CSED490Y: Optimization for Machine Learning

Week 08: Stochastic gradient descent

Namhoon Lee

POSTECH

Spring 2022

Admin

(NEW) Midway group presentation

- Present a research paper that is most relevant to your project.
- Explain how this paper is related to your idea.

Logistics

- Time: 15 minutes
- Date: Choose either $\{11, 18, 25\}$ of May – sign up here by this week
- Scores: 5% – it replaces one of two remaining quizzes

Stochastic gradient descent

$$\min_x f(x)$$

So far we have been assuming that we have access to the gradient $\nabla f(x)$. For example,

$$\text{(GD)} \qquad x_{t+1} = x_t - \eta \nabla f(x_t)$$

for which we call "oracle" for the gradient at any point x to perform GD.

Stochastic gradient descent

So far we have been assuming that we have access to the gradient $\nabla f(x)$. For example,

$$\text{(GD)} \qquad x_{t+1} = x_t - \eta \nabla f(x_t)$$

for which we call "oracle" for the gradient at any point x to perform GD.

In practice, we may not have access to the full gradient (*i.e.*, stochastic oracle).

- Gradient is noisy or inexact.
- Gradient is too expensive to compute.

Stochastic gradient descent

In stochastic setting, we assume that the gradient that oracle returns is not exact but only the expected value of it is.

Stochastic gradient descent

In stochastic setting, we assume that the gradient that oracle returns is not exact but only the expected value of it is.

A stochastic oracle for a differentiable function f takes as input a vector $x \in \mathbb{R}^d$ and outputs a random vector $g \in \mathbb{R}^d$ such that

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

where the expectation is taken with respect to the randomization of the oracle.

We say that the oracle is an unbiased estimator of the true gradient.

Examples

$x \in \mathbb{R}^d$

$i = \{1, \cdots, d\}$

$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ — i-th

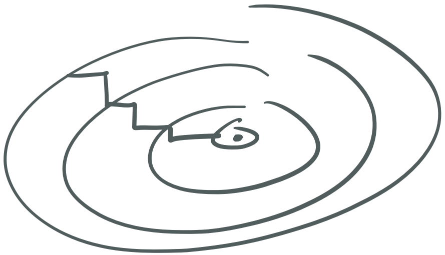
Random coordinate optimization

- Randomly sample a coordinate and update the corresponding variable at a time.

$$x_{t+1} = x_t - \eta \nabla_{i_t} f(x_t) e_{i_t}$$

where $\nabla_{i_t} f(x_t) = \frac{\partial f}{\partial x^{i_t}}(x_t)$, and e_{i_t} represents the i_t-th standard unit vector, *i.e.*, $e_{i_t}^j = 0$ if $j \neq i$ and $e_{i_t}^j = 1$ otherwise.

- can be faster than gradient descent if iterations are d times cheaper.



$O(1)$	$O(d)$
CD	GD

Examples

Finite sum optimization

- $f(x)$ is given as the sum of many terms.

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

Examples

Finite sum optimization

- $f(x)$ is given as the sum of many terms.

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

- Many machine learning problems fall into this category, *e.g.*, least squares:

$$f(x) = \frac{1}{n}\|Ax - b\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\underline{a_i}^\top x - b_i)^2$$

$A \in \mathbb{R}^{n \times d}$

$b \in \mathbb{R}^n \quad x \in \mathbb{R}^d$

Empirical risk minimization

In machine learning, we wish to minimize the expected risk

$$\min_{x} \mathbb{E}_{\xi}\left[f(x;\xi)\right]$$

but typically the distribution over ξ is unknown.

Empirical risk minimization

In machine learning, we wish to minimize the expected risk

$$\min_x \mathbb{E}_\xi \left[f(x;\xi) \right]$$

but typically the distribution over ξ is unknown.

So instead we minimize the empirical risk

$$\min_x f(x) = \frac{1}{n} \sum_i^n f_i(x)$$

hoping that n data (*i.e.* training data) may represent the distribution.

Deterministic vs Stochastic methods

Given a finite sum $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$,

Deterministic gradient method:

$$x_{t+1} = x_t - \eta \nabla f(x_t) = x_t - \eta \nabla \left(\frac{1}{n} \sum_{i=1}^{n} f_i(x_t) \right) = x_t - \frac{\eta}{n} \sum_{i=1}^{n} \nabla f_i(x_t)$$

Deterministic vs Stochastic methods

Given a finite sum $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$,

Deterministic gradient method:

$$x_{t+1} = x_t - \eta \nabla f(x_t) = x_t - \eta \nabla \left(\frac{1}{n} \sum_{i=1}^{n} f_i(x_t) \right) = x_t - \frac{\eta}{n} \sum_{i=1}^{n} \nabla f_i(x_t)$$

$O(n)$

- The cost of each update step is proportional to n; if n is large (a lot of data), performing GD can be very expensive.
- We know that this method converges with a fixed step size η.

Deterministic vs Stochastic methods

Given a finite sum $f(x) = \underline{\underline{\frac{1}{n}}} \sum_{i=1}^{n} f_i(x)$,

$\nabla f(x_t)$

↕

Stochastic gradient method:

$P(i_t = i) = \frac{1}{n}$

$$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t)$$

where $\underline{i_t} = \{1, 2, ..., n\}$ is selected uniformly at random.

Deterministic vs Stochastic methods

Given a finite sum $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$,

Stochastic gradient method:

$$x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t)$$

where $i_t = \{1, 2, ..., n\}$ is selected uniformly at random.

- The cost of each update is independent of n. $O(1)$
- The stochastic gradient is indeed an unbiased estimate of the full gradient; *i.e.*, with $p(i_t = i) = 1/n$

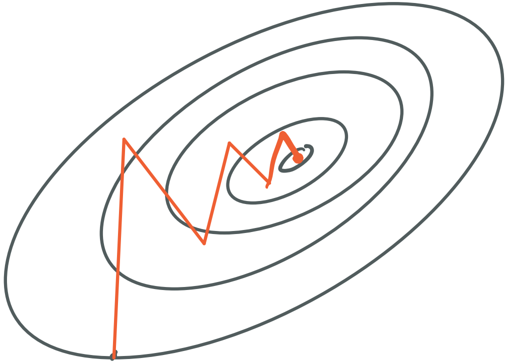
$$\mathbb{E}\left[\underbrace{\nabla f_{i_t}(x)}_{\text{random v.}}\right] = \sum_{i=1}^{n} p(i_t = i)\nabla f_i(x) = \sum_{i=1}^{n} \frac{1}{n}\nabla f_i(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x) = \nabla f(x)$$

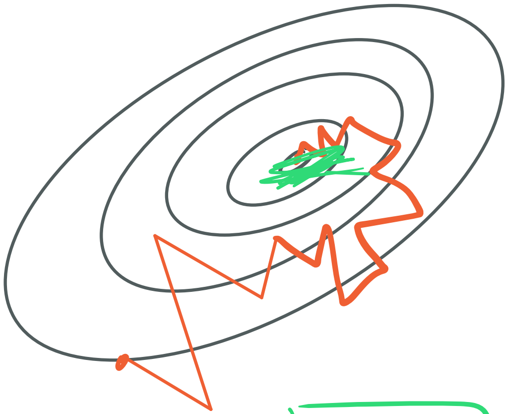
- This method requires a decreasing step size $\eta \to 0$ to converge.

Deterministic vs Stochastic methods

Illustrating determinstic vs stochastic methods (level sets)

$g_t \neq \nabla f(x_t)$



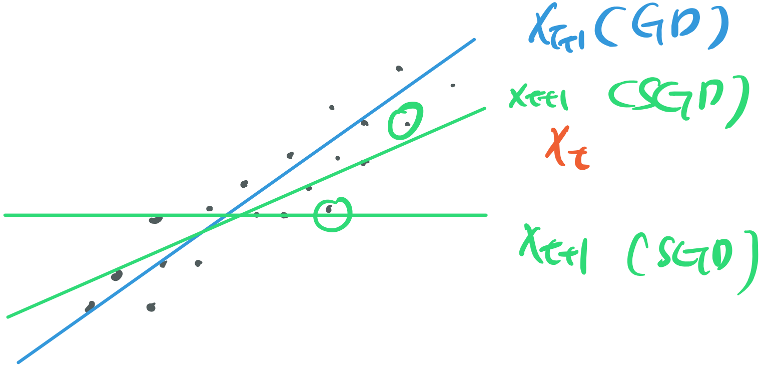


$$\boxed{\eta \to 0}$$

Deterministic vs Stochastic methods

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

Illustrating determinstic vs stochastic methods (linear regression)



Deterministic vs Stochastic methods

Comparing determinstic vs stochastic methods in convergence rate

Deterministic vs Stochastic methods

Comparing determinstic vs stochastic methods in convergence rate

For non-smooth case, the convergence rates are the same.

- $\mathcal{O}(1/\sqrt{t})$ for convex ⟸ $\underline{\underline{t}} \sim O(1/\varepsilon^2)$ × cost of 1 step iteration
- $\mathcal{O}(1/t)$ for strongly convex (not proved in the class)
- Same rate as deterministic method, but n times faster.

S / \ D

$O(1)$ $O(n)$

Deterministic vs Stochastic methods

Comparing determinstic vs stochastic methods in convergence rate

For non-smooth case, the convergence rates are the same.

- $\mathcal{O}(1/\sqrt{t})$ for convex
- $\mathcal{O}(1/t)$ for strongly convex (not proved in the class)
- Same rate as deterministic method, but n times faster.

∇f Lipschitz continuous $T \sim 1/\varepsilon^2$

For smooth case, stochastic method is slower.

$T \sim \frac{1}{\varepsilon}$

- $\mathcal{O}(1/\sqrt{t})$ for convex (whereas for deterministic $\mathcal{O}(1/t)$)
- $\mathcal{O}(1/t)$ for strongly convex (whereas for deterministic $\mathcal{O}(\rho^t)$)
- Even momentum methods do not improve this rate in stochastic setting.

Convergence rate

subgrad. $\|\underline{g_x}\|_2^2 \le G^2$ stochastic subgradient.

$\mathbb{E}[\|g\|_2^2] \le G^2$

Convergence rate proof for non-smooth case (1/2)

(SGD) $\underline{x_{t+1}} = \underline{x_t} - \eta \underline{g_t}$

$$\|x_{t+1} - x^*\|_2^2 \overset{(SGD)}{=} \|x_t - \eta g_t - x^*\|_2^2$$

$$= \|x_t - x^*\|_2^2 - 2\eta\langle g_t, x_t - x^*\rangle + \eta^2\|g_t\|_2^2$$

one step

$$\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2 \mid x_t\right] = \|x_t - x^*\|_2^2 - 2\eta\langle \underbrace{\mathbb{E}[g_t \mid x_t]}_{\nabla f(x_t)}, x_t - x^*\rangle + \eta^2\mathbb{E}\left[\|g_t\|_2^2 \mid x_t\right]$$

convexity of f

$$\le \|x_t - x^*\|_2^2 - 2\eta\left(f(x_t) - f(x^*)\right) + \eta^2\mathbb{E}\left[\|g_t\|_2^2 \mid x_t\right]$$

Convergence rate

$$\mathbb{E} = \mathbb{E}_{\xi_t}\mathbb{E}_{\xi_{t-1}}\cdots\mathbb{E}_{\xi_1}\left[\qquad\right]$$

Convergence rate proof for non-smooth case (2/2)

Total expectation

$$\mathbb{E}\left[\|x_{t+1}-x^*\|\right] \le \mathbb{E}\left[\|x_t-x^*\|\right] - 2\eta\left(\mathbb{E}\left[\boxed{f(x_t)}\right]-f(x^*)\right) + \eta^2\overset{\le G^2}{\mathbb{E}\left[\|g_t\|^2\right]}$$

$$\mathbb{E}[f(x_t)]-f(x^*) \le \frac{1}{2\eta}\left(\underline{\mathbb{E}[\|x_t-x^*\|]} - \underline{\mathbb{E}[\|x_{t+1}-x^*\|]}\right) + \frac{\eta G^2}{2}$$

sum both sides for T iterations, & divide by T.

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum x_t\right)\right]-f(x^*) \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} \Rightarrow O(1/\sqrt{T})$$

Thank you

Any questions?

Credits

A lot of material in this course is borrowed or derived from the following:

- Numerical Optimization, Jorge Nocedal and Stephen J. Wright.
- Convex Optimization, Stephen Boyd and Lieven Vandenberghe.
- Convex Optimization, Ryan Tibshirani.
- Optimization for Machine Learning, Martin Jaggi and Nicolas Flammarion.
- Optimization Algorithms, Constantine Caramanis.
- Advanced Machine Learning, Mark Schmidt.