IE 598: Big Data Optimization **Fall 2016**

Lecture 22: Stochastic Optimization II – November 08

Lecturer: Niao He *Scriber: Jialin Song*

Overview: In the last lecture we have introduced two approaches i.e., sample average approximation (SAA) and stochastic approximation (SA) to solve the stochastic optimization problems. In this lecture, we continue our discussion on stochastic approximation (SA), which is also known as stochastic gradient descent. We also extend our scope to solving non-smooth and general convex stochastic problems with mirror descent stochastic approximation. Finally, we discuss briefly on how to further improve stochastic gradient descent.

22.1 Recap of SAA and SGD

The stochastic optimization problem can often be formulated as follows

$$\min_{x\in X} f(x) := \mathbb{E}_\xi[F(x,\xi)] \tag{P}$$

We have discussed two approaches to tackle stochastic optimization problem:

Sample Average Approximation (SAA) First, we obtain N i.i.d. sample $\xi_1, \xi_2, ..., \xi_n$ of random variable ξ from a distribution P; we then adopt some first order or second order method to solve the following deterministic optimization problem

$$\min_{x\in X} f^N(x) = \frac{1}{N}\sum_{i=1}^{N} F(x,\xi_i) \tag{SAA}$$

Suppose the method produces a candidate solution x_t when solving (SAA) after t iterations.

- The convergence performance can be measured as the following way:

$$f(x_t) - f(x_*) = \underbrace{f(x_t) - f(x_*^N)}_{\text{optimization error}} + \underbrace{f(x_*^N) - f(x_*)}_{\text{estimation error}}$$

  Note that the optimization error depends on the algorithms utilized to solve the deterministic problem and sample size N, yet the estimation error only depends on N.

- In the last lecture, we show that if the sample size is

$$N = \mathcal{O}\left(\frac{\max_{x\in X} \mathrm{Var}[F(x,\xi)]}{\epsilon^2}\log(\frac{|X|}{\alpha})\right)$$

  then we have $\mathbb{P}(f(x_*^N) \le f(x_*) + \epsilon) \ge 1-\alpha$. Note that this only accounts to the estimation error.

Stochastic Approximation (SA, or SGD) The algorithm directly solves the original problem (P):

$$x_{t+1} = \Pi_X(x_t - \gamma_t \nabla F(x_t, \xi_t))$$

where $\nabla F(x_t, \xi_t)$ is called stochastic gradient.

- When f is L-smooth and μ-strongly convex, $x_* \in \text{int(X)}$, $\gamma_t = \mathcal{O}(\frac{1}{\mu t})$, we have the following result on the expectation of $f(x_t) - f(x_*)$

$$\mathbb{E}[f(x_t) - f(x_*)] \leq \mathcal{O}\left(\frac{\max_{x \in X} \mathbb{E}[\|\nabla F(x, \xi)\|_2^2]}{\mu^2 t}\right)$$

  Markov inequality gives us

$$\mathbb{P}(f(x_t) - f(x_*) \geq \epsilon) \leq \frac{\mathbb{E}[f(x_t) - f(x_*)]}{\epsilon} \leq \mathcal{O}\left(\frac{\max_{x \in X} \mathbb{E}[\|\nabla F(x, \xi)\|_2^2]}{\mu^2 t \epsilon}\right)$$

  If we set the total number of iterations $T = \mathcal{O}(\frac{M^2}{\mu^2 \epsilon \alpha})$, we have $\mathbb{P}(f(x_t) \leq f(x_*) + \epsilon) \geq 1 - \alpha$.

Deterministic Optimization vs. Stochastic Optimization Recall that for deterministic optimization problem

$$\min_{x \in X} f(x)$$

where f admits exact gradient information, we showed that

- $\diamond$ f is smooth and μ-strongly convex $\Longrightarrow \mathcal{O}((\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2t})$ (Accelerated Gradient Descent)
- $\diamond$ f is smooth and convex $\Longrightarrow \mathcal{O}(\frac{1}{t^2})$ (Accelerated Gradient Descent)
- $\diamond$ f is non-smooth and convex $\Longrightarrow \mathcal{O}(\frac{1}{\sqrt{t}})$ (Subgradient/Mirror Descent)

We see that the complexity depends heavily on the smoothness of the convex function. Also, in the case of stronly convex and smooth problems, there is a huge discrepancy between the linear rate by GD and sublinear rate by SGD. In this lecture, we intend to address the following questions?

- Is SGD an optimal algorithm for solving stochastic optimization problem?
- Does smoothness make any difference of the convergence rate?
- What happens when f is some general convex function?
- How can we improve upon SGD?

22.2 Lower Bounds Results on Stochastic Gradient Descent

In this section, we discuss the lower bound complexity results for solving stochastic optimization problems.

A Simple Example. Let us consider the one-dimensional function $f(x) = \mathbb{E}[\frac{1}{2}(x - \xi)^2]$, where $\xi \sim N(0, 1)$ is a standard normal random variable. Based on stochastic gradient descent we have

$$x_{t+1} = x_t - \gamma_t(x_t - \xi_t)$$

Let $x_1 = 0, \gamma_t = \frac{1}{t}$ and by induction we have

$$x_{t+1} = \frac{1}{t}\sum_{t=1}^{t} \xi_\tau$$

Hence, $x_{t+1} \sim N(0, \frac{1}{t})$ is a normal random variable with mean 0 and variance $\frac{1}{t}$.

Since $f(x) = \frac{1}{2}(x^2 + 1)$, it can be easily found taht the optimal solution is $x_* = 0$. Therefore

$$\mathbb{E}[\|x_{t+1} - x_*\|_2^2] = \frac{1}{t}$$

This implies that the $O(1/t)$ rate achieved by SGD is indeed tight.

Recall that for deterministic convex optimization problems, we show that for problems with sufficiently large dimensions, we cannot improve the $O(1/t^2)$ rate for smooth convex problems. In fact, for stochastic optimization problem with strongly convex objectives, we cannot get rates better than $\mathcal{O}(1/t)$, for any arbitrary dimensions.

To achieve a more general characterization for nonsmooth stochastic optimization problems, we introduce the notion of stochastic oracle.

Stochastic Oracle: given an input x, stochastic oracle returns $G(x, \xi)$ s.t.,

$$\mathbb{E}[G(x, \xi)] \in \partial f(x) \text{ and } \mathbb{E}[\|G(x, \xi)\|_p^2] \leq M^2$$

for some positive constant M and some $p \in [1, \infty]$. This characterizes the first and second moment of the estimator of true subgradient.

It was shown in [Nemirovsky and Yudin, 1983] that in the worst case,

- ◇ For convex problems, total number of stochastic oracles required is at least $T = \mathcal{O}(\frac{1}{\epsilon^2})$
- ◇ For strongly convex problems, total number of stochastic oracles required is at least $T = \mathcal{O}(\frac{1}{\epsilon})$

Theorem 22.1 ***[Agarwal et al., 2012]*** *Let $X = B_\infty(r)$ be a ℓ_∞ ball with radius bounded by r.*

(1) $\exists\ c_0 > 0$, $\exists$ a convex function f on R^d, $|f(x) - f(y)| < M\|x - y\|_\infty$, then for any algorithm making T stochastic oracles with $1 \leq p \leq 2$ and generating a solution x_T,

$$\mathbb{E}[f(x_T) - f(x_*)] \geq \min\left\{c_0 Mr\sqrt{\frac{d}{T}},\ \frac{Mr}{144}\right\}$$

(2) $\exists\ c_1,\ c_2 > 0$,$\exists$ a μ-strongly convex function, for any algorithm making T stochastic oracles with $p = 1$ and generating a solution x_T,

$$\mathbb{E}[f(x_T) - f(x_*)] \geq \min\left\{c_1\frac{M^2}{\mu^2 T},\ c_2,\ Mr\sqrt{\frac{d}{T}},\ \frac{M^2}{1152\mu^2 d},\ \frac{Mr}{144}\right\}$$

Moreover, such lower bounds can be attained by the Mirror Descent Stochastic Approximation algorithm, which we discuss in details below.

22.3 Stochastic Mirror Descent

Analogous to the deterministic optimization scenario, **Mirror Descent Stochastic Approximation (a.k.a. Stochastic Mirror Descent)** is adopted to solve non-smooth problems.

Let $\omega(x)$ be a continuously differentiable and 1-strongly convex function w.r.t. some norm $\|\cdot\|$. A simple example of a distance-generating function is $\omega(x) = \frac{1}{2}\|x\|_2^2$. Define function $V(x,y) = \omega(x) - \omega(y) - \nabla\omega(y)^T(x-y)$, which is called the Bregman distance.

The **mirror descent stochastic approximation** works as follows:

$$x_{t+1} = \arg\min_{x\in X}\{V(x,x_t) + \langle \gamma_t G(x_t,\xi_t),\ x\rangle\}$$

Theorem 22.3.1 [Nemirovski et al., 2009]

Let f be a convex function, $\Omega = \max_{x\in X} V(x,x_1)$. Let the candidate solution $\hat{x}_T =$ be the weighted average

$$\hat{x}_T = \frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t}$$

(1) If there exists $M > 0$, s.t., $\mathbb{E}[\|G(x,\xi)\|_*^2] \le M^2$, $\forall x \in X$, then

$$\mathbb{E}[f(\hat{x}_T) - f(x_*)] \le \frac{\Omega + \frac{M^2}{2}\sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}$$

(2) If there exists $M > 0$, s.t., $\mathbb{E}\left[\exp\left\{\|G(x,\xi)\|_*^2/M^2\right\}\right] \le \exp\{1\}$, then

$$\mathbb{P}\left\{f(\hat{x}_T) - f(x_*) \ge \frac{\sqrt{2}M\sqrt{\Omega}(12+2\Omega)}{\sqrt{T}})\right\} \le 2\exp\{-\Omega\}$$

Remark. Note that the condition in (2) essentially states that the stochastic subgradient has a light-tailed distribution. And if $T \ge \mathcal{O}\left(\frac{M^2\Omega}{\epsilon^2}\log^2(\frac{1}{\alpha})\right)$, we have $\mathbb{P}(f(\hat{x}_T) \le f(x_*) + \epsilon) \ge 1-\alpha$. We provide the simple proof for part (1) below.

Proof: Based on the optimality condition of the mirror descent stochastic approximation, we can have

$$\gamma_t(x_t - x_*)^T G(x_t,\xi_t) \le V(x_t,x_*) - V(x_{t+1},x_*) + \frac{\gamma_t}{2}\|G(x_t,\xi_t)\|_*^2 \tag{22.1}$$

Rewrite (22.1) as follows

$$\gamma_t(x_t - x_*)^T g(x_t) \le V(x_t,x_*) - V(x_{t+1},x_*) - \gamma_t(G(x_t,\xi_t) - g(x_t))^T(x_t - x_*) + \frac{\gamma_t}{2}\|G(x_t,\xi_t)\|_*^2 \tag{22.2}$$

where $g(x_t) \in \partial f(x_t)$. Taking summation over $t = 1,...,T$, we have

$$\sum_{t=1}^T \gamma_t(x_t - x_*)^T g(x_t) \le V(x_1,x_*) + \sum_{t=1}^T \frac{\gamma_t^2}{2}\|G(x_t,\xi_t)\|_*^2 - \sum_{t=1}^T \gamma_t(G(x_t,\xi_t) - g(x_t))^T(x_t - x_*) \tag{22.3}$$

Let's set $\hat{x}_T = \frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t}$, and consider the convexity of $f(x)$, we have

$$\sum_{t=1}^T \gamma_t(x_t - x_*)^T g(x_t) \ge \sum_{t=1}^T \gamma_t(f(x_t) - f(x_*)) \ge \left(\sum_{t=1}^T \gamma_t\right)(f(\hat{x}_T) - f(x_*)) \tag{22.4}$$

Combine (22.3) and (22.4), we can get

$$f(\hat{x}_T) - f(x_*) \leq \frac{V(x_1, x_*) + \sum_{t=1}^T \frac{\gamma_t^2}{2} \|G(x_t, \xi_t)\|_*^2 - \sum_{t=1}^T \gamma_t (G(x_t, \xi_t) - g(x_t))^T (x_t - x_*)}{\sum_{t=1}^T \gamma_t} \tag{22.5}$$

Taking expectations on both sides of (22.5), we can have

$$\mathbb{E}[f(\hat{x}_T) - f(x_*)] \leq \frac{\max_{x \in X} V(x, x_1) + \frac{M^2}{2} \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}$$

as desired. ■

22.4 Improving Stochastic Gradient Descent

Although we cannot improve the convergence rate the Stochastic Gradient Descent (SGD) or Stochastic Mirror Descent (SMD) methods, we can still try to accelerate the performance by improving the constant factors. We discuss several strategies below.

- ◇ **Reduce Variance:**
  - **Mini Batch sampling:** use a small batch of samples instead of one to estimate the gradient at every iteration

    $$G(x_t, \xi_t) \Longrightarrow \frac{1}{b} \sum_{i=1}^b G(x_t, \xi_{t,i})$$

    Consequently, the variance of the new stochastic gradient will be $O(b)$ times smaller, i.e. the constant term M^2 in the convergence now reduces to M^2/b.
  - **Importance Sampling:** Instead of sampling from $\xi \sim P$, we can obtain samples from another well defined random variable η with nominal distribution Q, and use a different stochastic gradient,

    $$G(x_t, \xi_t) \Longrightarrow G(x_t, \eta_t) \frac{P(\eta_t)}{Q(\eta_t)}$$

    The variance of the new stochastic gradient under properly chosen distribution Q could be smaller.
- ◇ **Adaptive Stepsize** The traditional fixed stepsize $\gamma_t = \frac{1}{\mu t}$ may be too small so that the efficiency of the stochastic gradient descent approach can be compromised. One may instead select the stepsize adaptively to optimize the progress at each iteration. For instance, in [YNS12], the authors propose to automatically update the stepsize based on the recursion

  $$\gamma_t = \frac{1}{\mu t} \Longrightarrow \gamma_t = \gamma_{t-1}(1 - c\gamma_{t-1})$$

- ◇ **Adaptation of Bregman Distance** One may also adaptively choose the Bregman distance and hope to improve the efficiency. For instance, the AdaGrad algorithm in [DHS11] propose the following

  $$\omega(x) = \frac{1}{2} x^T x \Longrightarrow \omega_t(x) = \frac{1}{2} x^T H_t x, \text{ where } H_t = \delta \mathbf{I} + [\sum_{t=1}^t g_t g_t^T]^{\frac{1}{2}}$$

  where $g_t = G(x_t, \xi_t)$.

References

[NY83] A. Nemirovski and D. Yudin, Problem Complexity and Method Efficiency in Optimization *Wiley, New York, 1983*

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust Stochastic Approximation Approach to Stochastic Programming, *SIAM J. Optim. 19(4) (2009), pp. 1574?1609*

[YNS12] F. Yousefian, A. Nedich and U.V. Shanbhag, On Stochastic Gradient and Subgradient Methods with Adaptive Steplength sequences, *Automatica 48 (1) 56-67, 2012*

[ABRW12] A. Agarwal, P. Bartlett, P. Ravikumar and M. Wainwright, Information-theoretic Lower Bounds on the Oracle Complexity of Convex Optimization *IEEE Transactions on Information Theory, 58(5), 2012*

[DHS12] J. Duchi, E. Hazan, and Y. Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research, 12(Jul), 2121-2159, 2012*