Convex Optimization

Part 6: Stochastic gradient methods

Namhoon Lee

POSTECH

28 Nov 2022

Deterministic oracle model

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

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

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

In practice, we may not have access to the true gradient.

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

Stochastic oracle model

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.

Coordinate optimization

Coordinate descent: only updates one variable at a time

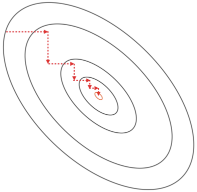


Figure: coordinate optimization; figure from Schmidt lecture note

- no better than GD either convergence or computation wise

Randomized coordinate descent: at iteration t randomly sample a coordinate i_t and perform

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

- can be faster than gradient descent if iterations are d times cheaper.
- can be applied to separable functions in general (*e.g.*, $f(x) = \|x\|_2^2$)

Analyzing coordinate optimization

We assume that each $\nabla_j f$ is L-Lipschitz ("coordinate wise Lipschitz")

$$|\nabla_j f(x + \gamma e_j) - \nabla_j f(x)| \leq L|\gamma|$$

which for twice differentiable functions is equivalent to $|\nabla^2_{jj} f(x)| \leq L$ for all j

- if gradient is L-Lipschitz then it's also coordinate wise L-Lipschitz

coordinate-wise Lipschitz assumption implies a coordinate-wise descent lemma

$$f(x_{t+1}) \leq f(x_t) + \nabla_j f(x_t)(x_{t+1} - x_t)_j + \frac{L}{2}(x_{t+1} - x_t)^2_j$$

GD with step size $\eta = 1/L$ gives a progress bound for updating coordinate j_t

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}|\nabla_{j_t} f(x_t)|^2$$

expected progress with random selection of j_t

$$
\begin{aligned}
\mathbb{E}[f(x_{t+1})] &\leq \mathbb{E}\left[f(x_t) - \frac{1}{2L}|\nabla_{j_t} f(x_t)|^2\right] \\
&\leq \mathbb{E}\big[f(x_t)\big] - \frac{1}{2L}\mathbb{E}\big[|\nabla_{j_t} f(x_t)|^2\big] \\
&\leq f(x_t) - \frac{1}{2L}\sum_{j=1}^{d} p(j_t = j)|\nabla_j f(x_t)|^2
\end{aligned}
$$

choose j_t uniformly at random, *i.e.*, $p(j_t = j) = 1/d$

$$
\mathbb{E}[f(x_{t+1})] \leq f(x_t) - \frac{1}{2dL}\|\nabla f(x_t)\|^2
$$

Under μ-strong convexity we get

$$\mathbb{E}\big[f(x_t)\big] - f^\star \leq \left(1 - \frac{\mu}{dL}\right)^t \big(f(x_0) - f^\star\big)$$

which means we have the iteration complexity $\mathcal{O}(d\frac{L}{\mu}\log(1/\epsilon))$

So compared to GD under strong convexity, coordinate descent requires d-times many iterations?

- if coordinate descent steps are d-times cheaper than both algorithm require $\mathcal{O}((L/\mu)\log(1/\epsilon))$
- but Lipschitz constant L are different: *i.e.*, L_f vs L_c and $L_c \leq L_f$
- extends to Lipschitz sampling, block-coordinate descent, etc.

Finite sum optimization

Consider minizing finite sum

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

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

many machine learning problems fall into this category, for example, consider the least squares objective

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

empirical risk minimization is in general finite-sum minimization

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 observation (n training data points) may represent the distribution.

Motivation: Big-N problems

for fitting a least squares model

- Gradient methods are effective when d is very large: *i.e.*, $\mathcal{O}(nd)$ per iteration instead of $\mathcal{O}(nd^2 + d^3)$ to solve as linear system

But what if number of training exampels n is very large?

- All Gmails, all products on Amazon, all homepages, all images, etc.

Deterministic vs Stochastic methods

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

Deterministic gradient method:

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

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

Stochastic gradient method:

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

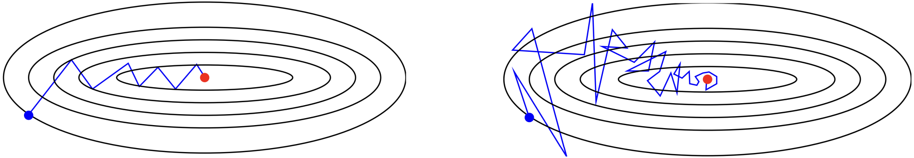
where $i_t = \{1, 2, ..., n\}$ is selected uniformly at random.

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

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

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

Deterministic vs Stochastic methods



(a) GD (b) SGD

Figure: Illustrating deterministic vs stochastic methods; figure from Schmidt lecture note

Deterministic vs Stochastic methods

Illustrating determinstic vs stochastic methods (least squares)

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.

For smooth case, stochastic method is slower.

- $\mathcal{O}(1/\sqrt{t})$ 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 for convex case

we can write

$$\begin{aligned}\|x_{t+1} - x^\star\|_2^2 &= \|x_t - \eta g_t - x^\star\|_2^2 \\ &= \|x_t - x^\star\|_2^2 - 2\eta\big\langle g_t, x_t - x^\star\big\rangle + \eta^2\|g_t\|_2^2\end{aligned}$$

take conditional expectation at iteration t

$$\begin{aligned}\mathbb{E}\big[\|x_{t+1} - x^\star\|_2^2 \mid x_t\big] &= \|x_t - x^\star\|_2^2 - 2\eta\big\langle \mathbb{E}[g_t \mid x_t], x_t - x^\star\big\rangle + \eta^2\mathbb{E}\big[\|g_t\|_2^2 \mid x_t\big] \\ &\le \|x_t - x^\star\|_2^2 - 2\eta\big(f(x_t) - f(x^\star)\big) + \eta^2\mathbb{E}\big[\|g_t\|_2^2 \mid x_t\big]\end{aligned}$$

take total expectation

$$\mathbb{E}\big[\|x_{t+1} - x^\star\|_2^2\big] \le \mathbb{E}\big[\|x_t - x^\star\|_2^2\big] - 2\eta\big(\mathbb{E}[f(x_t)] - f(x^\star)\big) + \eta^2\mathbb{E}\big[\|g_t\|_2^2\big]$$

assuming bounded gradient, *i.e.*, $\mathbb{E}\big[\|g_t\|_2^2\big] \le \sigma^2$ and re-arranging terms yields

$$\mathbb{E}\left[f\Big(\frac{1}{T}\sum_{i=1}^{T} x_t\Big)\right] - f^\star \le \frac{R^2}{2\eta T} + \frac{\eta\sigma^2}{2}$$

Convergence for smooth non-convex case

progress bound

$$f(x_{t+1}) \le f(x_t) - \eta_t \nabla f(x_t)^\top \nabla f_{i_t}(x_t) + \eta_t^2 \frac{L}{2} \|\nabla f_{i_t}(x_t)\|_2^2$$

take the expectation and assume η_t does not depend on i_t

$$\begin{aligned}
\mathbb{E}\big[f(x_{t+1})\big] &\le \mathbb{E}\big[f(x_t) - \eta_t \nabla f(x_t)^\top \nabla f_{i_t}(x_t) + \eta_t^2 \frac{L}{2} \|\nabla f_{i_t}(x_t)\|_2^2\big] \\
&\le f(x_t) - \eta_t \nabla f(x_t)^\top \mathbb{E}\big[\nabla f_{i_t}(x_t)\big] + \eta_t^2 \frac{L}{2} \mathbb{E}\big[\|\nabla f_{i_t}(x_t)\|_2^2\big]
\end{aligned}$$

under uniform sampling (unbiased gradient estimate) it gives

$$\mathbb{E}\big[f(x_{t+1})\big] \le f(x_t) - \eta_t \|\nabla f(x_t)\|_2^2 + \eta_t^2 \frac{L}{2} \mathbb{E}\big[\|\nabla f_{i_t}(x_t)\|_2^2\big]$$

- negative second term: always helps to decrease the objective, and the bigger gradient the more decrease
- positive third term: second moment (or variance) needs to be small

assume bounded variance: for all x

$$\mathbb{E}\big[\|\nabla f_i(x)\|_2^2\big] \leq \sigma^2$$

then the progress bound becomes

$$\mathbb{E}\big[f(x_{t+1})\big] \leq f(x_t) - \eta_t \|\nabla f(x_t)\|_2^2 + \eta_t^2 \frac{L\sigma^2}{2}$$

re-arranging the terms and summing for T iterations will give

$$\min_{t=1,\dots,T} \mathbb{E}\big[\|\nabla f(x_t)\|_2^2\big] \leq \frac{f(x_1) - f^\star}{\sum_{t=1}^T \eta_t} + \frac{L\sigma^2}{2} \frac{\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t}$$

Any questions?