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

Lecture 21: Stochastic Optimization – November 3

Lecturer: Niao He *Scriber: Juan Xu*

Overview In this lecture, we consider a new type of optimization problem involving uncertainty, which is called stochastic optimization. We introduce two approaches, sample average approximation and stochastic approximation, to solve stochastic optimization problems.

21.1 Introduction

When modeling and solving a realistic optimization problem, we always face uncertainty presenting in the problem, e.g., unknown parameters. The uncertainty makes both the realism and tractability of optimization problems more difficult. 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.

Stochastic Optimization Formulation

The stochastic optimization problem is often formulated as the following format

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

where $F(x,\xi)$ is a function involving our decision variable (vector) x and a random variable (vector) ξ. The random ξ 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.,

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

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

Example 1. (Newsvendor 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 D to represent the consumers' random demand for newspaper. Then, the Newsvendor model can be formulated as

$$\max_q \mathbb{E}_D \left[p \cdot \min(q, D) - c \cdot q\right].$$

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 $p \cdot \min(q, D)$ which means the vendor cannot sell more than the stocked number of newspaper and the consumers' demand.

Note that the maximization problem is equivalent to the following minimization problem by adding a minus sign to the objection function, so it is consistent with the stochastic optimization formulation,

$$\min_q -\mathbb{E}_D \left[p \cdot \min(q, D) - c \cdot q \right].$$

Also note that the objective function of the minimization problem $-\mathbb{E}_D \left[p \cdot \min(q, D) - c \cdot q \right]$ is a convex function now, and the values of the maximization and minimization problem are opposite each other.

Example 2. (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 $\mathbf{w}$ to denote the weights of stock and take it as our decision, and use $\mathbf{r}$ to denote the random returns. Then the Markowitz model is

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

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 $-\mathbb{E}_{\mathbf{r}} \left[\mathbf{w}^\top \cdot \mathbf{r} \right] + \lambda \cdot \mathrm{Var} \left[\mathbf{w}^\top \cdot \mathbf{r} \right]$ is a convex function.

Example 3. (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 $\mathcal{F}$ where x and y are random input data. The formulation is

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

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.

In this lecture, we introduce two approaches

1. sample average approximation
2. stochastic approximation

which can be used to solve stochastic optimization problems.

21.2 Sample Average Approximation

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

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

This i known as the **sample average approximation**.

For simplicity, we assume the feasible set $\mathbf{X}$ is a finite, then original stochastic optimization (P) and its sample average approximation (SAA) have nonempty sets of optimal solutions, denoted by $\mathbf{X}_*$ and $\mathbf{X}_*^N$ respectively. For each element in $\mathbf{X}_*$ and $\mathbf{X}_*^N$, it gives the optimal value of (P) and (SAA) respectively, i.e.,

$$\begin{aligned} f_* &= f(x_*) = \min_{x\in\mathbf{X}} f(x), \quad \forall x_* \in \mathbf{X}_*, \\ f_*^N &= f(x_*^N) = \min_{x\in\mathbf{X}} f^N(x), \quad \forall x_*^N \in \mathbf{X}_*^N. \end{aligned}$$

We define sets of ϵ-optimal solution set $\mathbf{X}_\epsilon$ and $\mathbf{X}_\epsilon^N$ for (P) and (SAA), respectively. That is,

$$\begin{aligned} &\forall \epsilon \geq 0, \quad \mathbf{X}_\epsilon := \{x \in \mathbf{X} : f(x) \leq f_* + \epsilon\}, \\ &\forall \epsilon \geq 0, \quad \mathbf{X}_\epsilon^N := \{x \in \mathbf{X} : f^N(x) \leq f_*^N + \epsilon\}. \end{aligned}$$

An extreme case is that when $\epsilon = 0$, set $\mathbf{X}_\epsilon$ coincides $\mathbf{X}_*$, and $\mathbf{X}_\epsilon^N$ coincides $\mathbf{X}_*^N$.

Proposition 21.1 *The following two properties hold:*

(1) $f_*^N \to f_*$ *w.p.1 as* $N \to \infty$*, and*

(2) $\forall \epsilon \geq 0, \mathbb{P}(\mathbf{X}_\epsilon^N \subseteq \mathbf{X}_\epsilon) = 1$ *as* $N \to \infty$*.*

Proof:

(1) Following from the SLLN (strong law of large numbers), $\forall x \in \mathbf{X}, f^N(x)$ converges to $f(x)$ w.p.1 as $N \to \infty$. In other words, $\forall x \in \mathbf{X}, |f^N(x) - f(x)| \to 0$ w.p.1. as $N \to \infty$. This means as $N \to \infty$, the set $\{x \in \mathbf{X} : |f^N(x) - f(x)| > 0\}$ has measure zero. Since set $\mathbf{X}$ is finite, and the union of a finite number of sets each of which has measure zero shows a measure of zero, we have $\delta_N := \max_{x\in\mathbf{X}} |f^N(x) - f(x)| \to 0$ w.p.1 as $N \to \infty$. Note that

$$|f_*^N - f_*| \leq \delta_N,$$

This is because

$$f_*^N - f_* = f^N(x_*^N) - f^N(x_*) + f^N(x_*) - f(x_*) \leq f^N(x_*) - f(x_*) \leq \delta_N \tag{21.1}$$

$$f_* - f_*^N = f(x_*) - f(x_*^N) + f(x_*^N) - f^N(x_*^N) \leq f(x_*^N) - f^N(x_*^N) \leq \delta_N \tag{21.2}$$

Hence, we have $|f_*^N - f_*| \to 0$, i.e. $f_*^N \to f_*$ w.p.1 as $N \to \infty$.

(2) For a given $\epsilon \geq 0$, we define

$$\rho(\epsilon) := \min_{x\in\mathbf{X}\setminus\mathbf{X}^\epsilon} f(x) - f_* - \epsilon.$$

By the definition of ϵ-optimal solution set $\mathbf{X}_\epsilon$ for (P), we have $\forall x \in \mathbf{X} \setminus \mathbf{X}_\epsilon, f(x) > f_* + \epsilon$. Because $\mathbf{X}$ is finite, then $\rho(\epsilon) > 0$.

Let N be large enough such that $\delta_N < \rho(\epsilon)/2$. Hence, $|f_*^N - f_*| \leq \delta_N < \rho(\epsilon)/2$, and furthermore, $f_*^N < f_* + \rho(\epsilon)/2$. In addition, we have $\forall x \in \mathbf{X} \setminus \mathbf{X}_\epsilon, |f^N(x) - f(x)| \leq \delta_N < \rho(\epsilon)/2$, which gives us $f^N(x) > f(x) - \rho(\epsilon)/2 \geq f_* + \epsilon + \rho(\epsilon)/2$. Combining the results above, it follows that $\forall x \in \mathbf{X} \setminus \mathbf{X}_\epsilon, f^N(x) > f_*^N + \epsilon$ and therefore x does not belong to the ϵ-optimal set $\mathbf{X}_\epsilon^N$ for (SAA). In other words, it tells us for $\forall \epsilon \geq 0, \mathbb{P}(\mathbf{X}_\epsilon^N \subseteq \mathbf{X}_\epsilon) = 1$ as $N \to \infty$.

■

This proposition tells us as $N \to \infty$, we can get a near-zero gap approximation for the optimal value f_* of (P) by the SAA formulation. What's more, as $N \to \infty$, for any $\epsilon \geq 0$, the ϵ-optimal solution set of (SAA) always lies in the ϵ-optimal solution set of (P).

Theorem 21.2 [*Kleywegt, Shapiro, Homem-De-Mellow*, 2001]
For $0 \leq \delta < \epsilon$, and integer N, we have

$$\mathbb{P}(\mathbf{X}_\delta^N \subseteq \mathbf{X}_\epsilon) \geq 1 - |\mathbf{X}| e^{-N \cdot \gamma(\delta, \epsilon)},$$

where $\gamma(\delta, \epsilon) \geq \frac{(\epsilon - \delta)^2}{4\sigma_{max}^2}$, and $\sigma_{max}^2 = \max\limits_{x \in \mathbf{X}} Var[F(x, \xi)]$.

Remark 1. When $\delta = 0$, for the optimal solution of the SAA problem to be an ϵ-optimal solution to (P) with probability $1 - \alpha$, the sample size N needs to be

$$N \geq \frac{4\sigma_{\max}^2}{\epsilon^2} \log\big(\frac{|\mathbf{X}|}{\alpha}\big).$$

We can see that even if the size of X increases exponentially, then lower bound of needed sample size N increases linearly. Also, the complexity depends linearly in the variance of $F(x, \xi)$. Overall, the sample complexity is of order $O(1/\epsilon^2)$.

Pros and Cons of SAA.

(+) : SAA is very general. As we see, SAA method doesn't make any assumptions about the convexity of our objective functions. Then 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.

(−) : Since the variance $\sigma_{\max}^2$ and $|X|$ is often unknown or hard to compute, in practice, it is difficult to detterminie an appropriate sample size N we need.

(−) : On the one hand, large sample size N can help us to get a high accuracy, while on the other hand, solving (SAA) problem with large N can be expensive. For instance, computing the gradient $\nabla f^N(x)$ requires to compute the sum of $O(N)$ gradients.

(−) : The SAA approach only works with batch data, and cannot handle streaming/online data.

21.3 Stochastic Approximation

Stochastic Approximation (SA) is another popular method to solve the stochastic optimization problem and it dates back to 1951 by Robbins and Monro [RM51]. 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:

$$x_{t+1} = \Pi_{\mathbf{X}}\big(x_t - \gamma_t \nabla F(x_t, \xi_t)\big),$$

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

Remark 2.

1. The stochastic gradient is unbiased, i.e., $\mathbb{E}\left[\nabla F(x,\xi)\right] = \nabla f(x)$.

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.

Theorem 21.3 [*Nemirovski, Juditsky, Lan, Shapiro, 2009*]
Assume $f(x)$ is μ-strongly convex, and $\exists M > 0$, s.t. $\mathbb{E}\left[||\nabla F(x,\xi)||_2^2\right] \leq M^2, \forall x \in \mathbf{X}$, 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* $C(\gamma) = \max\{\frac{\gamma^2 M^2}{2\mu\gamma - 1}, ||x_1 - x_*||_2^2\}$, *and*

(2) If $f(X)$ is L-smooth and $x_* \in int(\mathbf{X})$, then

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

Proof:

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

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

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

$$\begin{aligned} \mathbb{E}\left[||x_{t+1} - x_*||_2^2\right] \leq \quad & \mathbb{E}\left[||x_t - x_*||_2^2\right] - 2\gamma_t \mathbb{E}\left[\langle \nabla F(x_t, \xi_t), x_t - x_* \rangle\right] + \gamma_t^2 \mathbb{E}\left[|\nabla F(x_t, \xi_t)||_2^2\right] \\ = \quad & \mathbb{E}\left[||x_t - x_*||_2^2\right] - 2\gamma_t \mathbb{E}\left[\langle \nabla F(x_t, \xi_t), x_t - x_* \rangle\right] + \gamma_t^2 M^2. \end{aligned} \tag{21.1}$$

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

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

We also claim that $\mathbb{E}\left[\langle \nabla f(x_t), x_t - x_* \rangle\right] \geq \mu \mathbb{E}\left[||x_t - x_*||_2^2\right]$. By μ-strongly convexity of $f(x)$, we have

$$\begin{aligned} f(x) \text{ is } \mu\text{-strongly convex} \Leftrightarrow \quad & \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu ||x - y||_2^2 \quad \forall x, y \in \mathbf{X} \\ \Leftrightarrow \quad & \langle \nabla f(x), x - y \rangle \geq \mu ||x - y||_2^2 + \langle \nabla f(y), x - y \rangle \quad \forall x, y \in \mathbf{X} \end{aligned} \tag{21.2}$$

Note that by the optimality of x_*, we have $\langle \nabla f(x_*), x - x_* \rangle \geq 0$. Combining the optimality condition and Inequality (21.2), it follows that

$$\langle \nabla f(x_t), x_t - x_* \rangle \geq \mu ||x_t - x_*||_2^2 + \langle \nabla f(x_*), x_t - x_* \rangle \geq \mu ||x_t - x_*||_2^2 \quad \forall x_t \in \mathbf{X}.$$

Taking expectation of the inequality above, we get

$$\mathbb{E}\left[\langle \nabla F(x_t, \xi_t), x_t - x_* \rangle\right] = \mathbb{E}\left[\langle \nabla f(x_t), x_t - x_* \rangle\right] \geq \mu \mathbb{E}\left[||x_t - x_*||_2^2\right]. \tag{21.3}$$

Putting Inequality (21.3) back to Inequality (21.1), we get the following result

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

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

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

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

(2) Give any x_t and $\xi_{[t-1]}$, it can be shown that $f(x_t) - f(x_*) \leq \nabla f(x_*)^\top (x_t - x_*) + \frac{L}{2}||x_t - x_*||_2^2$ since we assume $f(x)$ is L-smooth. Optimality condition gives us that $\nabla f(x_*)^\top (x_t - x_*) \geq 0 \ \forall x_t \in \mathbf{X}$, thus $f(x_t) - f(x_*) \leq \frac{L}{2}||x_t - x_*||_2^2$. Taking expectation on both sides and combining the result in (1), we get

$$\mathbb{E}\left[f(x_t) - f_*\right] = \mathbb{E}\left[f(x_t) - f(x_*)\right] \leq \frac{L}{2}\mathbb{E}\left[||x_t - x_*||_2^2\right] \leq \frac{LC(t)}{2t}.$$

■

Remark 3.

1. Note that from Theorem 21.3, in order to get ϵ-accuracy, we need $\mathcal{O}(\frac{1}{\epsilon})$ number of samples in SA method, while we need $\mathcal{O}(\frac{1}{\epsilon^2})$ number of samples if using SAA method.
2. In the deterministic case (which is shown in Lecture 9), for strongly convex objective function, the error is $||x_t - x_*||_2^2 \leq \mathcal{O}\big((\frac{L-\mu}{L+\mu})^{2t}\big)$ which gives linear convergence rate. However, in the stochastic case, the expected error is $\mathbb{E}\left[||x_t - x_*||_2^2\right] \leq \mathcal{O}(\frac{1}{t})$ which gives a sublinear convergence rate.

We will discuss more on the sample complexity of SA in the next lecture.

References

[KS00] Kleywegt, A. and Shapiro, A. (2000). *Chapter 101 Stochastic Optimization.* http://www2.isye.gatech.edu/~anton/stochoptiebook.

[KSH01] Kleywegt, A, Shapiro, A and Homen-de-mello, T (2001). *The Sample Average Approximation Method for Stochastic Discrete Optimization*, (Vol. 12, No. 2, pp. 479-502). SIAM J. OPTIM.

[RM51] Robbins, H and Monro, S (1951). *A Stochastic Approximation Method*, (pp. 400–407). The annals of mathematical statistics.

[NJLS09] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009). *Robust Stochastic Approximation Approach to Stochastic Programming*, (Vol. 19, No. 4, pp. 1574-1609). SIAM J. OPTIM.