OPTIMAL CONVERGENCE RATE FOR RANDOM SEARCH

Yen Lin Chia

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305-4026, USA

Peter W. Glynn

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305-4026, USA

ABSTRACT

Many different applications require solving optimization problems of the form

$$\max_{\theta \in \Lambda} \alpha(\theta), \tag{1}$$

where $\Lambda \subseteq I\!R$ is the feasible set, and $\alpha(\cdot)$ is the objective function. When $\alpha(\cdot)$ is available in "closed form" and is smooth, (1) is generally solved numerically by applying derivative-based iterative algorithms; see, for example, Gill, Murray and Wright (1981).

We consider here the case where $\alpha(\theta)$ is defined as the expectation of a real-valued random variable, $X(\theta)$. We assume that $EX(\theta)$ is not available in closed form, and must be computed via (Monte Carlo) simulation. Perhaps the simplest possible algorithm for solving (1) is what we shall call "simple random search". Simple random search proceeds by first generating points $\theta_1, \theta_2, \ldots, \theta_m$ randomly from Λ. One then simulates n independent realizations of $X(\theta)$ at each $\theta \in \{\theta_1, \ldots, \theta_m\}$ and computes the sample mean, $\bar{X}_n(\theta_i) \stackrel{\triangle}{=} n^{-1} \sum_{j=1}^{n} X_j(\theta_i)$ $(1 \leq i \leq m)$, at each of the m random points. The maximum of the problem (1) is then estimated via $\{\bar{X}_n(\theta_i) : 1 \leq i \leq m\}$, and the maximizer is estimated via the empirical maximizer of $\{\bar{X}_n(\theta_i) : 1 \leq i \leq m\}$.

This algorithm does not take advantage of potential smoothness in $\alpha(\cdot)$, nor does it adapt its behavior in light of the information gained from previously generated observations. On the other hand, simple random search is trivial to implement, in part because it makes no effort to estimate derivatives (which are generally difficult to compute in the simulation setting; see for example, L'Ecuyer and Perron (1994). In addition, simple random search is convergent even in the presence of (many) local maxima. Such multi-modality significantly complicates the implementation of more intelligent derivative-based methods. For the above reasons, simple random search is frequently used by simulation practitioners to solve (1).

Despite the importance of simple random search, no analysis of the best possible convergence rate is presently available. Our main contribution here is to supply a relatively complete convergence theory for simple random search. Deriving the best possible rates of convergence (that optimally balance the relative magnitudes of m and n) also gives us a "benchmark" convergence rate for the simplest possible simulation-based optimization algorithm, against which all other algorithms can be compared.

To state our main result, we start by assuming that $\Lambda \subseteq I\!R^d$ is a compact set with a non-empty interior. Given a computer (time) budget c, we generate m i.i.d. random points $\theta_1, \theta_2, \ldots, \theta_m$ from a common density g that is concentrated on Λ. The density g is assumed to be positive and continuous on Λ. Given the random points $\{\theta_1, \theta_2, \ldots, \theta_m\}$, we perform n simulations at each of the m points. In view of the computer time constraint, we set $n = \lfloor c/m \rfloor$, so that the total number of simulations (aggregated across all m points) is approximately equal to c. More precisely, we simulate mn random variables

$\{X_j(\theta_i) : 1 \le i \le m, 1 \le j \le n\}$ having conditional distribution

$$P(X_j(\theta_i) \in dx_{ij},\ 1 \le i \le m,\ 1 \le j \le n \mid \theta_1, \theta_2, \ldots) = \prod_{i=1}^{m}\prod_{j=1}^{n} F(\theta_i, dx_{ij}),$$

where $F(\theta, \cdot)$ is the distribution function of $X(\theta)$. For $i \ge 1$, put

$$\alpha_n(\theta_i) = \frac{1}{n}\sum_{j=1}^{n} X_j(\theta_i),$$

so that $\alpha_n(\theta_i)$ is the sample mean of the $X_j(\theta_i)$'s associated with the ith random point θ_i. Then,

$$\hat{\alpha}(c) = \max_{1 \le i \le m} \alpha_n(\theta_i)$$

is the associated estimator of $\max\{\alpha(\theta) : \theta \in \Lambda\}$.

Our first result covers the case where $X(\theta) = \alpha(\theta)$ a.s., so that the function evaluations are deterministic. If we throw m points uniformly into Λ, then there is (in expectation) one point in each subset of Λ having volume $\mathrm{vol}(\Lambda)/m$. The radius of a d-sphere having volume of order $1/m$ is of order $m^{-1/d}$. This suggests that the closest sampled point θ_i to the maximizer θ^* of $\alpha(\cdot)$ is at a distance of order $m^{-1/d}$ from θ^*. If $\alpha(\cdot)$ is smooth, then $\alpha(\cdot)$ is locally quadratic around θ^*. Hence, the difference between $\alpha(\theta_i)$ and $\alpha(\theta^*)$ should be of order $m^{-2/d}$. This analysis suggests that the rate of convergence of $\hat{\alpha}(c)$ to $\max\{\alpha(\theta) : \theta \in \Lambda\}$ is of order $c^{-2/d}$ in the deterministic function evaluation setting. (Note that in this setting, $n = 1$ so that $m = \lfloor c \rfloor$.)

We now proceed to make this analysis precise.

Assumption 1 *$\alpha(\cdot)$ is twice continuously differentiable on Λ.*

Assumption 2 *$\alpha(\cdot)$ has a unique maximizer θ^* lying in the interior of Λ.*

Assumption 3 *The Hessian of $\alpha(\cdot)$, when evaluated at θ^* (and denoted $H(\theta^*)$), is negative definite.*

Theorem 1 *Assume 1, 2, and 3. If $X(\theta) = \alpha(\theta)$ a.s. for $\theta \in \Lambda$, then*

$$c^{2/d}(\alpha(\theta^*) - \hat{\alpha}(c)) \Rightarrow \mathit{Weibull}(a, d/2)$$

as $c \to \infty$, where $\mathit{Weibull}(a, d/2)$ is a Weibull random variable with shape parameter $d/2$ and scale parameter a given by

$$a = 2\pi\left(\frac{g(\theta^*)}{\Gamma(d/2+1)\sqrt{|\det\, H(\theta^*)|}}\right)^{2/d}.$$

The Weibull structure of the limit was previously identified by Archetti, Betrò, and Steffè (1977) and de Haan (1981). The new feature of the above result is its explicit computation of the scale parameter of the Weibull limit law.

Theorem 1 shows that when function evaluations are deterministic, then the rate of convergence to the maximum is of order $c^{-2/d}$. Our goal is next to identify the optimal rate of convergence in the setting of stochastic function evaluations. It seems intuitively clear that the optimum trade-off between m and n is attained when the error contributed by the finite number m of random points and the Monte Carlo error associated with the sample size n are roughly balanced. In view of our previous discussion of the rates attained in deterministic setting and the $n^{-1/2}$ associated with Monte Carlo estimators (due to the central limit theorem), this suggest that an optimal trade-off is attained when $m^{-2/d} \approx n^{-1/2}$. When expressed in terms of c, this leads to consideration of limit distributions for $\hat{\alpha}(c)$ in which the asymptotic regime is given by

$$\begin{aligned} m &\sim r c^{d/(d+4)} \\ n &\sim r^{-1} c^{4/(d+4)} \end{aligned} \tag{2}$$

as $c \to \infty$ (with $0 < r < \infty$).

To analyze this asymptotic regime, we make some additional assumptions:

Assumption 4 *The collection of distributions $\{F(\theta, \cdot) : \theta \in \Lambda\}$ is weakly continuous over Λ (i.e. if $\theta'_n \in \Lambda$ is such that $\theta'_n \to \theta'_\infty$, then $F(\theta'_n, \cdot) \Rightarrow F(\theta'_\infty, \cdot)$ as $n \to \infty$).*

Assumption 5 *var $X(\theta^*) > 0$.*

Set $\sigma(\theta) = \sqrt{\mathrm{var}X(\theta)}$. Our main result describes the behavior of $\hat{\alpha}(c)$ where m and n are balanced according to (2).

Theorem 2 *Assume 1 through 5. Suppose that $\sup\{E|X(\theta)|^p : \theta \in \Lambda\} < \infty$ for $p > \max(3, d/2)$. If m and n satisfy (2), then*

$$c^{2/(d+4)}(\hat{\alpha}(c) - \alpha(\theta^*)) \Rightarrow \beta$$

as $c \to \infty$, where

$$P(\beta \le x) = \exp\left(-\frac{2r^{(4+d)/4} g(\theta^*)\pi^{d/2}}{\Gamma(d/2)\sqrt{|\det H(\theta^*)|}} \mathit{Int}\right)$$

and

$$\mathit{Int} = \int_0^\infty P(N(0,1) > \frac{2x+y}{2\sigma(\theta^*)} y^{d/2-1}\, dy.$$

According to Theorem 2, the rate of convergence in the asymptotic regime (2) is $c^{-2/(d+4)}$ as $c \to \infty$. This makes clear that simple random search converges slowly when the number of decision variables d is large. On the other hand, when d is large, the rate is only marginally worse than the rate $c^{-2/d}$ obtained in the setting of deterministic function evaluations (see Theorem 1). This suggests that when d is large, the stochastic nature of the function evaluations only modestly degrades the performance of simple random search. The principal factor contributing to the slow convergence rate for d large is the fact that there is no learning effect that is present in the way points $\theta_1, \theta_2, \ldots$ are generated.

The above limit theorem suggests a potential procedure, to be developed elsewhere, for generating confidence intervals for the optimal value that leverages off the new limit distribution identified by Theorem 2.

REFERENCES

Archetti, F., B. Betrò, and S. Steffè. 1977. A theoretical framework for global optimization via random sampling. Technical Report, Cuaderni del Dipartimento di Ricerca Operative e Scienze Statische, Università di Pisa.

de Haan, L. 1981. Estimation of the minimum of a function using order statistics. *Journal of the American Statistical Association 76*(374), 467–469.

Gill, P. E., W. Murray, and M. H. Wright. 1981. *Practical Optimization*. Academic Press.

L'Ecuyer, P. and G. Perron. 1994. On the convergence rates of IPA and FDC derivative estimators. *Operations Research 42*(4), 643–656.

AUTHOR BIOGRAPHIES

YEN LIN CHIA earned her PhD in Management Science and Engineering at Stanford University and she is now working as a Senior Scientist at Novartis Pharmaceuticals in Cambridge, MA. Her e-mail address is <yenlin.chia@novartis.com>.

PETER W. GLYNN is a professor of operations research at Stanford University. His e-mail address is <glynn@stanford.edu>.