x_t - x^\star)] \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_i(x_t)\|^2] + \frac{1}{2\gamma}\|x_0 - x^\star\|^2 \quad (2)$$

By (1), $\mathbb{E}[f(x_t) - f(x^\star)] \leq \mathbb{E}[\nabla f_i(x_t)^\top (x_t - x^\star)]$, which by plugging in the assumptions of the theorem to (2) completes the proof.

Strong convexity

Theorem (Stochastic gradient descent under strong convexity): Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and strongly convex with parameter $\mu > 0$. Let $x^\star$ be the unique global minimum of f. With decreasing step size

$$\gamma_t = \frac{2}{\mu(t+1)}$$

stochastic gradient descent yields

$$\mathbb{E}\left[f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t x_t\right) - f(x^\star)\right] \le \frac{2B^2}{\mu(T+1)},$$

where $B = \max_{t=1}^{T} \mathbb{E}[\|\nabla f_i(x_t)\|]$.

Proof By following the “vanilla" analysis as in the classical gradient descent

$$\mathbb{E}[\nabla f_i(x_t)^\top (x_t - x^\star)] \le \frac{\gamma_t}{2}\mathbb{E}[\|\nabla f_i(x_t)\|^2] + \frac{1}{2\gamma_t}(\mathbb{E}[\|x_t - x^\star\|^2] - \mathbb{E}[\|x_{t+1} - x^\star\|^2])$$

Proof (cont'd)

Since the stochastic gradient is unbiased, see (1) and f is μ-strongly convex

$$\begin{aligned}\mathbb{E}[\nabla f_i(x_t)^\top(x_t - x^\star)] &\overset{\text{unbiased}}{=} \mathbb{E}[\underbrace{\nabla f(x_t)^\top(x_t - x^\star)}_{\geq f(x_t) - f(x^\star) + \frac{\mu}{2}\|x_t - x^\star\|^2}] \\ &\geq \mathbb{E}[f(x_t) - f(x^\star)] + \tfrac{\mu}{2}\mathbb{E}[\|x_t - x^\star\|^2]\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[f(x_t) - f(x^\star)] \leq \frac{\gamma_t}{2}\mathbb{E}[\|\nabla f_i(x_t)\|^2] + \frac{1}{2\gamma_t}(\mathbb{E}[\|x_t - x^\star\|^2] - \mathbb{E}[\|x_{t+1} - x^\star\|^2]) \\ - \frac{\mu}{2}\mathbb{E}[\|x_t - x^\star\|^2] \\ \leq \frac{\gamma_t}{2}B^2 + \frac{(\gamma_t^{-1} - \mu)}{2}\mathbb{E}[\|x_t - x^\star\|^2] - \frac{\gamma_t^{-1}}{2}\mathbb{E}[\|x_{t+1} - x^\star\|^2]\end{aligned}$$

Recall that the step size used is $\gamma_t = \frac{2}{\mu(t+1)}$, which leads to

Proof (cont'd)

$$
\begin{aligned}
& t\mathbb{E}[f(x_t) - f(x^\star)] \\
& \quad \leq \frac{tB^2}{\mu(1+t)} + \frac{t\mu(t-1)}{4}\mathbb{E}[\|x_t - x^\star\|^2] - \frac{t\mu(t+1)}{4}\mathbb{E}[\|x_{t+1} - x^\star\|^2] \\
& \quad \leq \frac{B^2}{\mu} + \frac{\mu}{4}\left(t(t-1)\mathbb{E}[\|x_t - x^\star\|^2] - t(t+1)\mathbb{E}[\|x_{t+1} - x^\star\|^2]\right)
\end{aligned}
$$

By summing over $t = 1, \ldots, T$, we obtain a telescopic sum

$$
\sum_{t=1}^{T} t\mathbb{E}[f(x_t) - f(x^\star)] \leq \frac{TB^2}{\mu} + \frac{\mu}{4}(-T(T+1)\mathbb{E}[\|x_{T+1} - x^\star\|^2]) \leq \frac{TB^2}{\mu}
$$

Define the parameter $\lambda_t = \frac{2t}{T(T+1)}$ and note that $\sum_{t=1}^{T} \lambda_t = 1$

Proof (cont'd)

Since f is convex Jensen's inequality ensures that

$$\begin{aligned} f\left(\sum_{t=1}^{T} \lambda_t x_t\right) - f(x^\star) &\leq \sum_{t=1}^{T} \lambda_t f(x_t) - f(x^\star) \\ &= \frac{2}{T(T+1)} \sum_{t=1}^{T} t f(x_t) - f(x^\star) \end{aligned}$$

Hence, taking the expectation ensures

$$\begin{aligned} \mathbb{E}\left[f\left(\sum_{t=1}^{T} \lambda_t x_t\right) - f(x^\star)\right] &\leq \frac{2}{T(T+1)} \sum_{t=1}^{T} t\mathbb{E}[f(x_t)] - f(x^\star) \\ &\leq \frac{2B^2}{(T+1)\mu}, \end{aligned}$$

which completes the proof.

Mini-batch variants

- stochastic gradient $g_t = \nabla f_i(x_t)$
- average of several stochastic gradients

$$\tilde{g}_t = \frac{1}{m}\sum_{j=1}^{m} g_t^j,$$

where $g_t^j = \nabla f_{i_j}(x_t)$ and the set of (distinct) i_j indices for $j = 1, \ldots, m$ is called a mini-batch of size m

- All g_t^j are defined at the same iterate $x_t \implies$ parallelization over m processors

Mini-batch SGD:

1) sample $i_j \subset \{1, \ldots, n\}^m$ uniformly at random
2) $x_{t+1} = x_t - \gamma_t g_t^j$

- Mini-batch SGD reduces the variance

$$\mathbb{E}[\|\tilde{g}_t - \nabla f(x_t)\|^2] \leq \frac{1}{m}\mathbb{E}[\|g_t^1\|^2] + \frac{1}{m}\|\nabla f(x_t)\|^2 \leq \frac{2B^2}{m}$$

Main take-away points

- **Definitions:** stochastic gradient, stochastic gradient descent, mini-batch variants
- **Properties of SGD:** Unbiasedness, convergence of SGD for Lipschitz functions, convergence of SGD und strong convexity