Stochastic Gradient Method

SIE 449/549: Optimization for Machine Learning

Afrooz Jalilzadeh

Department of Systems and Industrial Engineering
University of Arizona

Motivation

$\min \quad E[f(x,\xi)]$

- Consider the following problem of minimizing an average of functions:

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

- The gradient of the above objective function will be:

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

- If we apply gradient descent algorithm, we would be repeating the following steps:

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k)\right)$$

$n \gg 1$

- However, gradient descent will be very costly if we have n in the order of, say, 1 millions. Instead, we can apply the following algorithm:

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

where $i_k \in \{1, \ldots, n\}$ is some chosen index at iteration k

SG

- Consider the following problem of minimizing an average of functions:

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

- The following algorithm is called the **stochastic gradient** method or **SG**:

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

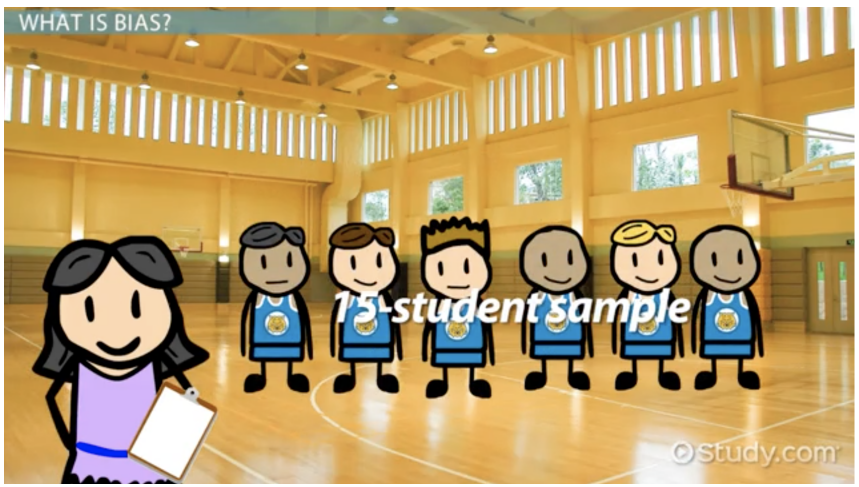
where $i_k \in \{1, \ldots, n\}$ is some chosen index at iteration k

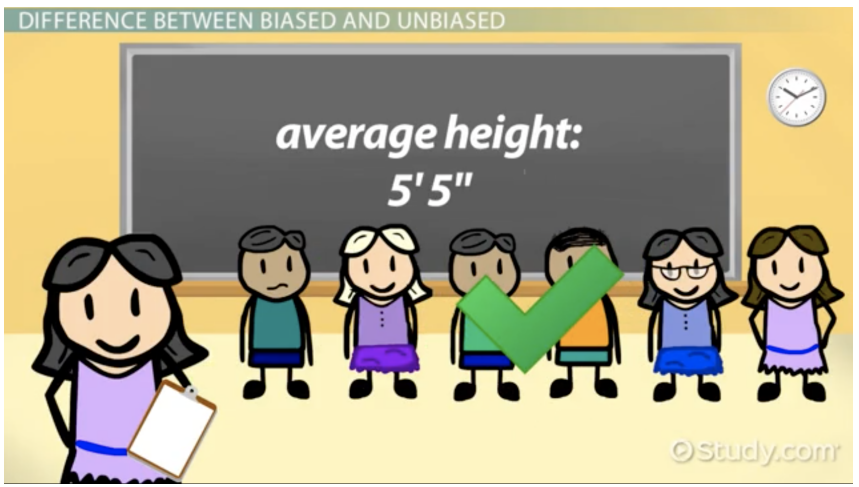
- The main appeal and motivations of using SG are:
  - The iteration cost of SG is independent of n
  - SG can be a big savings in terms of memory usage

- Is SG a descent method? NO!

Unbiased estimator

- If you were going to check the average heights of a high school by looking at a sample of students
- You wouldn't only call for members of basketball team! Not a good estimate!
- If the actual average is $5'5''$ we want a sample with an average of around $5'5''$





- $\hat{\theta}$ is said to be unbiased for a function θ if it equals θ in expectation:

$$\mathbb{E}[\hat{\theta}] = \theta$$

Choosing index i_k

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

- **Randomized Rule:** Every iteration we choose $i_k \in \{1, \ldots, n\}$ uniformly at random
- For randomized rule, we have

$$\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x)$$

$$E\left[\nabla f_{i_k}(x)\right] = \sum_{i=1}^{n} p(i = i_k)\, \nabla f_{i_k}(x) = \frac{1}{n} \sum_{k=1}^{n} \nabla f_{i_k}(x) = \nabla f(x).$$

- So, we can view SG as an **unbiased estimate** of the gradient at each step
- Each iteration updated by:

$$\text{gradient} + \text{zero-mean noise}$$

$$x_{k+1} = x_k - \alpha_k \underbrace{\nabla f_{i_k}(x_k)}_{\nabla f(x_k) + w}$$

error

Example: Least Square

$f(x) = \frac{1}{n}\sum f_i(x)$

- Suppose $A \in \mathbb{R}^{n\times m}$, $x \in \mathbb{R}^{m\times 1}$ and $b \in \mathbb{R}^{n\times 1}$:

$$\min_x \|Ax - b\|^2$$

$f(x) = \|Ax-b\|^2$

$\nabla f(x) = 2A^T(Ax-b)$

- Rewrite the problem in the form of $\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

$$\underbrace{\begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_n^T \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}}_{x} - \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}}_{b} = \begin{bmatrix} a_1^Tx - b_1 \\ \vdots \\ a_n^Tx - b_n \end{bmatrix} \Rightarrow \|Ax-b\|^2 = \frac{1}{n}\sum_{i=1}^{n} \overbrace{n(a_i^Tx-b_i)^2}^{f_i(x):=}$$

$$= \frac{1}{n}\sum f_i(x)$$

- Then, the sample gradient $\nabla f_i(x)$ is

$f_i(x) = n(a_i^Tx-b_i)^2 \qquad \nabla f_i(x) = 2n\,a_i(a_i^Tx-b).$

- In terms of time complexity, we have:

Full Gradient: $O(nm)$

Stochastic Gradient: $O(m)$

Example: Logistic Regression

- Given $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$, consider the following logistic regression problem:

$$\min_x \frac{1}{n} \sum_{i=1}^{n} \underbrace{\log(1 + \exp(-a_i^T x))}_{f_i(x)}$$

- Then, the sample gradient $\nabla f_i(x)$ is

$$\nabla f_i(x) = \frac{-a_i \exp(-a_i^T x)}{1 + \exp(-a_i^T x)}$$

- In terms of time complexity, we have:

Full Gradient: $O(nm)$.

Stochastic Gradient: $O(m)$

Clearly, the stochastic steps are much more affordable!

Choice of Step Size

- In gradient descent, step size can be constant
- Can we use fixed step size for SG?
- SG with fixed step size cannot converge to global/local minimizers

Intuition:

- Consider $\min_x f(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
- If x^* is the minimizer, then

$$\nabla f(x^*) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^*) = 0$$

- In SG, we may have:

$$\nabla f_i(x^*) \neq 0$$

- Even if we got minimizer, SG will move away from it

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

$x_{k+1} = x_k - \alpha \nabla f_{i_k}(x_k)$

Diminishing Step Size

- **Diminishing Stepsize:** Decrease α_k in each iteration:
- α_k satisfies the following two conditions

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

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

Stochastic gradient:

- Pros:
  - Cheaper computation per iteration
  - Faster convergence in the beginning
- Cons:
  - Less stable, slower final convergence
  - Hard to tune step size

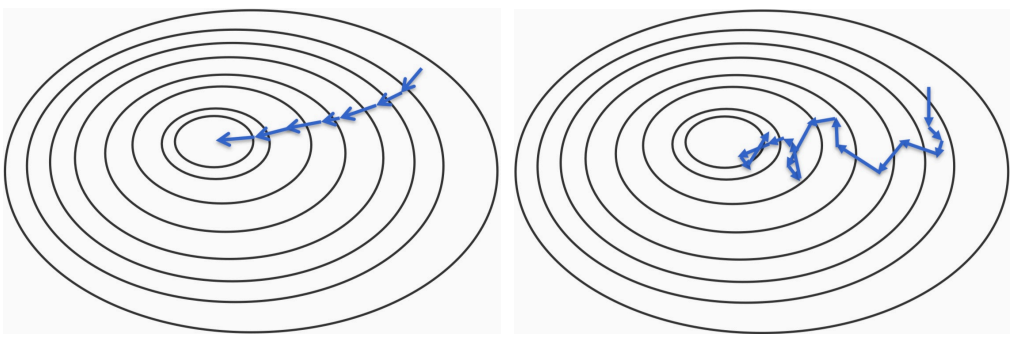


Figure 1: Convergence of GD vs SG

Convergence Analysis

Properties of Expectation

- For any random variables X and Y and constant α:
  - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
  - $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$
  - If $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$
  - If $X = c$ for some real number c, then $\mathbb{E}[X] = c$
  - If X and Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- **Jensen's inequality**: Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function and X a random variable with finite expectation. Then:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(x)]$$

- **Tower rule:** $\mathbb{E}\left[\mathbb{E}[X \mid Y]\right] = \mathbb{E}[X]$

Convergence Analysis

Assumption 1

Define $w_k \triangleq \nabla f(x_k) - \nabla f_{i_k}(x_k)$*:*

(i) *Conditional unbiasedness:* $\mathbb{E}[w_k \mid x_k] = 0$

$$\mathbb{E}\left[\nabla f(x_k) - \nabla f_{i_k}(x_k) \mid x_k\right] = \nabla f(x_k) - \mathbb{E}\left[\nabla f_{i_k}(x_k) \mid x_k\right] = 0 \qquad \mathbb{E}\left[\nabla f_{i_k}(x_k) \mid x_k\right] = \nabla f(x_k)$$

unbiased.

(ii) *Conditional boundedness of the variance:* $\mathbb{E}[\|w_k\|^2 \mid x_k] \leq \sigma^2$.

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}^2[Y] \qquad \mathrm{Var}(w_k \mid x_k) = \mathbb{E}[w_k w_k^T \mid x_k] - \underbrace{\mathbb{E}^2[w_k \mid x_k]}_{=0}$$

Convergence rate: Define $\bar{x}_T = \dfrac{1}{T}\sum_{k=0}^{T-1} x_{k+1}$ and choose $\alpha \leq \dfrac{1}{2L}$, then we have:

$$\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \frac{1}{2\alpha T}\|x_0 - x^*\|^2 + \alpha\sigma^2$$

What is the best choice of α to minimize the upper bound?

$$x_{k+1} = x_k - \alpha_k \nabla f_i(x_k).$$

Convergence Analysis

Convexity: $f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k)$

Theorem 1

Consider $\min_x f(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x)$, *where* f *is convex and has a Lipschitz gradient with constant* L. *Suppose Assumption 1 holds and* $\{x_k\}_{k=1}^T$ *be the sequence generated by SG after* T *iterations. Define* $\bar{x}_T = \frac{1}{T}\sum_{k=0}^{T-1} x_{k+1}$:

(i) Choose $\alpha = \min\{\frac{1}{2L}, \frac{1}{\sqrt{T}}\}$, *then* $\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \mathcal{O}(1/\sqrt{T})$.

(ii) Choose $\alpha_k = \min\{\frac{1}{2L}, \frac{1}{\sqrt{k}}\}$, *then* $\mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \mathcal{O}(\log T/\sqrt{T})$.

proof. Convexity of f: $f(x_k) \leq f(x^*) + \langle \nabla f(x_k), x_k - x^* \rangle$

Lipschitz: $f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$ +

$f(x_{k+1}) \leq f(x^*) + \langle \nabla f(x_k), x_{k+1} - x^* \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$

$x_{k+1} = x_k - \alpha \nabla f_i(x_k) \longrightarrow \nabla f_i(x_k) = \frac{1}{\alpha}(x_k - x_{k+1})$ $\quad w_k := \nabla f(x_k) - \nabla f_i(x_k)$

$$f(x_{k+1}) \leq f(x^*) + \langle \nabla f(x_k) - \nabla f_i(x_k), x_{k+1} - x^* \rangle + \langle \nabla f_i(x_k), x_{k+1} - x^* \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$= f(x^*) + \underbrace{\langle w_k, x_{k+1} - x^* \rangle}_{\langle w_k, x_{k+1} \pm x_k - x^* \rangle} + \frac{1}{\alpha}\langle x_k - x_{k+1}, x_{k+1} - x^* \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$f(x_{k+1}) \leq f(x^*) + \langle w_k, x_k - x^* \rangle + \langle w_k, x_{k+1} - x_k \rangle + \frac{1}{2\alpha}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) + \left(\frac{L}{2} - \frac{1}{2\alpha}\right)\|x_{k+1} - x_k\|^2$$

Convergence Analysis

Young's ineq. $\uparrow$

Taking conditional Expectation from both sides, using $\langle w_k, x_{k+1}-x_k\rangle \le \alpha\|w_k\|^2 + \frac{1}{4\alpha}\|x_{k+1}-x_k\|^2$

$$E\left[f(x_{k+1})-f(x^*)\,|\,x_k\right] \le \underbrace{E\left[\langle w_k, x_k - x^*\rangle\,|\,x_k\right]}_{=0} + \alpha \underbrace{E\left[\|w_k\|^2\,|\,x_k\right]}_{\le \sigma^2} + \underbrace{\frac{1}{4\alpha}}_{a} E\left[\|x_{k+1}-x_k\|^2\,|\,x_k\right]$$

$$+ \frac{1}{2\alpha}E\left[\|x_k-x^*\|^2\,|\,x_k\right] - \frac{1}{2\alpha}E\left[\|x_{k+1}-x^*\|^2\,|\,x_k\right] + \underbrace{\left(\frac{1}{2}-\frac{1}{2\alpha}\right)}_{b} E\left[\|x_{k+1}-x_k\|^2\,|\,x_k\right]$$

we need $a+b\le 0 \Longrightarrow$ choose $\boxed{\alpha \le \frac{1}{2}}$

Taking another Expectation:

$$E\left[f(x_{k+1})-f(x^*)\right] \le \alpha\sigma^2 + \frac{1}{2\alpha}E\left[\|x_k-x^*\|^2\right] - \frac{1}{2\alpha}E\left[\|x_{k+1}-x^*\|^2\right]$$

$$\underbrace{\frac{1}{T}\sum_{k=0}^{T-1} E\left[f(x_{k+1})-f(x^*)\right]}_{= E\left[\frac{1}{T}\sum_{k=0}^{T-1}\left(f(x_{k+1})-f(x^*)\right)\right]} \le \alpha\sigma^2 + \frac{1}{2\alpha T}E\left[\|x_0-x^*\|^2\right]$$

$$E\left[f(\bar{x}_T)-f(x^*)\right] \le$$

Choose $\alpha = \frac{1}{\sqrt{T}}$, then $E\left[f(\bar{x}_T)-f(x^*)\right] \le \frac{\sigma^2}{\sqrt{T}} + \frac{1}{2\sqrt{T}}\|x_0-x^*\|^2 = O\left(1/\sqrt{T}\right)$

Mini-batch SG

$\min \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ $\quad x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

- Idea: Use the average of several $\nabla f_i(x_k)$ values in each iteration to get a better estimate of the full gradient $\nabla f(x)$
- Called *mini-batching*: useful when multiple gradients can be evaluated in parallel
- During the $k-$th update, we choose a random subset $I_k \subset \{1, ..., n\}$, where $|I_k| = b << n$, and use the following update rule:

$$x_{k+1} = x_k - \frac{\alpha_k}{b} \sum_{i \in I_k} \nabla f_i(x_k)$$

- As before, the gradient estimate is unbiased:

$$E\left[\frac{1}{b}\sum_{i\in I} \nabla f_i(x)\right] = \nabla f(x)$$

- Its variance is reduced by a factor $1/b$, though at the cost of b times more expensive running time at each iteration

$$E\left[\left\|\frac{1}{b}\sum_{i\in I}\nabla f_i(x_k) - \nabla f(x)\right\|^2 | x_k\right] = \frac{1}{b^2}\sum_{i\in I} E\left[\left\|\nabla f_i(x_k) - \nabla f(x_k)\right\|^2 | x_k\right] \leq \frac{\sigma^2}{b}$$

- One may regard Mini-batch SG as a compromise between SG and full GD

Mini-batch SG

Convergence rate: Choosing $\alpha_k = \mathcal{O}(1/\sqrt{k})$, one can show that:

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \mathcal{O}(\sqrt{b/k} + b/k) = \mathcal{O}(1/\sqrt{k})$$

- While mini-batches help reducing the variance, they do not necessarily lead to faster convergence in the theoretical domain
- Mini-batch SG turns out performing not too bad in practice

Algorithm 1 Mini-batch Stochastic Gradient Method

Initialization: pick $x_0 \in \mathbb{R}^m$ arbitrarily, choose $b << n$

for $k = 0, 1, 2, \ldots, T-1$ **do**

choose subset $I_k \subset \{1, ..., n\}$, where $|I_k| = b << n$, uniformly at random

choose $\alpha_k = \min\{\frac{1}{2L}, \frac{1}{\sqrt{T}}\}$ or $\alpha_k = 1/\sqrt{k}$

set $x_{k+1} = x_k - \frac{\alpha_k}{b} \sum_{i \in I_k} \nabla f_i(x_k)$

end for
