Introductory Lectures on Optimization

Stochastic Optimization (1)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

December 16, 2024

Outline

Part I
Stochastic Optimization Formulation

Introduction

Uncertainty: When modeling and solving a realistic optimization problem, we always face uncertainty presenting in the problem, e.g., unknown parameters.

Uncertainty makes optimization problems intractable.

Stochastic Optimization: In this section, we introduce the formulation with uncertainty and some examples stimulating from real-life decision making request to give a good description of stochastic optimization.

Introduction

The stochastic optimization problem is often formulated as the following format

$$\min_{x\in\mathcal{X}} \left\{ f(x) \triangleq \mathbb{E}_\xi[F(x,\xi)] \right\}, \tag{1}$$

where $F(x,\xi)$ is a function involving our decision variable (vector) x, and a random variable (vector) ξ which is some well-defined random variable with support $\Omega \subseteq \mathbb{R}^d$ and a distribution P. We minimize $f(x)$ to get the optimization solution, where the function $f(x)$ without uncertainty is given by taking the expectation of $F(x,\xi)$ over ξ, i.e.,

$$\mathbb{E}_\xi[F(x,\xi)] = \int_{\xi\in\Omega} F(x,\xi) dP(\xi).$$

Remark. Note that $F(x,\xi)$ is convex for any $\xi \in \Omega$, and by the calculus of convex function, it can imply that $f(x)$ is convex. However, vice versa is not true.

Example: Newspaper-vendor model

A newspaper vendor needs to decide how many copies of today's newspaper to stock in order to meet the uncertain demand and to maximize profit meanwhile. Suppose the number of newspaper to be stocked q is the vendor's decision variable, the purchase price per newspaper that the vendor needs to pay is denoted by c, the selling price per newspaper that consumers need to pay is denoted by p. We use ξ to represent the consumers' random demand for newspaper. Then, the Newspaper-vendor model can be formulated as

$$\max_{q} \ \mathbb{E}_{\xi}[p \cdot \min(q, \xi) - c \cdot q].$$

The cost is $c \cdot q$ which the vendor needs to pay for the stocked q copies of newspaper. The revenue the vendor can get is $p \cdot \min(q, \xi)$ which means the vendor cannot sell more than the stocked number of newspaper and the consumers' demand.

Example: Markowitz Model

Suppose an investor wants to invest in different stocks with random returns in order to maximize the returns generating by buying these stocks, and to minimize the variance of the random returns meanwhile. We use w to denote the weights of stock and take it as our decision, and use r to denote the random returns. Then the Markowitz model is

$$\max_{w \geq 0, \sum w_i = 1} \mathbb{E}_r[w^\top \cdot r] - \lambda \cdot \mathrm{Var}[w^\top \cdot r],$$

where λ is a parameter used to penalize the variance of the random returns. We again translate the maximization problem into a minimization problem by adding a minus sign before the objective function, and it can be show that $-\mathbb{E}_r[w^\top \cdot r] + \lambda \cdot \mathrm{Var}[w^\top \cdot r]$ is a convex function.

Example: Expected Risk Minimization

In many machine learning problems, we hope to minimize the expected loss given by the loss function $l(f(x), y)$ through choosing a suitable function f from the function set F where x and y are random input data. The formulation is

$$\min_{f \in \mathcal{F}} \mathbb{E}_{x,y}[l(f(x), y)]$$

Notice that the three examples are all stochastic optimization problems. A question appears: how to solve stochastic optimization problems? It is difficult to use the methods we have learned for deterministic problems here because it often is intractable to compute the gradient of $f(x)$ which involves an integration. Suppose $F(x, \xi)$ is differential for any $\xi \in \Omega$, the gradient $\nabla f(x) = \int_{\xi \in \Omega} \nabla F(x, \xi) dP(\xi)$ can be difficult to compute.

Part II
Sample Average Approximation

Monte Carlo Sampling

A natural way to address the stochastic optimization problem, is to use Monte Carlo sampling. Let $\xi_1, \ldots, \xi_N$ be independently and identically distributed (i.i.d.) random sample of the random variable (vector) ξ. We consider the following estimation of the original problem

$$\min_{x \in \mathcal{X}} f^N(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi^i). \tag{2}$$

Here $\mathcal{X}$ is a nonempty closed subset of $\mathbb{R}^n$, ξ is a random vector whose probability distribution P is supported on a set $\Xi \subset \mathbb{R}^d$, and $F : \mathcal{X} \times \Xi \to \mathbb{R}$ This is known as the sample average approximation.

Monte Carlo Sampling

Let us observe that we can write the sample average function as the expectation

$$f^N(x) = \mathbb{E}_{P_N}[F(x, \xi)]$$

taken with respect to the empirical distribution (measure)

$$P_N := N^{-1} \sum_{j=1}^{N} \delta(\xi^j).$$

Therefore, for a given sample, the SAA problem can be considered as a stochastic programming problem with respective scenarios $\xi^1, ..., \xi^N$, each taken with probability $1/N$.

Monte Carlo Sampling

As with data vector ξ, the sample $\xi_1, \ldots, \xi_N$ can be considered from two points of view: as a sequence of random vectors or as a particular realization of that sequence. Which of these two meanings will be used in a particular situation will be clear from the context.

1. The SAA problem is a function of the considered sample and in that sense is random.
2. For a particular realization of the random sample, the corresponding SAA problem is a stochastic programming problem with respective scenarios $\xi^1, \ldots, \xi^N$ each taken with probability $1/N$.

We always assume that each random vector ξ^j in the sample has the same (marginal) distribution P as the data vector ξ. If, moreover, each ξ^j, $j = 1, \ldots, N$, is distributed independently of other sample vectors, we say that the sample is independently identically distributed (iid).

Statistic Properties

By the Law of Large Numbers (LLN) we have that, under some regularity conditions, $f^N(x)$ converges pointwise w.p. 1 to $f(x)$ as $N \to \infty$. In particular, by the classical LLN this holds if the sample is iid. Moreover, under mild additional conditions the convergence is uniform.

We also have that $\mathbb{E}[f^N(x)] = f(x)$, i.e., $f^N(x)$ is an unbiased estimator of $f(x)$.

Therefore, it is natural to expect that the optimal value and optimal solutions of the SAA converge to their counterparts of the true problem as $N \to \infty$.

We denote by f_* and $\mathcal{X}_*$ the optimal value and the set of optimal solutions, respectively, of the true problem and by f_*^N and $\mathcal{X}_*^N$ the optimal value and the set of optimal solutions, respectively, of the SAA problem.

Statistic Properties

Remark.
Basically, pointwise convergence means that for each x and ϵ, you can find an N such that by N iterations some distance is samller than ϵ . Here the N is allowed to depend both on x and ϵ.

In uniform convergence the requirement is strengthened. Here for each ϵ you need to be able to find an N such that by N iterations some distance is samller than ϵ for all x in the domain of the function. In other words N can depend on ϵ but not on x.

For example, the functions $f_n(x) = \frac{x}{n}$ converge pointwise to the zero function on $\mathbb{R}$, but do not converge uniformly. For example, if we choose $\epsilon = 1$, then the convergence condition boils down to $N > |x|$. For each $x \in \mathbb{R}$ we can find such an N easily, but there's no N that works simultaneously for every x.

Consistency of SAA Estimators

Theorem 1

Suppose that there exists a compact set $C \subset \mathbb{R}^n$ such that

1. the set $\mathcal{X}_*$ of optimal solutions of the true problem is nonempty and is contained in C,
2. the function $f(x)$ is finite valued and continuous on C,
3. $f^N(x)$ converges to $f(x)$ w.p. 1 as $N \to \infty$, uniformly in $x \in C$, and
4. w.p. 1 for N large enough the set $\mathcal{X}_*^N$ is nonempty and $\mathcal{X}_*^N \subset C$.

Then $f_*^N \to f_*$ and $\mathbb{D}(\mathcal{X}_*^N - \mathcal{X}_*) \to 0$ w.p. 1 as $N \to \infty$.

Asymptotics of the SAA Optimal Value

Two assumptions for asymptotics of the SAA optimal value as follows.

(A1) For some point $\tilde{x} \in \mathcal{X}$ the expectation $\mathbb{E}[F(\tilde{x}, \xi)^2]$ is finite.

(A2) There exists a measurable function $C : \Xi \to \mathbb{R}_+$ such that $\mathbb{E}[C(\xi^2)]$ is finite and

$$|F(x, \xi) - F(x', \xi)| \leq C(\xi) \left\|x - x'\right\|$$

for all $x, x' \in \mathcal{X}$ and a.e. $\xi \in \Xi$.

Asymptotics of the SAA Optimal Value

Theorem 2

Let f_*^N be the optimal value of SAA problem. Suppose that the sample is iid, the set $\mathcal{X}$ is compact, and assumptions (A1) and (A2) are satisfied. Then the following holds:

$$f_*^N = \inf_{x \in \mathcal{X}_*} f^N(x) + o_p(N^{-\frac{1}{2}}).$$

If, moreover, $\mathcal{X}_* = \{\bar{x}\}$ is a sigleton, then

$$N^{\frac{1}{2}} \cdot (f_*^N - f_*) \xrightarrow{D} \mathcal{N}(0, \delta^2(\bar{x})).$$

(1) convergence in distribution; (2) This theorem implies the bias $\mathbb{E}[f_*^N - f_*]$ is of order $o(N^{-\frac{1}{2}})$. (3) It is written that $Z_k = o_p(Y_k)$ if for any $\epsilon > 0$ it holds that $\lim_{k\to\infty} P(|Z_k/Y_k| > \epsilon) = 0$.

Asymptotics of the SAA Optimal Value

Remark.

A sequence $X_1, X_2, \ldots$ of real-valued random variables, with cumulative distribution functions $F_1, F_2, \ldots$, is said to converge in distribution, or converge weakly, or converge in law to a random variable X with cumulative distribution function F if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every number $x \in \mathbb{R}$ at which F is continuous.

Conclusion

1. SAA is very general. Even if our objective is a non-convex or even comes from discrete optimization, we can still use SAA method to solve the stochastic optimization problem, and the related results still hold. In addition, we can combine any existing algorithms to solve SAA problems.
2. In practice, it is difficult to determine an appropriate sample size N that we need.
3. On the one hand, large sample size N can help us to get a high accuracy, while on the other hand, solving problem (2) with large N can be expensive. For instance, computing the gradient $\nabla f^N(x)$ requires to compute the sum of $O(N)$ gradients.
4. The SAA approach only works with batch data, and cannot handle streaming/online data.

Part III
Stochastic Gradient Descent

Stochastic Gradient Descent

Stochastic Approximation (SA) is another popular method to solve the stochastic optimization problem and it dates back to 1951 by Robbins and Monro [1951].

Assume $F(x, \xi)$ is differential with x for any $\xi \in \Omega$, the idea of Classic Stochastic Approximation is that give a sample realization ξ_t, we update the decision variable (vector) x_{t+1} at iteration $t + 1$ following the rule:

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

which is also known as the stochastic gradient descent (SGD) method.

Stochastic Gradient Descent

Remarks:

1. The stochastic gradient is unbiased, i.e., $\mathbb{E}[\nabla F(x, \xi)] = \nabla f(x)$.
2. We need a decreasing sequence $\{\gamma_t\}$ and $\gamma_t \to 0$ as t goes to infinity to ensure convergence. Because at optimality, we have $x_* = x_* - \gamma \nabla F(x_*, \xi)$. However, since $\nabla f(x_*, \xi)$ is random, we cannot guarantee that $\nabla F(x_*, \xi) = 0$, $\forall \xi \in \Omega$. Hence, we need $\gamma_t \to 0$ as $t \to \infty$.
3. For the SA algorithm, the iterate $x_t = x_t(\xi_{[t-1]})$ is a function of the $i.i.d.$ historic sample $\xi_{[t-1]} = (\xi_1, \ldots, \xi_{t-1})$ of the generated random process, so x_t and $f(x_t)$ are random variables. We cannot use the previous error functions to measure the optimality, e.g. $[f(x_t) - f_*]$ and $\|x_t - x_*\|_2^2$. Instead, a more appropriate criterion would be consider the expectation or high probability results.

Main Theoretical Result

Consider

$$\min_x f(x) := \mathbb{E}[F(x, \xi)].$$

1. f: μ-strongly convex, L-smooth,
2. $g(x^t, \xi^t)$: an unbiased estimate of $\nabla f(x^t)$ given $\{\xi^0, \ldots, \xi^{t-1}\}$.
3. for all x,

$$\mathbb{E}\left[\left\|g(x, \xi)^2\right\|_2\right] \leq \delta_g^2 + c_g \left\|\nabla f(x)\right\|_2^2.$$

Main Theoretical Result

Theorem 3

Under the above assumptions, if $\gamma_t = \gamma \leq \frac{1}{Lc_g}$, then SGD achieves

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{\gamma L \delta_g^2}{2\mu} + (1 - \gamma\mu)^t (f(x_1) - f(x^*)).$$

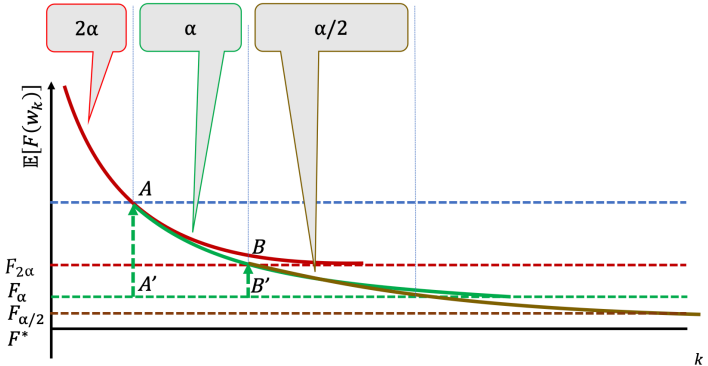
See Theorem 4.6 (Strongly Convex Objective, Fixed Stepsize) of [Bottou et al., 2018] for the proof.

Main Theoretical Result

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{\gamma L \delta_g^2}{2\mu} + (1 - \gamma\mu)^t (f(x_1) - f(x^*)).$$

1. Fast (linear) convergence at the very beginning.
2. Converges to some neighborhood of x^*: variation in gradient computation prevents further progress.
3. When gradient computation is noiseless (i.e. $\delta_g = 0$), it converges linearly to optimal points.
4. Smaller stepsizes γ yield better converging points.

Main Theoretical Result



The above picture is from section 4.2 of [Bottou et al., 2018].

Main Theoretical Result

Theorem 4

Assume $f(x)$ is μ-strongly convex, and exists $M > 0$, subject to $\mathbb{E}\left[\|\nabla F(x,\xi)\|_2\right] \leq M^2$, $\forall x \in \mathcal{X}$, then SA method with $\gamma_t = \gamma/t$ at iteration t where $\gamma > 1/2\mu$ satisfies the following two properties:

1. $\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$, where $C(\gamma) = \max\left\{\frac{\gamma^2 M^2}{2\mu\gamma - 1}, \|x_1 - x_*\|_2^2\right\}$, and
2. If $f(x)$ is L-smooth and $x_* \in \text{int}(\mathcal{X})$, then

$$\mathbb{E}[f(x_t) - f_*] \leq \frac{LC(\gamma)}{2t}.$$

Main Theoretical Result

$$\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$$

Proof. (1) For any given x_t and $\xi_{[t-1]}$, we want to calculate x_{t+1} by a sample ξ_t generated in this iteration, and the distance of x_{t+1} to the optimal x_* is

$$\begin{aligned}
\|x_{t+1} - x_*\|_2^2 &= \left\| \prod_{\mathcal{X}}(x_t - \gamma_t \nabla F(x_t, \xi_t)) - \prod_{\mathcal{X}}(x_*) \right\|_2^2 \quad (\text{ by definition}) \\
&\leq \|x_t - \gamma_t \nabla(x_t, \xi_t) - x_*\|_2^2 \quad (\text{ by non-expansion of projection}) \\
&= \|x_t - x_*\|_2^2 - 2\gamma_t \langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle + \gamma_t^2 \left\| \nabla F(x_t, \xi_t) \right\|_2^2 .
\end{aligned}$$

Main Theoretical Result

$$\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$$

Proof. (1) (continued) Because $\xi_{[t-1]}$ are samples generated from a random process, and x_t is a function of $\xi_{[t-1]}$, we take expectation on both sides of the above inequality to get

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x_*\|_2^2\right] &\leq \mathbb{E}\left[\|x_t - x_*\|_2^2\right] - 2\gamma_t \mathbb{E}[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle] \\
&\quad + \gamma_t^2 \mathbb{E}[\|\nabla F(x_t, \xi_t)\|_2^2] \\
&\leq \mathbb{E}\left[\|x_t - x_*\|_2^2\right] - 2\gamma_t \mathbb{E}[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle] \\
&\quad + \gamma_t^2 M^2. \qquad (3)
\end{aligned}$$

Main Theoretical Result

$$\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$$

Proof. (1) (continued) We claim that $\mathbb{E}[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle] = \mathbb{E}[\langle \nabla f(x_t),\ x_t - x_* \rangle]$, which is shown below. Because $x_t = x_t(\xi_{[t-1]})$ is independent of ξ_t, we have

$$\begin{aligned}
\mathbb{E}[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle] &= \mathbb{E}\left[\mathbb{E}\left[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle | \xi_{[t-1]}\right]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}[\nabla F(x_t, \xi_t) | \xi_{[t-1]}],\ x_t - x_* \rangle\right] \\
&= \mathbb{E}\left[\langle \nabla f(x_t),\ x_t - x_* \rangle\right] \quad \text{(Independent)}.
\end{aligned}$$

Remark: Tower property $\mathbb{E}_Y\left[\mathbb{E}_{X|Y}[X|Y]\right] = \mathbb{E}_X[X]$.

Main Theoretical Result

Remark. Here, $X = \xi_{[t]}$ and $Y = \xi_{[t-1]}$. Thus,

$$\begin{aligned}
\mathbb{E}_{\xi_{[t]}}[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle] &= \mathbb{E}_{\xi_{[t-1]}}\left[\mathbb{E}_{\xi_{[t]}|\xi_{[t-1]}}\left[\langle \nabla F(x_t, \xi_t),\ x_t - x_* \rangle | \xi_{[t-1]}\right]\right] \\
&= \mathbb{E}_{\xi_{[t-1]}}\left[\langle \mathbb{E}_{\xi_{[t]}|\xi_{[t-1]}}[\nabla F(x_t, \xi_t)|\xi_{[t-1]}],\ x_t - x_* \rangle\right] \\
&= \mathbb{E}_{\xi_{[t-1]}}\left[\langle \mathbb{E}_{\xi_t}[\nabla F(x_t, \xi_t)],\ x_t - x_* \rangle\right] \\
&= \mathbb{E}_{\xi_{[t-1]}}\left[\langle \nabla f(x_t),\ x_t - x_* \rangle\right] \quad \text{(Independent)}.
\end{aligned}$$

Main Theoretical Result

$$\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$$

Proof. (1) (continued) We also claim that $\mathbb{E}[\langle \nabla f(x_t),\ x_t - x_* \rangle] \geq \mu \mathbb{E}[\|x_t - x_*\|_2^2]$, we have, for $\forall x, y \in \mathcal{X}$,

$$\begin{aligned} f(x) \text{ is } \mu\text{-strongly convex } &\Leftrightarrow \langle \nabla f(x_t) - \nabla f(x_*),\ x_t - x_* \rangle \geq \mu \|x_t - x_*\|_2^2 \\ &\Leftrightarrow \langle \nabla f(x_t),\ x_t - x_* \rangle \geq \mu \|x_t - x_*\|_2^2 + \langle \nabla f(x_*),\ x_t - x_* \rangle. \end{aligned}$$

By the optimality of x_*, we have $\langle \nabla f(x_*),\ x - x_* \rangle \geq 0$. Combining the optimality condition and the inequality above, it follows that

$$\langle \nabla f(x_t),\ x_t - x_* \rangle \geq \mu \|x_t - x_*\|_2^2 + \langle \nabla f(x_*),\ x_t - x_* \rangle \geq \mu \|x_t - x_*\|_2^2 . \qquad (4)$$

Main Theoretical Result

$$\mathbb{E}[\|x_t - x_*\|_2^2] \leq \frac{C(\gamma)}{t}$$

Proof. (1) (continued) Putting (4) back to (3), we have

$$\mathbb{E}[\|x_{t+1} - x_*\|_2^2] \leq (1 - 2\mu\gamma_t)\mathbb{E}[\|x_t - x_*\|_2^2] + \gamma_t^2 M^2$$

Remember that we choose $\gamma_t = \gamma/t$ for iteration t, and $\gamma \geq 1/2\mu$, the iequality above is equivalent with

$$\mathbb{E}[\|x_{t+1} - x_*\|_2^2] \leq (1 - \frac{2\mu\gamma}{t})\mathbb{E}[\|x_t - x_*\|_2^2] + \frac{\gamma^2 M^2}{t^2}.$$

By induction, we conclude that $\mathbb{E}[\|x_{t+1} - x_*\|_2^2] \leq \frac{C(\gamma)}{t}$, where $C(\gamma) = \max\{\frac{\gamma^2 M^2}{2\mu\gamma - 1}, \|x_1 - x_*\|_2^2\}$.

Main Theoretical Result

$$\mathbb{E}[f(x_t) - f_*] \leq \frac{LC(\gamma)}{2t}.$$

Proof. (2)

Give any x_t and ξ_{t-1}, it can be shown that, $\forall x_t \in \text{int}(\mathcal{X})$,

$$f(x_{t+1}) - f(x_*) \leq \frac{L}{2} \|x_{t+1} - x_*\|_2^2 .$$

Taking expectation on both sides and combining the result in (1), we get

$$\mathbb{E}[f(x_{t+1}) - f_*] \leq \frac{L}{2}\mathbb{E}[\|x_{t+1} - x_*\|_2^2] \leq \frac{LC(\gamma)}{2t}.$$

□

Main Theoretical Result

Remark 3.

1. Note that from Theorem (4), in order to get ϵ-accuracy, we need $O(\frac{1}{\epsilon})$ number of samples in SA (SGD) method, while we need $O(\frac{1}{\epsilon^2})$ number of samples if using SAA method. (注：一个是优化精度，一个是泛化精度。)
2. In the deterministic case, for strongly convex objective function, the error is $\|x_t - x_*\|_2^2 \leq O((\frac{L-\mu}{L+\mu})^{2t})$ which gives linear convergence rate. However, in the stochastic case, the expected error is $\mathbb{E}[\|x_t - x_*\|_2^2] \leq O(\frac{1}{t})$ which gives a sublinear convergence rate.

Main Theoretical Result

For general convex settings, consider

$$\min_x f(x) := \mathbb{E}[F(x, \xi)].$$

Assume that

1. f: convex,
2. $g(x_t, \xi_t)$: an unbiased estimate of $\nabla f(x_t)$ given $\{\xi_0, \ldots, \xi_{t-1}\}$,
3. for all x,
$$\mathbb{E}\left[\left\|g(x, \xi)^2\right\|_2\right] \leq \delta_g^2.$$

Main Theoretical Result

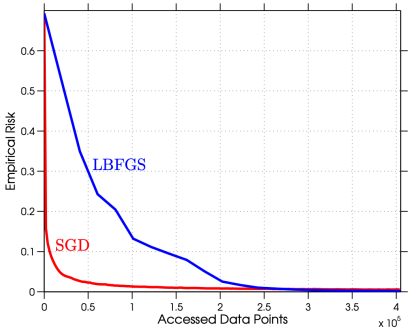
Suppose that we return a weighted average $\tilde{x}_t = \sum_{k=0}^{t} \frac{\gamma_k}{\sum_{j=0}^{t} \gamma_k} x_k$.

Theorem 5 (Theorem 11.3 of Chen)

Under the above assumptions, one has

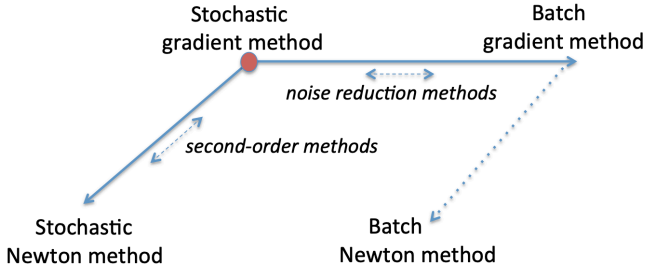
$$\mathbb{E}[f(\tilde{x}_t) - f(x_*)] \leq \frac{1}{2} \frac{\mathbb{E}\left[\|x_0 - x_*\|_2^2\right] + \delta_g^2 \sum_{k=0}^{t} \gamma_k^2}{\sum_{k=0}^{t} \gamma_k}.$$

Main Theoretical Result



(From [Bottou et al., 2018]) Empirical risk R_n as a function of the number of accessed data points (ADP) for a batch L-BFGS method and the stochastic gradient (SG) method on a binary classification problem with a logistic loss objective and the RCV1 dataset. SG was run with a fixed stepsize of $\gamma_k \equiv 4$.

Main Theoretical Result



(From [Bottou et al., 2018]) Schematic of a two-dimensional spectrum of optimization methods for machine learning. The horizontal axis represents methods designed to control stochastic noise; the second axis, methods that deal with ill conditioning.

Lower Bounds on SGD

A Simple Example.

Let us consider the one-dimensional function $f(x) = \mathbb{E}\left[\frac{1}{2}(x-\xi)^2\right]$, where $\xi \sim \mathcal{N}(0,1)$ 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

$$x_{t+1} = \frac{1}{t}\sum_{t=1}^{t} \xi_t.$$

Lower Bounds on SGD

A Simple Example. (continued)

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

Since $\mathbb{E}[\xi^2] = \mathrm{Var}[\xi] + \mathbb{E}[\xi]^2$, $f(x) = \frac{1}{2}(x^2 + 1)$. It can be easily found that 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.

See Theorem 4 (at most)

$$\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$$

Lower Bounds on SGD

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.,

$$\mathbb{E}[G(x, \xi)] \in \partial f(x),$$

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.

Lower Bounds on SGD

In fact, for stochastic optimization problem with strongly convex objectives, we cannot get rates better than $O(1/t)$, for any arbitrary dimensions. It was shown in [Nemirovski and Yudin, 1983] that in the worst case,

1. For convex problem, total number of stochastic oracles required is at least

$$T = O\left(\frac{1}{\epsilon^2}\right).$$

2. For strongly convex problem, total number of stochastic oracles required is at least

$$T = O\left(\frac{1}{\epsilon}\right).$$

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.

Lower Bounds on SGD

Theorem 6 ([Agarwal et al., 2010], (for $p > 2$, see the original paper.))

Let $X = B_\infty(r)$ be a ℓ_∞ ball with radius bounded by r.

- $\exists c_0 > 0, \exists$ a convex funciton f on $\mathbb{R}^d$, $|f(x) - f(y)| < M \|x - y\|_q$ where $1/p + 1/q = 1$, then for any algorithm making t stochastic oracles with $1 \leq p \leq 2$ and generating a solution x_t,

$$\mathbb{E}[f(x_t) - f(x_*)] \geq \min \left\{ c_0 M r \sqrt{\frac{d}{t}}, \frac{Mr}{144} \right\}.$$

- $\exists c_1, c_2 > 0, \exists$ a μ-strongly convex function f, $|f(x) - f(y)| < M \|x - y\|_q$ where $1/p + 1/q = 1$, then for any algorithm making t stochastic oracles with $p = 1$ and generating a solution x_t,

$$\mathbb{E}[f(x_t) - f(x_*)] \geq \min \left\{ c_1 \frac{M^2}{\mu^2 t}, c2Mr\sqrt{\frac{d}{t}}, \frac{M^2}{1152\mu^2 d}, \frac{Mr}{144} \right\}.$$

References I

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Yuxin Chen. Large-scale optimization for data science, ele522. URL `http://www.princeton.edu/~yc5/ele522_optimization/`.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *arXiv e-prints*, pages arXiv–1009, 2010.

References II

Ernest K Ryu and Wotao Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, 2022.

A Kleywegt and A Shapiro. Chapter 101: stochastic optimization. *Work done at the School of Industrial and Systems Engineering at Georgia Institute of Technology*, 2000.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

Niao He. Big data optimizaton course, ece236c. URL `http://niaohe.ise.illinois.edu/IE598_2016/index.html`.

Thank You!

Email:qianhui@zju.edu.cn