Limit Theorems for Random Search

Peter W. Glynn

Stanford University

Conference in Honour of Soren Asmussen, Sandbjerg Estate, Denmark
August 1-5, 2011

This talk has a number of elements that are aligned with Søren's work:

simulation

sums of random variables

extreme values

large deviations

asymptotics

The Setting

$$\min_{\theta \in \mathbb{R}^d} \alpha(\theta)$$

where

$$\alpha(\theta) = \mathbb{E}X(\theta)$$

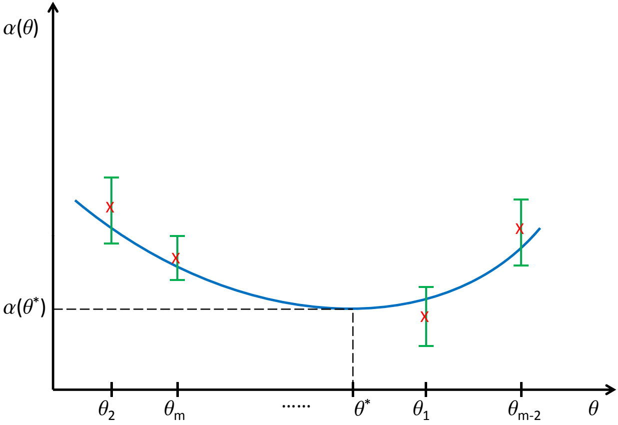
must be computed via (Monte Carlo) simulation

Assume that $\alpha(\cdot)$ is smooth

The Class of Algorithms to be Studied

- Randomly sample m points $\theta_1, \ldots, \theta_m$ from $\mathbb{R}^d$
- Perform simulations at each of the m points
- Estimate the minimum value of $\alpha(\cdot)$ from the observations

Not much intelligent adaptation built into these algorithms (i.e. simple random search)


$\alpha(\theta)$
$\alpha(\theta^*)$
θ_2
θ_m
......
θ^*
θ_1
θ_{m-2}
θ

Why Study?

- Because of their simplicity, they are easy to implement (and are used in practice)

- They can be viewed as “benchmark algorithms” (Any “good algorithm” should beat the rates of convergence associated with these random search algorithms.)

- They are tractable mathematically, and provide insights into more complex algorithms

Outline of Talk

- Simple Random Search
  - Consistency
  - Optimal Convergence Rate
  - Large Deviations

- Simple Random Search with Gradient Information

- Simple Random Search with Point-dependent Sample Size

A More Detailed Description of Simple Random Search

- Randomly and independently sample m points $\theta_1, \ldots, \theta_m$ from $\mathbb{R}^d$ from a continuous positive density g
- At each point θ_i, randomly generate n iid copies of $X(\theta_i)$ (independent of the simulations at the other θ-values), thereby computing $\overline{X}_n(\theta_i)$
- Given a computer (time) budget c, let $n = \lfloor c/m \rfloor$
- Use the minimum of $\overline{X}_n(\theta_i)$ as an estimator of the minimum $\alpha(\theta^*)$, where θ^* is the minimizer of $\alpha(\cdot)$
- Note that our estimator of the minimum is:

$$\hat{\alpha}(c) = \min_{1 \leq i \leq m} \overline{X}_n(\theta_i)$$

An extreme value statistic (but with a distribution depending on n)

Consistency

If the number of points m is too large relative to the sample size n, the extreme value may not be consistent as an estimator for $\alpha(\theta^*)$

Light-tailed Case: $\sup_\theta \mathbb{E} \exp(\gamma |X(\theta)|) < \infty$ for some $\gamma > 0$

Theorem

1. If $\log m/n \to 0$ as $c \to \infty$, then
$$\hat{\alpha}(c) \Rightarrow \min_\theta \alpha(\theta) \quad \text{as } c \to \infty.$$
2. If $\log m/n \to \infty$ as $c \to \infty$, then
$$\hat{\alpha}(c) \Rightarrow s \quad \text{as } c \to \infty,$$
where $s = \min_\theta s(\theta)$, and $s(\theta)$ is the left end-point of support of $X(\theta)$

Theorem (continue)

3. Suppose $\log m/n \to \tau \in (0, \infty)$ as $c \to \infty$. Assume that for each $\theta \in \mathbb{R}^d$, there exists a root $\tilde{\gamma} = \tilde{\gamma}(\theta) > 0$ satisfying

$$\tilde{\gamma} \frac{\partial}{\partial \gamma} \psi(\theta, \tilde{\gamma}) - \psi(\theta, \tilde{\gamma}) = \tau,$$

where $\psi(\theta, \gamma) \triangleq \log \mathbb{E} \exp(\gamma X(\theta))$. Furthermore, suppose that ψ is twice differentiable on $\mathbb{R}^d \times [0, \gamma_0]$, where $\gamma_0 > \sup_\theta \tilde{\gamma}(\theta)$. Then,

$$\hat{\alpha}(c) \Rightarrow \min_\theta \frac{\partial}{\partial \gamma} \psi(\theta, \tilde{\gamma}) \quad \text{as } c \to \infty.$$

Heavy-tailed Case: With stable noise $(1 < \nu < 2)$, $m/n^{\nu-1}$ must converge to zero in order that our method consistently estimate $\alpha(\theta^*)$

Assumptions

1. α has a unique minimizer θ^*
2. The Hessian of α, when evaluated at θ^* (denoted $H(\theta^*)$), is positive definite

Theorem (Archetti et 1977, de Haan 1978, Chia and G 2011)

Assume 1 and 2. If $X(\theta) = \alpha(\theta)$ a.s. for all θ, then

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

as $c \to \infty$, where Weibull$(a, d/2)$ is a Weibull rv 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 Optimal Convergence Rate in the Noisy Setting

Heuristic Argument:

- Noise in the function evaluations: $n^{-1/2}$
- Closest point to θ^*: $m^{-1/d}$
- Function value relative to $\alpha(\theta^*)$ at closest point: $m^{-2/d}$
- For optimal rate, balance two errors: $n^{-1/2} \approx m^{-2/d}$
- With $mn = c$:

$$m \sim rc^{d/(d+4)}$$
$$n \sim r^{-1}c^{4/(d+4)}$$

for $r \in (0, 1)$

Assumptions

3. The collection of distributions $\{F(\theta, \cdot) : \theta \in \mathbb{R}^d\}$ is weakly continuous over $\mathbb{R}^d$
4. $\mathrm{var}(X(\theta^*)) > 0$

Theorem (Chia and G 2011)

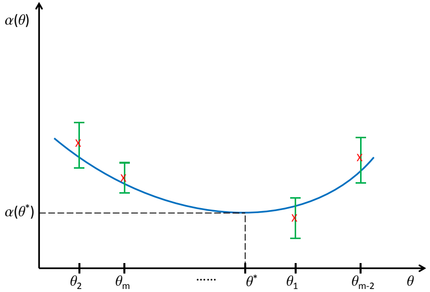
Assume 1 through 4. Suppose $\sup_\theta \mathbb{E}|X(\theta)|^p < \infty$ for $p > \max(3, d^3/2)$. Then,

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

as $c \to \infty$, where, letting $\sigma(\theta^*) = \sqrt{\mathrm{var}(X(\theta^*))}$,

$$\mathbb{P}(\beta \leq x) = \exp\Bigg(- \frac{2r^{(d+4)/4} g(\theta^*)\pi^{d/2}}{\Gamma(d/2)\sqrt{|\det H(\theta^*)|}} \times \int_0^\infty \mathbb{P}(\mathcal{N}(0,1) > \frac{2x+y}{2\sigma(\theta^*)}) y^{d/2-1} \mathrm{d}y \Bigg)$$

Large Deviations Analysis



- Large deviations below can be caused by unusually large deviations at any one of $\theta_1, \ldots, \theta_m$
- Large deviations above requires unusual behavior at all m of $\theta_1, \ldots, \theta_m$ ("cheapest way" is that we were unlucky in the placement of the m sample points)

The Lower Large Deviations Result

Theorem (Subramanian and G 2011)

Let $\psi(\theta; t) = \log \mathbb{E} e^{tX(\theta)}$ and $\mathcal{I}(\theta; x)$ be the large deviations rate function for $n^{-1} \sum_{i=1}^{n} X_i(\theta)$. Then,

$$\begin{aligned}
&\mathbb{P}(\hat{\alpha}(c) < \alpha(\theta^*) - x) \\
=&\frac{md}{2} \left(\frac{\pi \psi''(\theta^*; \theta(x))}{xn} \right)^{d/2} \\
&\times \frac{\exp(-n\mathcal{I}(\theta^*; x))}{\sqrt{2\pi \psi''(\theta^*; \theta(x)) |\det H(\theta^*)|}} g(\theta^*)(1 + o(1)),
\end{aligned}$$

as $c \to \infty$.

The Upper Large Deviations Result

Theorem (Subramanian and G 2011)

$$\mathbb{P}(\hat{\alpha}(c) > \alpha(\theta^*) + x) = (1-p)^m \exp\left(-m \sum_{j=1}^{k} \frac{b_j}{n^j} + o(1)\right),$$

as $c \to \infty$, where k is the smallest integer such that $mn^{-k+1} \to 0$ as $c \to \infty$.

Note that

$$
\begin{aligned}
&\mathbb{P}(\hat{\alpha}(c) \geq \alpha(\theta^*) + x) \\
=&\mathbb{P}(\overline{X}_n(\theta_i) \geq \alpha(\theta^*) + x)^m \\
=&\mathbb{P}(\alpha(\theta_i) \geq \alpha(\theta^*) + x)^m \exp\left(m \log\left(1 - \frac{p_n - p}{1 - p}\right)\right)
\end{aligned}
$$

where

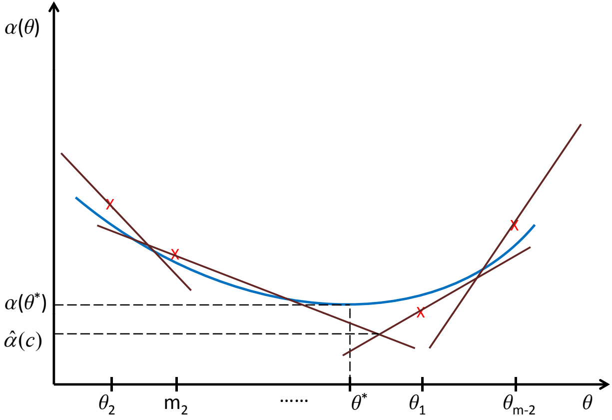
$$
p_n = \mathbb{P}(\overline{X}_n(\theta_i) \leq \alpha(\theta^*) + x) = p + \sum_{j=1}^{k} \frac{a_j}{n^j} + o(n^{-k})
$$

(Lee and G 99)

Simple Random Search with Gradient Information

- At each point $\theta_1, \ldots, \theta_m$, estimate both $\alpha(\theta_i)$ and $\nabla\alpha(\theta_i)$
- Let $\nabla\overline{X}_n(\theta_i)$ be our estimator for $\nabla\alpha(\theta_i)$ based on:
  - likelihood ratio gradient estimation (often unbiased)
  - infinitesimal perturbation analysis (often unbiased)
  - (noisy) finite difference approximation based on central differences (always biased)
- Assume $\alpha(\cdot)$ is strictly convex

The Estimator



Theorem (Wu and G 2011)

Suppose that $n \sim \beta c^{2p}$ for $\beta > 0$ as $c \to \infty$, where $p = 2/(d+4)$. Then,

$$c^p(\hat{\alpha}(c) - \alpha(\theta^*)) \Rightarrow W$$

as $c \to \infty$.

- W can be described in terms of a limiting Poisson random field with randomly generated hyperplanes/function values at each Poisson point
- Heart of the argument: Showing that the estimator ultimately depends on "local behavior" of Poisson random field
- Generalizes to setting of biased gradient estimators

Simple Random Search with Point-dependent Sample Size

What happens if you do not use common sample size n across all the θ_i's?

More intelligent approach:

- Begin sampling simultaneously at each θ_i value
- Continue sampling until it is clear the θ_i value is clearly not optimal
- Focus sampling on the "best"
- Note that it is pointless to let $n^{-1/2} \ll m^{2/d}$, even for the most promising points

Conclusions:

- One gets a convergence rate arbitrarily close to $c^{-2/d}$
- Optimal rate is close to that in noiseless setting

Extensions

- What about if one applies common random numbers for the simulations at each of the points $\theta_1, \ldots, \theta_m$?
- What happens in the presence of constraints?
- What about similarly descriptive limit theorems for more intelligent search?