Stochastic Discrete Optimization

Shoudao Wang

November 8, 2021

A Hierarchical Framework of Stochastic Optimization

According to *Peter Glynn, 1986*, stochastic optimization can be viewed in terms of three structure:

- Infinite-Dimensional Stochastic Optimization
  - e.g. determining a time-varying policy.
- Finite-Dimensional Stochastic Optimization
  - Continuous Parameter Stochastic Optimization
    - e.g. optimization over a subset of Euclidean space.
  - Discrete Parameter Stochastic Optimization
    - e.g. optimization over some alternatives.

The Difference Between Discrete and Continuous Stochastic Optimization

Although it seems that the discrete optimization seems easier, since discrete optimization has less candidate. But in fact the opposite is true.

- Continuous Optimization is easier than Discrete Optimization in some sense.
- Discrete problem's solution is tailor-made to the application for most case, while Continuous algorithms are more robust and can be applied to general problem.

Problem Definition

In the fields of manufacturing engineering, operations research, and management science, we often find a discrete optimization problem in which an objective function g is minimized over a nonempty discrete finite feasible set S:

$$\min\{g(s)|s \in S\}, \tag{1}$$

where $g : S \to \mathbb{R}$ and $S = \{s_1, s_2, \cdots, s_\kappa\}$ is a finite feasible set.

Problem Definition

In practice the objective function $g(s)$ is often the expectation of the performance of a system that is subject to stochastic phenomena. We can define it as:

$$g(s) = E[h(s, Y(s))], \tag{2}$$

where E denotes the expectation, h is a function of s and y, and $Y(s)$ is a random vector dependent on s.

In such problems, a closed-form formula is often not available for the objective function $g(s)$, and one is forced to estimate $g(s)$ by Monte Carlo-type simulation.

Difficulties in Discrete Problems

According to *Nelson and Hong, 2015*, There are three fundamental types of errors that occur in discrete optimization problems;

- The optimal solution is never simulated.
- The best solution that was simulated is not selected.
- We do not have a good estimate of the objective function value of the solution we do select.

So how to address these issues is a main subject of the proposed methods.

Optimality Conditions

Let $S^* = \arg\min g(x) : x \in S$ be the solution of problem (1). The finiteness of S implies that there exists a positive constant $\sigma > 0$ such that

$$g^* \leq g(y) - \sigma \qquad \text{for} \quad \text{all} \quad \text{y} \in \text{S}\backslash\text{S}^*, \tag{3}$$

where $g^* = \min_{x \in S}(x)$ is the objective value.

Optimality Conditions

Although the optimal solution S^* is clearly defined, defining optimality conditions is not easy.

- The objective function $g(x)$ cannot be calculated exactly.
- Typically $g(x)$ and $Y(x)$ are unknown functions that are embedded in simulation models.
- Although S is a finite set, it often has a large number of feasible solutions.

Optimality Conditions

Despite these difficulties, researchers have established various optimality conditions for discrete optimization problems that are either theoretically convenient or practically useful.

- When $|S|$ is small, a practical approach is to analyze the probability of correct selection(PCS). i.e.

$$\mathrm{P}\left(\mathbf{x}^* \in \Theta^*\right) \geq 1 - \alpha$$

- When $|S|$ is large, we can relax the goal. Denote T as the top t solutions and $\hat{S}$ is the final n solutions. Our goal is

$$\mathrm{P}(|T \cap \hat{S}| \geq 1) \geq 1 - \alpha$$

- Another optimality condition in global convergence algorithms is $\lim_{m\to\infty} \mathrm{P}\left(\mathbf{x}_m^\star \in S^\star\right) = 1$.

Problem structure and assumptions

Recall the problem we defined in (1) and (2), one can easily come up with a idea that the objective function $g(s)$ can be replaced by its estimate $\hat{g}_\ell(s)$ based on ℓ simulation experiments. But there are two main problems lies here:

- It is not obvious how large the sample size ℓ should be to guarantee the convergence of the optimization technique.
- If the feasible set S is large, then the simulation effort is unacceptable large.

The algorithm proposed in this paper solved this problem by transfer this problem into a maximization problem of a probability. This new problem can be solved by constructing a Markov Chain whose stationary probability distribution converges to the optimal solution.

Problem structure and assumptions

Denote the global optimum set by

$$S^* = \left\{ s \in S \mid g(s) \leqq g\left(s'\right), \forall s' \in S \right\} \tag{4}$$

Recall that $g(s) = \mathbb{E}[H(s, Y(s)]$. Here $H(s)$ is a random variable.

The assumption we need here is that $H(s)$ has a limited variance, i.e.

$$E\left[H(s)^2\right] < \infty, \forall s \in S \tag{5}$$

Translation to a maximization problem

The paper transfer the minimization problem into a maximization by introducing a stochastic ruler.

Let $\Theta(a, b)$ denote the uniformly distributed random variable. Here a and b represent a lower and upper bound for $\{H(s)|s \in S\}$. The probability $P(s, a, b)$ is defined as

$$P(s, a, b) = P[H(s) \leqq \Theta(a, b)] \tag{6}$$

Translation to a maximization problem

We can intuitively see that minimizing $g(s) = \mathbb{E}[H(s)]$ is equivalent to maximizing the probability $P(s, a, b)$ provided the interval (a, b) is sufficiently wide.

Hence we can transfer the original problem(1) into the following maximization problem:

$$\max\{P(s, a, b) \mid s \in S\} \tag{7}$$

The global optimum solution set for this maximization problem is

$$S^*(a, b) = \left\{s \in S \mid P(s, a, b) \geqq P\left(s', a, b\right) \forall s' \in S\right\} \tag{8}$$

The following theorem rigorously delineates the relationship between the original minimization problem and the above maximization problem.

Theorem 1

There exist a real number $\overline{a}$ and $\overline{b}$ such that $\overline{a} < \overline{b}$ and for any $a < \overline{a}$ and any $b < \overline{b}$, the following conclusion hold:

1. *If $g(s) < g(s^{'})$ then $P(s, a, b) > P(s^{'}, a, b)$,*
2. $0 < P(s, a, b) < 1$, *for all $s \in S$*
3. $S^*(a, b) \subset S^*$ *and* $S^*(a, b) \neq \varnothing$.

Translation to a maximization problem

The Theorem (1) mainly states the following points:

- The maximization problem has at least one solution
- Any solution of maximization problem is a solution of the original minimization problem.

Actually the converse also holds.

Theorem 2

Suppose there exist reals $a(s)$ *and* $b(s)$ *such that*

$$a(s) \leq H(s) \leq b(s) \quad w.p.1 \tag{9}$$

If $a < \min\{a(s)|s \in S\}$ *and* $b > \max\{b(s)|s \in S\}$, *then* $S^\star(a, b) = S^\star$.

Definition and assumptions on Computational method

Since we have the maximization problem now, we now have to find a way to solve it. The paper solve by constructing a Markov chain that converges to a global solution to the problem.

Before diving into the algorithm, we need some definition and assumption first.

Definition 1

For each $s \in S$, there exists a subset $N(s)$ of $S - \{s\}$, which is called *the set of neighbors* of s.

.

The search is organized in such a way that the next solution candidate is found among the neighbors of the present candidate.

Definition and assumptions

To ensure that our search will eventually cover all the elements of S, we make the following assumption.

Assumption 1

For any pair $(s, s^{'})$ in $S \times S$, $s^{'}$ is *reachable* from s; i.e., there exists a finite sequence, $\{n_i\}_{i=0}^{\ell}$ for some ℓ, such that

$$s_{n_0} = s, \quad s_{n_\ell} = s', \quad s_{n_{i+1}} \in N(s_{n_i}), \quad i = 0, 1, 2, \cdots, \ell - 1.$$

.

Now we impose a structure to the selection of a candidate .

Definition 2

A function $R : S \times S \to [0, 1]$ is said to be a transition probability for S and N if

1. $R(s, s') > 0 \Leftrightarrow s' \in N(s)$.
2. $\sum_{s' \in S} R(s, s') = 1$.

Definition and assumptions

Now we introduce the following simplification.

Assumption 2

The neighbor system N and the transition probability R for S are *symmetric*, i.e.,

1. $s' \in N(s) \Leftrightarrow s \in N(s')$ and
2. $R(s, s') = R(s', s)$.

In the algorithm, we make use of a sequence of positive integers tending to infinity.

Assumption 3

A sequence $\{M_k\}$ of positive integers satisfies $M_k \to \infty$ as $k \to \infty$.

The Stochastic Algorithm

Aside from N, R, and $\{M_k\}$ defined above, the proposed stochastic algorithm requires parameters, a and b, and an initial guess $s_0 \in S$ for the optimal solution.

THE STOCHASTIC ALGORITHM.
Data: $N, R, \{M_k\}, a, b, s_0 \in S$.
Step 0: Set $X_0 = s_0$ and $k = 0$.
Step 1: Given $X_k = s$, choose a candidate Z_k from $N(s)$ with probability distribution

$$P[Z_k = s' / X_k = s] = R(s, s'), s' \in N(s).$$

Step 2: Given $Z_k = s'$, set

$$X_{k+1} = \begin{cases} Z_k, & \text{with probability } p_k, \\ X_k, & \text{with probability } (1 - p_k), \end{cases}$$

where

$$p_k = \{P[H(s') \leqq \Theta(a, b)]\}^{M_k} = \{P(s', a, b)\}^{M_k}.$$

Remark. Since we are interested in cases in which the probability $P(s', a, b)$ given above in Step 2 is not explicitly computable, we suggest a subalgorithm for implementing Step 2 immediately following the algorithm.

Step 3: Set $k = k + 1$ and go to Step 1.

The Stochastic Algorithm

The implementation of Step 2 of the above algorithm may be accomplished by the following subalgorithm where $P(s', a, b)$ need not be computed.

1. Set $c = 1$;
2. Draw a sample $h(s')$ from $H(s')$. Next draw a sample θ from $\Theta(a, b)$.
   - If $h(s') > \theta$, then set $X_{k+1} = X_k$, break.
   - Else if $c > M_k$, set $X_{k+1} = Z_k = s'$, break.
   - Else set $c = c + 1$ and continue Step 2 from beginning.

The Stochastic Algorithm

The random process $\{X_k\}$ produced by the Stochastic Algorithm is a discrete-time Markov chain defined over states S, and its state transition probabilities are given by

$$P_{ss'}(M_k) = P\left[X_{k+1} = s'/X_k = s\right]$$
$$= \begin{cases} R(s,s')\left\{P(s',a,b)\right\}^{M_k}, & \text{if } s' \in N(s) \\ 1 - \sum_{s'' \in N(s)} R(s,s'')\left\{P(s'',a,b)\right\}^{M_k}, & \text{if } s' = s \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

We make use of the state transition probability matrix, which is a matrix consisting of the above probabilities:

$$P(M_k) = (P_{ss'}(M_k)) \quad (11)$$

Analysis for the stationary process

We now suspend the Assumption 3 and set M_k to a positive integer M. For each $s \in S$, define

$$\pi_s(M) = \frac{\{P[H(s) \leqq \Theta(a,b)]\}^M}{\sum_{s' \in S}\{P[H(s') \leqq \Theta(a,b)]\}^M} = \frac{\{P(s,a,b)\}^M}{\sum_{s' \in S}\{P(s',a,b)\}^M}$$

Theorem 3

The vector $\pi(M)$ consisting of $\pi_s(M)$ is the stationary probability distribution for the Markov chain $\{X_k\}$ generated by the stochastic algorithm, i.e.,

$$\pi(M)P(M) = \pi(M)$$

The limiting behavior of the stationary distribution

We now investigate the behavior of the stationary probability distribution $\{\pi_s(M)|s \in S\}$ as M goes to infinity.

Definition 3

Given a finite set S, the set $\Pi(S)$ of positive unit vectors is called the *set of probability vectors* for S, below:

$$\Pi(S) = \left\{ \pi \in [0,1]^{\kappa} \mid \pi_s \geqq 0, \|\pi\| = \sum_{s \in S} \pi_s = 1 \right\},$$

where $\kappa = |S|$ represents the cardinality of S.

Definition 4

A probability vector $\pi^\star$ for S is called *optimal* if $\pi^\star = 0$ for any $s \notin S^\star$.

The limiting behavior of the stationary distribution

Theorem 4

The probability vector $\pi(M)$ converges, as M goes to infinity, to an optimal probability vector $\pi^\star$. Furthermore

$$\pi_s^* = \begin{cases} 1/\left|S^*(a,b)\right|, & \text{if } s \in S^*(a,b) \\ 0, & \text{otherwise} \end{cases}$$

where $|S^*(a,b)|$ represent the cardinality of $S^\star(a,b)$

Proposition 1

1. For each $s \in S^\star(a,b)$, if $M < M'$ then $\pi(M) \leq \pi_s(M')$.
2. For each $s \notin S^\star$ there exists an integer M_s such that if $M_s \leq M < M'$ then $\pi_s(M) \geq \pi_s(M')$.

Rate of convergence

Theorem 5

Suppose that reals c and r, integer k_0, and a sequence $\{M_k\}$ are selected as in Theorem 7.1 in the paper. Then for a sufficiently large integer m,

$$\|x(mr) - \pi^*\| \leqq O\left(1/m^t\right)$$

where $t = \min\{\hat{t}, \bar{t}\} = \min\{(\rho/r^c/2), \eta c/2\} > 0.$

Advantages and disadvantages of the algorithm

Advantages:

- This algorithm is globally convergent in theory.
- When there are large number of alternatives, this algorithm can be used while R&S can not.
- Since in each iteration it retains no past data, this algorithm is memory free.

Disadvantage:

- It's hard to determine when to stop for this algorithm.
- The computation effort goes up as iteration goes up.
- It is not a adaptive method. Lack of past information result in a poor performance in practice.

Thank you!