Optimization Theory

Lecture 13

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Large Scale Optimization

In machine learning, we usually learn model parameter $\mathbf{x} \in \mathbb{R}^d$ from

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

where n may be very large.

More generally, we also consider the stochastic optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \mathbb{E}_\xi[F(\mathbf{x};\xi)],$$

where the random variable $\xi \sim \mathcal{D}$.

Stochastic Subgradient Descent

We consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \mathbb{E}_\xi[F(\mathbf{x};\xi)],$$

where each $F(\mathbf{x};\xi)$ is convex but possible nonsmooth.

Algorithm 1 Stochastic Subgradient Descent

1: **Input:** $\mathbf{x}_0$, $\{\eta_t\}_{t=0}^{T-1}$
2: **for** $t = 0, \dots, T-1$
3: draw $\xi_t \sim \mathcal{D}$
4: let $\mathbf{g}_t \in \partial F(\mathbf{x}_t;\xi_t)$
5: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_t$
6: **end for**
7: **Output:** $\bar{\mathbf{x}}_T = \left(\sum_{t=0}^{T-1}\eta_t\right)^{-1}\sum_{t=0}^{T-1}\eta_t\mathbf{x}_t$

Stochastic Subgradient Descent

Suppose each $F(\mathbf{x};\xi)$ is convex and G-Lipschitz such that $\|\mathbf{g}\|_2 \leq G$ for any $\mathbf{g} \in \partial F(\mathbf{x};\xi)$, then stochastic subgradient descent holds

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \sum_{t=0}^{T-1} G^2\eta_t^2}{2\sum_{t=0}^{T-1} \eta_t}.$$

Taking $\eta_t = \eta_0/\sqrt{T}$, we have

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] \leq f(\hat{\mathbf{x}}) + \frac{\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2 + \eta_0^2 G^2}{2\eta_0\sqrt{T}}.$$

Stochastic Subgradient Descent

We additionally suppose each $f(\mathbf{x})$ is μ-strongly convex and take $\eta_t = 2/(\mu(t+1))$, then we have

$$\mathbb{E}[f(\bar{\mathbf{x}}_T)] \leq f(\hat{\mathbf{x}}) + \frac{2G^2}{\mu(T-1)},$$

where

$$\bar{\mathbf{x}}_T = \sum_{t=0}^{T-1} \frac{t\mathbf{x}_t}{T(T-1)}.$$

Stochastic Gradient Descent

We consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \mathbb{E}_\xi[F(\mathbf{x};\xi)],$$

where each $F(\mathbf{x};\xi)$ is L-smooth and convex.

We consider mini-batch stochastic gradient descent.

Algorithm 2 Mini-Batch Stochastic Gradient Descent

1: **Input:** $\mathbf{x}_0$, $\{\eta_t\}_{t=0}^{T-1}$, b
2: **for** $t = 0, \dots, T-1$
3: draw $\xi_{t,1}, \dots, \xi_{t,b} \overset{\text{i.i.d}}{\sim} \mathcal{D}$
4: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot \frac{1}{b}\sum_{i=1}^{b} \nabla F(\mathbf{x}_t; \xi_{t,i})$
5: **end for**
6: **Output:** $\bar{\mathbf{x}}_T$ be weighed average of $\{\mathbf{x}_t\}_{t=0}^{T-1}$

Running mini-batch SGD with $\eta_t = \eta \le 1/(3L)$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*) \le \frac{3\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2}{2\eta T} + \frac{3V^*\eta}{b},$$

where $V^* = \mathbb{E}_\xi \left\|\nabla F(\mathbf{x}^*;\xi) - \nabla f(\mathbf{x}^*)\right\|_2^2$ and

$$\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}_t.$$

We consider μ-strongly convex case

1. Taking $\eta_t = \eta \leq 1/(2L)$, we have

$$\mathbb{E}\left\|\mathbf{x}_T - \mathbf{x}^*\right\|_2^2 \leq (1 - 2\eta\mu)^T \left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{2\eta V^*}{2b\mu}.$$

2. Taking $\eta_t = 2/(8L + \mu t)$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \frac{4L}{T}\left\|\mathbf{x}_0 - \mathbf{x}^*\right\|_2^2 + \frac{4V^*\ln(T+1)}{b\mu T}.$$

Outline

1. Stochastic Gradient Decent

2. Variance Reduction Methods

Stochastic Gradient Descent

We consider the finite-sum problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

where $f:\mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and $f_i:\mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth.

The convergence of SGD

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f_i(\mathbf{x}_t)$$

requires η_t converging to zero.

Variance Reduction Methods

We hope the iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{v}_t$$

such that $\mathbf{v}_t$ converges to $\mathbf{0}$ when $\mathbf{x}_t$ converges to $\mathbf{x}^*$.

There are several variance reduction methods:

1. SAG (Stochastic Average Gradient)
2. SVRG (Stochastic Variance Reduced Gradient)
3. SAGA (What is the full name?)
4. Katyusha (A Russian of Soviet era folk-based song)
5. SARAH (StochAstic Recursive grAdient algoritHm)
6. SPIDER (Stochastic Path-Integrated Differential EstimatoR)

Stochastic Average Gradient (SAG)

The SAG iterations take the form

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_{i,t},$$

where at each iteration a random index i_t is selected and we set

$$\mathbf{g}_{i,t} = \begin{cases} \nabla f_i(\mathbf{x}_t) & \text{if} \quad i = i_t, \\ \mathbf{g}_{i,t-1} & \text{otherwise.} \end{cases}$$

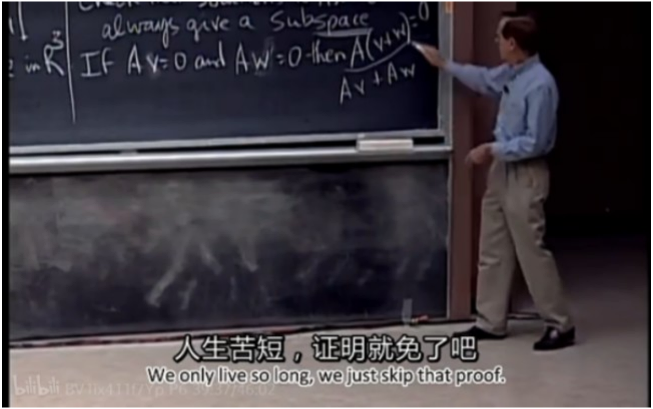
Taking $\eta_t = 1/(16L)$, we have

$$\mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*) \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{n}\right\}\right)^t C_0$$

for some constant $C_0 > 0$.

Stochastic Average Gradient (SAG)

"The analysis of SAG is notoriously difficult, which is perhaps due to the estimator of gradient being biased." — Francis Bach



Stochastic Variance Reduced Gradient (SVRG)

We keep a snap shot point $\tilde{\mathbf{x}}$ and maintain

$$\tilde{\boldsymbol{\mu}} = \nabla f(\tilde{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\mathbf{x}}).$$

We apply the update

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\boldsymbol{\mu}}),$$

where i is randomly sampled from $\{1, \dots, n\}$.

If $\mathbf{x}_t$ and $\tilde{\mathbf{x}}$ tends to $\mathbf{x}^*$, then

$$\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\boldsymbol{\mu}} \to 0.$$

We also have

$$\mathbb{E}_i\left[\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\boldsymbol{\mu}}\right] = \nabla f(\mathbf{x}_t).$$

Stochastic Variance Reduced Gradient (SVRG)

Algorithm 3 Stochastic Variance Reduced Gradient

Input: $\mathbf{x}_0$, η, m
$\tilde{\mathbf{x}}^{(0)} = \mathbf{x}_0$
for $s = 0, \dots, S-1$
  $\tilde{\boldsymbol{\mu}} = \nabla f(\tilde{\mathbf{x}}^{(s)})$
  $\mathbf{x}_0 = \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{(s)}$
  for $t = 0, \dots, m-1$
    draw i_t from $\{1, \dots, n\}$ uniformly
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \tilde{\boldsymbol{\mu}})$,
  end for
  Option I: $\tilde{\mathbf{x}}^{(s)} = \mathbf{x}_m$
  Option II: $\tilde{\mathbf{x}}^{(s)} = \mathbf{x}_t$ for randomly chosen $t \in \{0, \dots, m-1\}$
end for
Output: $\tilde{\mathbf{x}}^{(S)}$

Stochastic Variance Reduced Gradient (SVRG)

Assume $\eta = \Theta(1/L)$ and m is sufficient large so that

$$\rho = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$

then SVRG holds that

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)\right] \leq \rho^s(f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*)).$$

The incremental first-order oracle complexity to achieve

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}^{(s)}) - f(\mathbf{x}^*)\right] \leq \epsilon$$

is at most $\mathcal{O}((\kappa + n)\log(1/\epsilon))$.