Fu 2015
Handbook of Simulation Optimization

Daniele Zago

June 27, 2021

Contents

1 Summary

Simulation optimization refers to optimization in the setting where f cannot be computed exactly, but is estimated with some noise. We therefore use the objective function

$$f(\boldsymbol{x}) = \mathbb{E}\big[Y(\boldsymbol{x})\big],$$

or some other performance measure based on some stochastic observation Y.

2 Discrete Optimization via Simulation

We want to optimize a function

$$\operatorname*{argmin}_{\boldsymbol{x}\in\Theta} g(\boldsymbol{x}) = \mathbb{E}\big[Y(\boldsymbol{x})\big],$$

where $\boldsymbol{x}$ takes values in $\{1, 2, \ldots, k\}$ or is an integer-ordered variable in $\mathbb{R}^d$. There are three types of error in these problems:

1. We *never simulate from the optimal solution.*
2. Sampling variability means that *the best solution might not have the best objective function value.*
3. There is a natural bias towards solutions whose estimated performance is *lower than the true expected value.*

Optimality conditions are defined in terms of probability of converging to the optimum, whether $g_n^* \xrightarrow{P} g^*$ or $g_n \xrightarrow{\text{a.s.}} g^*$. This can be valid either for a global optimal $g^* \in \Theta^*$ or a local optimum $g^* \in \mathcal{L}$. This means that the algorithm is going to stop in a finite amount of time.

Once we found a local neighborhood $N(\boldsymbol{x})$, we can perform a statistical test

$$\begin{cases} H_0 : g(\boldsymbol{x}) \leq \min\limits_{\boldsymbol{y}\in N(\boldsymbol{x})} g(\boldsymbol{y}) \\ H_1 : g(\boldsymbol{x}) > \min\limits_{\boldsymbol{y}\in N(\boldsymbol{x})} g(\boldsymbol{y}) \end{cases}$$

Bechhofer's procedure

One of the simplest procedures, which is not sequential. Assuming $Y_j(\boldsymbol{x}_i) \sim \mathcal{N}\big(g(\boldsymbol{x}_i), \sigma_i^2\big)$ such that $g(\boldsymbol{x}_i) + \delta = g(\boldsymbol{x}_2) = \ldots = g(\boldsymbol{x}_k)$, then we can set $n = \lceil 2h^2\sigma^2/\delta^2 \rceil$ and select $Y(\boldsymbol{x}_i)$ such that the sample mean is the smallest.

Paulson's procedure

Sequential procedure based on the same assumptions, we set the initial solution set to be $I = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ and $r = 0$. Define $a = \log\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{\delta - \lambda}$, and iterate

1. $r = r + 1$, for each $\boldsymbol{x}_i \in I$ compute $\overline{Y}(x_i; r)$ with one observation.
2. $I_{\text{old}} = I$ and
$$I = \left\{ \boldsymbol{x}_i \in I_{\text{old}} : \overline{Y}(\boldsymbol{x}_i; r) \leq \min_{\ell \in I_{\text{old}}} \overline{Y}(\boldsymbol{x}_\ell; r) + \left(\frac{a}{r} - \lambda\right)_+ \right\}.$$
Iterate 1. and 2. until $|I| = 1$.

This method is based on a large deviation bounds to account for multiple comparisons of the data. Other methods are Gupta's procedure, NSGS procedure, KN procedure, ...

3 Response Surface Methodology

RSM is based on a stepwise procedure that alternates these steps (Box and Wilson, 1951)

1. Fit a linear model with a full factorial design to get an approximate local gradient.
2. Perform a steepest ascent in the direction of the gradient, until performance degrades.

Once we get close to the optimum, we use a second-order polynomial to estimate the optimal solution. The factors levels are fixed in a star-shaped pattern with the addition of a center value, in order to estimate the curvature of the parabola.

4 Stochastic Approximations

A Stochastic Approximation (SA) algorithm is an update scheme for finding the zero of a stochastic function inside a set Θ,

$$\min_{x\in\Theta} f(x) = \mathbb{E}\big[Y|x\big], \quad Y \sim P$$

which has a general formulation of the following type:

$$x_{n+1} = \pi_\Theta\left(x_n - a_n\widehat{\nabla f}(x_n)\right), \tag{1}$$

where $a_n \xrightarrow{n\to\infty} 0$ is a sequence of step sizes, and π_Θ is the projection onto Θ in case the update goes outside the set.

$\widehat{\nabla f}(x)$ is an estimator of the gradient ∇f, and depending on its characteristics, we have different types of resulting algorithms

1. Robbins and Monro (1951) use direct unbiased estimate of $\widehat{\nabla f}$, with convergence $\mathcal{O}(n^{-1/2})$.
2. Kiefer and Wolfowitz (1952) use the asymptotically-unbiased finite difference estimate of $\widehat{\nabla f}$, with convergence $\mathcal{O}(n^{-1/3})$.

4.1 Indirect gradient estimators

Different methods can be used for various problems, depending on the specific needs. The simplest method is the *symmetric difference estimator*

$$\widehat{\nabla f}_i(x) = \frac{Y(\boldsymbol{x} + c_i\boldsymbol{e}_i) - Y(\boldsymbol{x} - c_i\boldsymbol{e}_i)}{2c_i}.$$

A more sophisticated method for high-dimensional estimation is the *Simultaneous Perturbation* (SP) of Spall (1992), which estimates the gradient by

$$\widehat{\nabla f}_i(x) = \frac{Y(\boldsymbol{x} + c_i\Delta) - Y(\boldsymbol{x} - c_i\Delta)}{2c_i\Delta_i}, \tag{2}$$

where $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_d)$ is such that $\mathbb{E}\big[\Delta\big] = \mathbf{0}$, and has finite inverse second moments.

Note The finite inverse second moments prohibits the use of the Gaussian distribution for Δ, instead the most common perturbation is $\Delta_i \overset{i.i.d}{\sim}$ Rademacher(0.5).

If we use instead the *random direction* gradient estimator,

$$\widehat{\nabla f}_i(x) = \frac{Y(\boldsymbol{x} + c_i\Delta) - Y(\boldsymbol{x} - c_i\Delta)}{2c_i}\Delta_i,$$

then the finite inverse second moments requirement becomes instead $\mathbb{E}\big[\Delta\Delta^\top\big] = I_d$, and a Gaussian distribution can be employed.

4.2 Robbins-Monro algorithm

Thm. (Convergence of RM algorithm)
Assume $\nabla f(x)$ has a unique root x^* and that $\widehat{\nabla f}(x)$ is unbiased, $\mathbb{E}\big[\widehat{\nabla f}(x)\big] = \nabla f(x)$. Then, by applying the iteration (1) with the following conditions

1. *$\{a_n\}$ is such that $\sum_{n=1}^{\infty} a_n = \infty$ and $\sum_{n=1}^{\infty} a_n^2 < \infty$;*
2. *$\nabla f(x) \geq 0$ for $x > x^*$ and $\nabla f(x) \leq 0$ for $x < x^*$;*
3. *There exists $C > 0$ such that $\mathbb{P}\big(\big|\widehat{\nabla f}(x)\big| \leq C\big) = 1$ for all x;*

we have that $x_n \xrightarrow{L^2} x^*$ as $n \to \infty$.

Remarks

- Usually, we choose $a_n = \frac{\vartheta_a}{(n+A)^\alpha}$ for $\alpha \in (1/2, 1)$, $\vartheta_a > 0$, $A \geq 0$.
- There are variants that improve bounded search inside Θ, such as considering $\Theta = \cup_m \Theta_m$.
- *Simultaneous Perturbation Stochastic Approximation* (SPSA): the SP method (2) is used to estimate $\widehat{\nabla f}(x)$ inside the Robbins-Monro iteration (Spall, 1992). We can average multiple pairs of evaluations in Equation (2) to reduce noise. Also, we can approximate the Hessian $\nabla^2 f(x)$ to accelerate convergence.

References

Box, G. E. P. and Wilson, K. B. (1951). "On the Experimental Attainment of Optimum Conditions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 13.1, pp. 1–45.

Kiefer, J. and Wolfowitz, J. (1952). "Stochastic Estimation of the Maximum of a Regression Function". In: *The Annals of Mathematical Statistics* 23.3, pp. 462–466.

Robbins, H. and Monro, S. (1951). "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407.

Spall, J. (1992). "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation". In: *IEEE Transactions on Automatic Control* 37.3, pp. 332–341.