Limit Theorems for Simulation-Based Optimization via Random Search

YEN LIN CHIA, Nektar Therapeutics
PETER W. GLYNN, Stanford University

This article develops fundamental theory related to the use of simulation-based nonadaptive random search as a means of optimizing a function that can be expressed as an expectation. Our results establish rates of convergence that express the trade-off between exploration and estimation, and fully characterize the limit distributions that arise. Our rates of convergence results should be viewed as a baseline against which to compare more intelligent algorithms.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Optimization, simulation, random search, limit theorems

ACM Reference Format:
Chia, Y. L. and Glynn, P. W. 2013. Limit theorems for simulation-based optimization via random search. *ACM Trans. Model. Comput. Simul.* 23, 3, Article 16 (July 2013), 18 pages.
DOI:http://dx.doi.org/10.1145/2499913.2499915

1. INTRODUCTION

Many different applications require solving optimization problems of the form

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

where $\Lambda \subseteq I\!R^d$ 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 et al. [1981].

In this article, we focus on 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, $\alpha_n(\theta_i) \overset{\triangle}{=} n^{-1} \sum_{j=1}^{n} X_j(\theta_i)$ $(1 \le i \le m)$, at each of the m random points. The maximum of the problem (1) is then estimated via $\max\{\alpha_n(\theta_i) : 1 \le i \le m\}$, and the maximizer is estimated via the empirical maximizer of $\{\alpha_n(\theta_i) : 1 \le i \le m\}$.

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

Author's address: Y. L. Chia; email: yenlin@gmail.com.

DOI:http://dx.doi.org/10.1145/2499913.2499915

compute in the simulation setting; see, for example, Robbins and Monro [1951], Kiefer and Wolfowitz [1952] and L'Ecuyer and Perron [1994]). In addition, simple random search is convergent even in the presence of (many) local maxima. Such multimodality significantly complicates the implementation of more intelligent derivative-based methods. For these reasons, simple random search is a potentially attractive algorithm from a practitioner viewpoint.

Despite the attractiveness of simple random search, no analysis of the best possible convergence rate is presently available. Our main contribution in this article 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.

The major contributions of this article are theoretical in nature and include the following

(a) We establish that the optimal trade-off between exploration (m) and estimation (n) occurs when m is of order $n^{d/4}$, so that in low dimensions ($d \leq 4$), one needs accurate estimation ($n \gg m$) whereas in high-dimensional settings ($d \geq 5$), lots of exploration ($m \gg n$) is necessary; see Theorems 3.2 and 3.3. A related result can be found in Yakowitz et al. [2000], in a setting in which quasi-random sequences are used to determine the placement of the θ_i's.

(b) We present the first derivation of the limit distributions that arise in the context of simple random search; see Theorems 3.2 and 3.3.

This article is organized as follows Section 2 carefully introduces the mathematical setting for our discussion and reviews previous theory on conditions under which random search is consistent, while Section 3 presents the large-sample limit theorems that form the core results of this article. Section 4 provides some brief concluding remarks.

2. PROBLEM FORMULATION AND BASIC THEORY

We start by reviewing the basic theory of simple random search. We assume that $\Lambda \subseteq \mathbb{R}^d$ is a compact set with a nonempty interior. For each $\theta \in \Lambda$, there exists an integrable random variable $X(\theta)$ for which $EX(\theta) = \alpha(\theta)$.

Given a computer (time) budget c, we generate m independent identically distributed (iid) 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; g can then be extended to $\mathbb{R}^d$ by setting its values to zero on $\mathbb{R}^d \setminus \Lambda$. 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 $\{X_j(\theta_i) : 1 \leq i \leq m, 1 \leq j \leq n\}$ having conditional distribution

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

where $F(\theta, \cdot)$ is the distribution function of $X(\theta)$. For $i \geq 1$, 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 $\max\{\alpha(\theta) : \theta \in \Lambda\}$.

Intuitively, it seems clear that if n is too small relative to m, then the noise that is present in the sample mean $\alpha_n(\theta_i)$ will dominate, and the simple random search algorithm will be inconsistent (i.e., nonconvergent). The remainder of this section discusses what is known about consistency for random search, and provides context for the more precise convergence rate results of Section 3.

Conditions under which $\hat{\alpha}(c)$ converges to $\max\{\alpha(\theta) : \theta \in \Lambda\}$ have been previously studied. To describe these results, put $s(\theta) = \sup\{y : P(X(\theta) \le y) < 1\}$, and $s = \sup\{s(\theta) : \theta \in \Lambda\}$, so that s is the maximal value that can be taken on by any of the $X(\theta)$'s. We assume the following.

Assumption 1. $\alpha : \Lambda \to I\!R$ is continuous.

Assumption 2. There exists $t > 0$ such that $\sup\{E \exp(t|X(\theta)|) : \theta \in \Lambda\} < \infty$.

Assumption 3. If $s < \infty$, then there exists, for each $\varepsilon > 0$, a choice of $\theta_0 \in \Lambda$ and $\delta > 0$ such that

$$\inf_{\|\theta - \theta_0\| < \delta} P(X(\theta) > s - \varepsilon) > 0.$$

If $s = \infty$, there exists, for each $x > 0$, a choice of $\theta_0 \in \Lambda$ and $\delta > 0$ such that

$$\inf_{\|\theta - \theta_0\| < \delta} P(X(\theta) > x) > 0.$$

Note that Assumption 2 ensures (via application of Markov's inequality) that the tail of $X(\theta)$ converges to zero exponentially fast uniformly in θ, and is a strong version of what is known in the probability literature as a statement that the $X(\theta)$'s are light-tailed. This assumption is critical to the validity of Theorem 2.1 below. When the $X(\theta)$'s are heavy-tailed, a different consistency theory holds.

For $\theta \in \Lambda, \gamma \in I\!R$, set $\psi(\theta, \gamma) = \log[E \exp(\gamma X(\theta))]$. Part (i) of the following theorem is Theorem $2'$, p. 145, in Devroye [1978]; see also Proposition 4.2 of Ensor and Glynn [1997]. Part (ii) is Proposition 4.1 of that paper and (iii) is Theorem 4.1 of Ensor and Glynn [1997]. (These proofs are given when $\Lambda = [0, 1]^d$, but go over without change in the current setting.)

THEOREM 2.1. *If Assumption* 1*, Assumption* 2*, and Assumption* 3 *hold, then:*

(i) If $(\log m)/n \to 0$ *as* $c \to \infty$*, then*

$$\hat{\alpha}(c) \Rightarrow \max_{\theta \in \Lambda} \alpha(\theta)$$

as $c \to \infty$.

(ii) If $(\log m)/n \to \infty$ *as* $c \to \infty$*, then*

$$\hat{\alpha}(c) \Rightarrow s$$

as $c \to \infty$.

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

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

Furthermore, suppose that ψ *is twice continuously differentiable on* $\Lambda \times [0, \gamma_0]$, *where* $\gamma_0 > \sup_{\theta \in \Lambda} \tilde{\gamma}(\theta)$. *Then,*

$$\hat{\alpha}(c) \Rightarrow \max_{\theta \in \Lambda} \frac{\partial}{\partial \gamma} \psi(\theta, \tilde{\gamma}(\theta))$$

as $c \to \infty$.

Theorem 2.1 establishes that when the noise has a tail that decays exponentially rapidly, then $(\log m)/n \to 0$ as $c \to \infty$ is typically required for consistency of simple random search. For a discussion of consistency in the presence of heavy tails, see Theorem 3.1 of Ensor and Glynn [1997].

3. LIMIT DISTRIBUTIONS FOR THE MAXIMUM

Section 2 establishes conditions under which simple random search is consistent. In this section, our focus is on developing an understanding of the optimal trade-off between m and n for a given value of c.

We first consider 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 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; the three assumptions needed for Theorem 3.1 are completely natural when considering global optimization of smooth objective functions (whether via random search or otherwise).

Assumption 4. $\alpha(\cdot)$ is three times continuously differentiable on Λ.

Assumption 5. $\alpha(\cdot)$ has a unique maximizer θ^* lying in the interior of Λ.

Assumption 6. The Hessian of $\alpha(\cdot)$, when evaluated at θ^* (and denoted $H(\theta^*)$), is negative definite.

Since $H(\theta^*)$ is symmetric and negative definite, the eigenvalues of $H(\theta^*)$ are negative real numbers (see Strang [1986]). We denote the d eigenvalues as

$$-\lambda_1, -\lambda_2, \ldots, -\lambda_d,$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$.

THEOREM 3.1. *If Assumption* 4, *Assumption* 5, *and Assumption* 6 *hold and* $X(\theta) = \alpha(\theta)$ *a.s. for* $\theta \in \Lambda$, *then*

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

as $c \to \infty$, *where* Weibull$(a, d/2)$ *is a* Weibull *random variable with shape parameter* $d/2$ *and scale parameter* $1/a$ *where*

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

PROOF. For $x > 0$, observe that

$$\begin{aligned} & P(\alpha(\theta^*) - \hat{\alpha}(c) \geq xc^{-2/d}) \\ = \; & [P(\alpha(\theta_1) \leq \alpha(\theta^*) - xc^{-2/d})]^m \\ = \; & \exp\{m \log[1 - P(\alpha(\theta_1) > \alpha(\theta^*) - xc^{-2/d})]\}. \end{aligned} \tag{2}$$

Since $\alpha(\cdot)$ has a unique maximizer and g is continuous on Λ, $P(\alpha(\theta_1) > \alpha(\theta^*) - xc^{-2/d}) \searrow 0$ as $c \to \infty$. So

$$\begin{aligned} & \log(1 - P(\alpha(\theta_1) > \alpha(\theta^*) - xc^{-2/d})) \\ = \; & -P(\alpha(\theta_1) > \alpha(\theta^*) - xc^{-2/d})(1 + o(1)) \end{aligned} \tag{3}$$

as $c \to \infty$. But

$$\begin{aligned} & P(\alpha(\theta_1) > \alpha(\theta^*) - xc^{-2/d}) \\ = \; & \int_\Lambda I(xc^{-2/d} > \alpha(\theta^*) - \alpha(y))g(y)\, dy. \end{aligned} \tag{4}$$

Furthermore, Assumption 4 ensures that

$$\alpha(\theta^*) - \alpha(y) = -\frac{1}{2}(y - \theta^*)^T H(\theta^*)(y - \theta^*) + o(\| y - \theta^* \|^2) \tag{5}$$

as $y \to \theta^*$. The uniqueness of the maximizer guarantees that for each $\varepsilon > 0$,

$$\begin{aligned} & \{y : xc^{-2/d} > -\frac{1}{2}(1 + \varepsilon)(y - \theta^*)^T H(\theta^*)(y - \theta^*)\} \\ \subseteq \; & \{y \in \Lambda : xc^{-2/d} > \alpha(\theta^*) - \alpha(y)\} \\ \subseteq \; & \{y : xc^{-2/d} > -\frac{1}{2}(1 - \varepsilon)(y - \theta^*)^T H(\theta^*)(y - \theta^*)\} \end{aligned}$$

for c large enough. To see why this is true, note that (5) implies that if $Q(y) = -(1/2)(y - \theta^*)^T H(\theta^*)(y - \theta^*)$, then

$$\alpha(\theta^*) - \alpha(y) = Q(y) + o(\|y - \theta^*\|^2).$$

But since $Q(y) \geq (\lambda_1/2)\|y - \theta^*\|^2$, it follows that there exists $\delta > 0$ for which $\alpha(\theta^*) - \alpha(y) \geq (1 - \varepsilon)Q(y)$ for $\|y - \theta^*\| \leq \delta$. For c sufficiently large, the uniqueness of the maximizer θ^* and continuity of $\alpha(\cdot)$ guarantee that $\{y \in \Lambda : xc^{-2/d} > \alpha(\theta^*) - \alpha(y)\} \subseteq \{y \in \Lambda : \|y - \theta^*\| < \delta\}$. So, for such values of c, $\{y \in \Lambda : xc^{-2/d} \geq (1 - \varepsilon)Q(y)\} \supseteq \{y \in \Lambda : xc^{-2/d} > \alpha(\theta^*) - \alpha(y)\}$. The other set inclusion can be similarly argued.

Hence, an upper bound on (4) is

$$\int_{\mathbb{R}^d} I\left(xc^{-2/d} > -\frac{1}{2}(1 - \varepsilon)(y - \theta^*)^T H(\theta^*)(y - \theta^*)\right) g(y)\, dy. \tag{6}$$

Because $-H(\theta^*)$ is a symmetric matrix, it can be diagonalized via an orthogonal matrix A (see Strang [1986, p. 254]). In particular, there exists a $d \times d$ matrix A such that

$A^TA = I$, with $-H(\theta^*) = A^TDA$, where $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$. Hence, this integral can be re-written as

$$\begin{aligned}
&\int_{\mathbb{R}^d} I\left(\frac{2x}{1-\varepsilon} > (c^{1/d}A(y-\theta^*))^T D(c^{1/d}A(y-\theta^*))\right) g(y)\, dy \\
&= c^{-1}\int_{\mathbb{R}^d} I\left(\frac{2x}{1-\varepsilon} > z^TDz\right) g(\theta^* + c^{-1/d}A^Tz)\, dz \\
&= \frac{1}{c\sqrt{\lambda_1\lambda_2\cdots\lambda_d}}\int_{\mathbb{R}^d} I\left(\frac{2x}{1-\varepsilon} > \sum_{i=1}^d \tilde{z}_i^2\right) g(\theta^* + c^{-1/d}A^TD^{-1/2}\tilde{z})\, d\tilde{z}.
\end{aligned}$$

(We used the fact that $|\det A| = 1$ in the change-of-variables in the first equality.)

Note that

$$g(\theta^* + c^{-1/d}A^TD^{-1/2}\tilde{z}) \to g(\theta^*)$$

as $c \to \infty$ uniformly over the region $\{\tilde{z} : 2x(1-\varepsilon)^{-1} > \sum_{i=1}^d \tilde{z}_i^2\}$ because $\|c^{-1/d}A^TD^{-1/2}\tilde{z}\| \le c^{-1/d}\{2x/[\lambda_1(1-\varepsilon)]\}^{1/2}$ for every $\tilde{z}$ in that region. Also,

$$\int_{\mathbb{R}^d} I\left(\frac{2x}{1-\varepsilon} > \sum_{i=1}^d \tilde{z}_i^2\right) d\tilde{z}$$

is the volume of a d-sphere having radius $(2x/(1-\varepsilon))^{1/2}$. It therefore equals

$$\left(\frac{2\pi x}{1-\varepsilon}\right)^{d/2} \frac{1}{\Gamma(d/2+1)};$$

see, for example, Zhigljavsky [1991, pg. 78]. Finally, $\lambda_1\lambda_2\cdots\lambda_d = \det D = |\det H(\theta^*)|$, so we conclude that the integral (6) equals

$$\frac{1}{c\sqrt{|\det H(\theta^*)|}\Gamma(d/2+1)} g(\theta^*) \left(\frac{2\pi x}{1-\varepsilon}\right)^{d/2} (1+o(1)),$$

as $c \to \infty$. By appealing to (2),(3), (4), and (6), and using the fact that $m = \lfloor c \rfloor$, in this context, we therefore find that

$$\begin{aligned}
&\liminf_{c\to\infty} P(\alpha(\theta^*) - \hat{\alpha}(c) \ge xc^{-2/d}) \\
&\ge \exp\left(-\left(\frac{ax}{1-\varepsilon}\right)^{d/2}\right).
\end{aligned}$$

Similarly, we find that

$$\begin{aligned}
&\limsup_{c\to\infty} P(\alpha(\theta^*) - \hat{\alpha}(c) \ge xc^{-2/d}) \\
&\le \exp\left(-\left(\frac{ax}{1+\varepsilon}\right)^{d/2}\right).
\end{aligned}$$

Since ε was arbitrary, this proves the theorem. □

The Weibull structure of the limit was previously identified by Archetti et al. [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 3.1 shows that when function evaluations are deterministic, then the rate of convergence to the maximum is of order $c^{-2/d}$. This result makes clear that simple random search degrades rapidly when the dimension d is large, even when the function can be evaluated without (random) error.

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 suggests 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

$$\begin{aligned} m &\sim rc^{d/(d+4)} \\ n &\sim r^{-1}c^{4/(d+4)} \end{aligned} \tag{7}$$

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

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

Assumption 7. 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 8. $\operatorname{var} X(\theta^*) > 0$.

Assumption 7 is a mild assumption in practice, and merely asserts that the distribution of the (random) noise varies continuously in the decision variable θ, while Assumption 8 is a nondegeneracy assumption that asserts that nonzero noise is present when evaluating the objective function at the maximizer.

Set $\sigma(\theta) = \sqrt{\operatorname{var} X(\theta)}$. Our next result describes the behavior of $\hat{\alpha}(c)$ where m and n are balanced according to (7).

THEOREM 3.2. *Suppose that Assumption* 4 *through Assumption* 8 *hold and* $\sup\{E|X(\theta)|^p : \theta \in \Lambda\} < \infty$ *for* $p > \max(3, d^2)$. *If m and n satisfy (7), then*

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

as $c \to \infty$, where

$$P(\beta \le x) = \exp\left(-\frac{r^{(4+d)/4} g(\theta^*)\pi^{d/2}}{\Gamma(d/2)\sqrt{|\det H(\theta^*)|}} \int_0^\infty P\left(N(0,1) > \frac{2x+y}{2\sigma(\theta^*)}\right) y^{d/2-1}\, dy\right).$$

PROOF. We start by observing that

$$\begin{aligned} & P(n^{1/2}(\hat{\alpha}(c) - \alpha(\theta^*)) \le x) \\ =\ & P(\max_{1 \le i \le m} \alpha_n(\theta_i) - \alpha(\theta^*) \le xn^{-1/2}) \\ =\ & [P(\alpha_n(\theta_1) \le \alpha(\theta^*) + xn^{-1/2})]^m \\ =\ & \exp(m \log[1 - P(\alpha_n(\theta_1) > \alpha(\theta^*) + xn^{-1/2})]\,). \end{aligned} \tag{8}$$

For $\delta > 0$, let $B_0(\delta) = \{\theta :\| \theta - \theta^* \| \leq \delta\}$, where $\| \cdot \|$ is the Euclidean norm on $I\!R^d$. Then, for $\varepsilon > 0$,

$$\begin{aligned} &P(\alpha_n(\theta_1) > \alpha(\theta^*) + xn^{-1/2}) \\ &= \int_{B_0(n^{-\varepsilon})} P(\alpha_n(\theta) > \alpha(\theta^*) + xn^{-1/2}) g(\theta)\, d\theta \\ &\quad + \int_{[B_0(n^{-\varepsilon})]^c \cap \Lambda} P(\alpha_n(\theta) > \alpha(\theta^*) + xn^{-1/2}) g(\theta)\, d\theta. \end{aligned} \tag{9}$$

The uniform boundedness of $E|X(\theta)|^p$ (with $p > 2$) ensures that $(X^2(\theta) : \theta \in \Lambda)$ and $(X(\theta) : \theta \in \Lambda)$ are uniformly integrable families of rv's. In view of the weak continuity of $P(X(\theta) \in \cdot)$ in θ, it follows that $\sigma^2(\theta)$ is continuous in θ over Λ. Put $\tilde{\alpha}_n(\theta) = n^{1/2}(\alpha_n(\theta) - \alpha(\theta))$. The positivity of $\sigma^2(\theta^*)$ allows us to write, for n so large that $B_0(n^{-\epsilon}) \subset \{\theta \in \Lambda : \sigma^2(\theta) > 0\}$,

$$\begin{aligned} &\int_{B_0(n^{-\varepsilon})} P(\alpha_n(\theta) > \alpha(\theta^*) + xn^{-1/2}) g(\theta)\, d\theta \\ &= \int_{B_0(n^{-\varepsilon})} P\left(\frac{\tilde{\alpha}_n(\theta)}{\sigma(\theta)} > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)} \right) g(\theta)\, d\theta. \end{aligned}$$

The Berry-Esseen theorem asserts that for $\theta \in B_0(n^{-\varepsilon})$,

$$\begin{aligned} &\left| P\left(\frac{\tilde{\alpha}_n(\theta)}{\sigma(\theta)} > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)} \right) - P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)} \right) \right| \\ &\leq 3n^{-1/2} \frac{\sup_{\theta \in \Lambda} E[\, |X(\theta) - \alpha(\theta)|^3]}{\inf_{\theta \in B_0(n^{-\varepsilon})} [\, \sigma(\theta)]^3}; \end{aligned}$$

see, for example, p. 542 of Feller [1971]. Hence, the first integral on the right-hand side of (9) equals

$$\begin{aligned} &\int_{B_0(n^{-\varepsilon})} P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)} \right) g(\theta)\, d\theta + n^{-1/2} O(\mathrm{vol}(B_0(n^{-\varepsilon}))) \\ &= \int_{B_0(n^{-\varepsilon})} P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)} \right) g(\theta)\, d\theta + O(n^{-1/2 - d\varepsilon}). \end{aligned} \tag{10}$$

Turning next to the second integral on the right-hand side of (9), we start by noting that Markov's inequality shows that

$$\begin{aligned} &P(\alpha_n(\theta) > \alpha(\theta^*) + xn^{-1/2}) \\ &= P(\tilde{\alpha}_n(\theta) > n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x) \\ &\leq |n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x|^{-p} E|\tilde{\alpha}_n(\theta)|^p. \end{aligned} \tag{11}$$

But, according to Petrov [1995, p. 62],

$$E|\tilde{\alpha}_n(\theta)|^p \leq C(p) E[|X_1(\theta) - \alpha(\theta)|^p], \tag{12}$$

where $C(p)$ is a universal constant (not depending on the distribution of $X_1(\theta)$). Also, the compactness of Λ guarantees that for each n, there exists $\theta_n \in \Lambda$ that maximizes α over $[B_0(2n^{-\varepsilon})]^c \cap \Lambda$. Clearly,

$$\tilde{\theta}_n = \theta^* + n^{-\varepsilon}(\theta_n - \theta^*)/\|\theta_n - \theta^*\| \in \{\theta : n^{-\varepsilon} \le \|\theta - \theta^*\| \le 2n^{-\varepsilon}\}$$

and note that $\tilde{\theta}_n - \theta^* = r_n(\theta_n - \theta^*)$ with $0 < r_n \le 1/2$. Furthermore, we claim that $\alpha(\tilde{\theta}_n) > \alpha(\theta_n)$ for n sufficiently large. Note that if this is false, then exists a subsequence $\theta_{n'}$ for which $\alpha(\tilde{\theta}_{n'}) \le \alpha(\theta_{n'})$. Since θ^* is the unique maximizer of α, it is evident that $\theta_{n'} \to \theta^*$. Let $z_{n'} = A(\theta_{n'} - \theta^*)$, $\tilde{z}_{n'} = A(\tilde{\theta}_{n'} - \theta^*)$ (where A is the orthogonal matrix introduced in the proof of Theorem 3.1), and observe that $\tilde{z}_{n'} = r_{n'} z_{n'}$. Also,

$$\begin{aligned} &\alpha(\theta_{n'}) - \alpha(\tilde{\theta}_{n'}) \\ &= -(1 - r_{n'}^2) \sum_{i=1}^{d} \lambda_i \frac{z_{n',i}^2}{2} + o(\|z_{n'}\|^2), \end{aligned}$$

which is negative for n sufficiently large, yielding a contradiction. We conclude that

$$\begin{aligned} &\min\{|\alpha(\theta) - \alpha(\theta^*)| : \theta \in [B_0(n^{-\varepsilon})]^c \cap \Lambda\} \\ &= \min\{|\alpha(\theta) - \alpha(\theta^*)| : n^{-\varepsilon} \le \|\theta - \theta^*\| \le 2n^{-\varepsilon}\} \\ &= \min\{(1/2)|(\theta - \theta^*)^T H(\theta^*)(\theta - \theta^*) + o(n^{-2\varepsilon})| : n^{-\varepsilon} \le \|\theta - \theta^*\| \le 2n^{-\varepsilon}\}(1 + o(1)) \\ &= O(n^{-2\varepsilon}) \end{aligned} \tag{13}$$

as $n \to \infty$. It follows from (11), (12), and (13) that

$$\begin{aligned} &\sup_{\theta \in [B_0(n^{-\varepsilon})]^c \cap \Lambda} P(\alpha_n(\theta) > \alpha(\theta^*) + x n^{-1/2}) \\ &= O(n^{-p(1/2 - 2\varepsilon)}) \end{aligned}$$

as $n \to \infty$, and consequently

$$\begin{aligned} &\int_{[B_0(n^{-\varepsilon})]^c \cap \Lambda} P(\alpha_n(\theta) > \alpha(\theta^*) + x n^{-1/2}) g(\theta)\, d\theta \\ &= O(n^{-p(1/2 - 2\varepsilon)}). \end{aligned} \tag{14}$$

Put $\varepsilon = 1/4 - 1/(8d)$. Relations (10) and (14) then guarantee that

$$P(\alpha_n(\theta_1) > \alpha(\theta^*) + x n^{-1/2}) \to 0$$

as $n \to \infty$, so that

$$\begin{aligned} &m \log[1 - P(\alpha_n(\theta_1) > \alpha(\theta^*) + x n^{-1/2})] \\ &\sim -mP(\alpha_n(\theta_1) > \alpha(\theta^*) + x n^{-1/2}). \end{aligned} \tag{15}$$

as $n \to \infty$. Also, with this choice for ε, (10) and (14) in turn establish that

$$\begin{aligned}
&mP(\alpha_n(\theta_1) > \alpha(\theta^*) + xn^{-1/2}) \\
&= m\int_{B_0(n^{-\varepsilon})} P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)}\right) g(\theta)\,d\theta + O(mn^{-1/2-d\varepsilon}) \\
&\quad + O(mn^{-p(1/2-2\varepsilon)}) \\
&= m\int_{B_0(n^{-\varepsilon})} P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)}\right) g(\theta)\,d\theta + O(n^{-3/8}) \\
&\quad + O(n^{d/4-p/(4d)}) \\
&= m\int_{B_0(n^{-\varepsilon})} P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta)) + x}{\sigma(\theta)}\right) g(\theta)\,d\theta + o(1) \qquad (16)
\end{aligned}$$

as $c \to \infty$.

To evaluate this integral, using the fact that the symmetric negative definite matrix $H(\theta^*)$ can be diagonalized via an orthogonal similarity transformation, we now do a change of variables so that $y = n^{1/4}A(\theta - \theta^*)$. Let $B(\delta) = \{y \in I\!R^d : \|y\| \le \delta\}$. Since $|\det(A)| = 1$,

$$\begin{aligned}
&\int_{B_0(n^{-\varepsilon})} P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta))}{\sigma(\theta)} + \frac{x}{\sigma(\theta)}\right) g(\theta)\,d\theta \\
&= n^{-d/4}\int_{B(n^{1/4-\varepsilon})} \left[P\left(N(0,1) > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta^* + n^{-1/4}A^Ty)) + x}{\sigma(\theta^* + n^{-1/4}A^Ty)}\right)\right. \\
&\qquad \left. \times\, g(\theta^* + n^{-1/4}A^Ty)\right] dy.
\end{aligned}$$

Uniformly in $y \in B(n^{1/4-\varepsilon})$,

$$\begin{aligned}
&\left(\frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta^* + n^{-1/4}A^Ty)) + x)}{\sigma(\theta^* + n^{-1/4}A^Ty)}\right) \\
&= \frac{-y^TAH(\theta^*)A^Ty + 2x}{2\sigma(\theta^*)}(1 + o(1)) \\
&= \frac{(y^TDy + 2x)}{2\sigma(\theta^*)}(1 + o(1))
\end{aligned}$$

and

$$g(\theta^* + n^{-1/4}A^Ty) = g(\theta^*)(1 + o(1))$$

as $n \to \infty$. Consequently,

$$\begin{aligned}
&m\int_{B_0(n^{-\varepsilon})} P\left(\frac{\tilde{\alpha}_n(\theta)}{\sigma(\theta)} > \frac{n^{1/2}(\alpha(\theta^*) - \alpha(\theta))}{\sigma(\theta)} + \frac{x}{\sigma(\theta)}\right) g(\theta)\,d\theta \\
&\sim r^{(4+d)/4}\int_{B(n^{1/4-\varepsilon})} P\left\{N(0,1) > \left(\frac{2x + \sum_{i=1}^d \lambda_i y_i^2}{2\sigma(\theta^*)}\right)\right\} g(\theta^*)(1 + o(1))\,dy. \qquad (17)
\end{aligned}$$

Because $P(N(0,1) > (x + 2^{-1} \sum_{i=1}^{d} \lambda_i y_i^2)/\sigma(\theta^*))$ is integrable over $I\!R^d$ and dominates the integrand of (17) uniformly in n, it follows that the integral (17) converges to

$$g(\theta^*) r^{(4+d)/4} \int_{I\!R^d} P\left(N(0,1) > \frac{2x + \sum_{i=1}^{d} \lambda_i y_i^2}{2\sigma(\theta^*)}\right) dy \tag{18}$$

as $n \to \infty$. To simplify the limit distribution further, let $z_i = \sqrt{\lambda_i} y_i$. Then, it follows that

$$\begin{aligned} &\int_{I\!R^d} P\left(N(0,1) > \frac{2x + \sum_{i=1}^{d} \lambda_i y_i^2}{2\sigma(\theta^*)}\right) dy \\ &= \frac{1}{\sqrt{\lambda_1 \lambda_2 \ldots \lambda_d}} \int_{I\!R^d} P\left(N(0,1) > \frac{2x + \sum_{i=1}^{d} z_i^2}{2\sigma(\theta^*)}\right) dz \\ &= \frac{1}{\sqrt{|\det H(\theta^*)|}} \int_{I\!R^d} P\left(N(0,1) > \frac{x}{\sigma(\theta^*)} + \frac{1}{2\sigma(\theta^*)} \sum_{i=1}^{d} z_i^2\right) dz. \end{aligned} \tag{19}$$

Finally, we convert from Cartesian coordinates to hyperspherical coordinates using the transformation

$$\begin{aligned} r &= \sqrt{\textstyle\sum_{i=1}^{d} z_i^2}, \\ \phi_1 &= \text{arc cot}\left(z_1/\sqrt{z_2^2 + \ldots + z_d^2}\right), \\ \phi_2 &= \text{arc cot}\left(z_2/\sqrt{z_3^2 + \ldots + z_d^2}\right), \\ &\vdots \\ \phi_{d-2} &= \text{arc cot}\left(z_{d-2}/\sqrt{z_{d-1}^2 + z_d^2}\right), \\ \phi_{d-1} &= \text{arc cot}\left(z_{d-1}/z_d\right). \end{aligned}$$

With this transformation, $\phi_i \in [0, \pi]$ for $1 \le i \le d-2$, $\phi_{d-1} \in [0, 2\pi]$, and

$$|\det J(r, \phi_1, \ldots, \phi_{d-1})| = r^{d-1} \sin^{d-2}(\phi_1) \sin^{d-3}(\phi_2) \cdots \sin(\phi_{d-2});$$

see Edwards [1995, p. 268].

Note that

$$\begin{aligned} \text{vol}(B(1)) &= \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} \int_0^{1} |\det J(r, \phi_1, \cdots, \phi_{d-1})| \, dr \, d\phi_1 \cdots d\phi_{d-1} \\ &= \frac{1}{d} \int_0^{2\pi} \left[\prod_{k=1}^{d-2} \int_0^{\pi} \sin^{d-1-k}(\phi_k) \, d\phi_k\right] d\phi_{d-1}. \end{aligned}$$

Then, it follows from p. 342 of Edwards [1995] that

$$\begin{aligned} (\textit{surface area of } B(1)) &= d \cdot \text{vol}(B(1)) \\ &= \int_0^{2\pi} \left[\prod_{k=1}^{d-2} \int_0^{\pi} \sin^{d-1-k}(\phi_k) \, d\phi_k\right] d\phi_{d-1}. \end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_{\mathbb{R}^d} P\left(N(0,1) > x[\sigma(\theta^*)]^{-1} + (2\sigma(\theta^*))^{-1}\sum_{i=1}^{d} z_i^2\right) dz\\
&= \int_0^\infty \int_0^{2\pi} \left[\prod_{k=1}^{d-2}\int_0^\pi \sin^{d-1-k}(\phi_k)d\phi_k\right] d\phi_{d-1} P\left(N(0,1) > \frac{2x+r^2}{2\sigma(\theta^*)}\right) r^{d-1}dr\\
&= (\textit{surface area of } B(1)) \cdot \int_0^\infty P\left(N(0,1) > \frac{2x+r^2}{2\sigma(\theta^*)}\right) r^{d-1}dr\\
&= \frac{d\pi^{d/2}}{\Gamma(d/2+1)} \int_0^\infty P\left(N(0,1) > \frac{2x+r^2}{2\sigma(\theta^*)}\right) r^{d-1}dr\\
&= \frac{d\pi^{d/2}}{2\Gamma(d/2+1)} \int_0^\infty P\left(N(0,1) > \frac{2x+y}{2\sigma(\theta^*)}\right) y^{d/2-1}dy\\
&= \frac{\pi^{d/2}}{\Gamma(d/2)} \int_0^\infty P\left(N(0,1) > \frac{2x+y}{2\sigma(\theta^*)}\right) y^{d/2-1}dy. \qquad (20)
\end{aligned}$$

The theorem then follows from (8), (14), (15),(16), (17), and (20). □

According to Theorem 3.2, the rate of convergence in the asymptotic regime (7) 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 3.1). 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 the points $\theta_1, \theta_2, \ldots$ are generated. By learning effect, we mean that one can devise algorithms that "learn" the shape of the objective function over time, and direct their sampling effort (in terms of point placement and sample sizes used) to those regions of the feasible set that appear most promising.

A key difference in the setting of random search with "noise" (as opposed to random search in the presence of deterministic function evaluations) is that the error $\hat{\alpha}(c) - \alpha(\theta^*)$ can be of either positive or negative sign, whereas the error in the deterministic case can only be negative; see Theorem 3.1. Furthermore, unlike a normal distribution, the asymptotic distribution of the error is not symmetric about zero. Note that the limit distribution takes the form

$$P(\beta \le x) = \exp\left(-w\int_0^\infty P\left[N(0,1) > \frac{2x+y}{2\sigma(\theta^*)}\right] y^{d/2-1}\, dy\right),$$

where $w = (r^{(4+d)/4} g(\theta^*)\pi^{d/2})/(\Gamma(d/2)\sqrt{|\det H(\theta^*)|})$. The standard deviation $\sigma(\theta^*)$ acts as a scale parameter, so that larger values of $\sigma(\theta^*)$ make the error distribution more diffuse. But the behavior of the error distribution as a function of the parameter w is much more complex, as w affects the shape of the distribution in a nontrivial manner.

Note that if one sends $w \to 0$, then $P(\beta \leq x) \to 1$ for each x, establishing that $\beta \overset{P}{\to} -\infty$ as $w \to 0$. On the other hand, if $w \to \infty$, $P(\beta \leq x) \to 0$ for each x, establishing that $\beta \overset{P}{\to} \infty$ as $w \to \infty$. In other words, the magnitude of the error is not monotone as a function w, suggesting there is a value w^* that minimizes (for example) $E|\beta|$ for fixed $\sigma(\theta^*)$. So, a w that is either "too large" or "too small" creates an asymptotic error for random search that is substantial. For example, if r is too small, then the number m of sampled points is too small (relative to n) and it is less likely that one of the sampled points will be close to θ^*, so that the value of the objective function at the closest sampled point will be significantly smaller than $\alpha(\theta^*)$. This creates a strong negative bias in $\hat{\alpha}(c)$ so that the error $\hat{\alpha}(c) - \alpha(\theta^*)$ tends to be quite negative (which, of course, is in agreement with the fact that $\beta \overset{P}{\to} -\infty$ as $w \to 0$). On the other hand, if r is large, then we will have more sampled points close to θ^*, with smaller sample sizes at each point. Consequently, there will be more points, with objective values very close to $\alpha(\theta^*)$, at which their associated approximately normal "noise" (with larger standard deviation, due to n being reduced) will be competing to be the maximum. The maximum of this collection of normal random variables will therefore be made larger, pushing the probability mass of the error $\hat{\alpha}(c) - \alpha(\theta^*)$ towards larger values (as is consistent with the observation that $\beta \overset{P}{\to} \infty$ as $w \to \infty$). Similarly, if $g(\theta^*)$ is small, there will be fewer sampled points close to θ^*, increasing the tendency of the error to be negative (as in our discussion of r), while if $g(\theta^*)$ is (very) big, there will be many sampled points close to θ^*, with all the associated (approximately normally distributed) errors competing to be the maximum, thereby increasing the tendency of the error to be positive (as for r). Finally, if $|\det H(\theta^*)|$ is large, the objective function is very peaked near θ^*, so that the sampled points need to be close to θ^* in order that the bias (as an estimate of $\alpha(\theta^*)$) be small. Thus, the error $\hat{\alpha}(c) - \alpha(\theta^*)$ tends to be negative (due to large negative bias) when $|\det H(\theta^*)|$ is large (which is again consistent with $\beta \overset{P}{\to} -\infty$ as $w \to 0$). On the other hand, if $|\det H(\theta^*)|$ is small, the objective function is "flat" in a neighborhood of θ^*, increasing the magnitude of the error, due to the presence of more sampled points at which the associated "noise" can compete to be the maximum (as is suggested by $\beta \overset{P}{\to} \infty$ as $w \to \infty$).

We turn next to identifying the convergence rate of the simple random search when m and n are chosen so that the algorithm is consistent but fails to satisfy (7).

THEOREM 3.3. *If Assumption* 2 *and Assumptions* 4 *through* 8 *hold, then*

(i) If $m \sim rc^q$ as $c \to \infty$, with $0 < q < d/(d+4)$ and $0 < r < \infty$, then

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

as $c \to \infty$, where a is defined as in Theorem 3.1.

(ii) If $m \sim rc^q$ as $c \to \infty$, with $d/(d+4) < q < 1$ and $0 < r < \infty$, then

$$\left(\frac{c^{1-q}}{\log c}\right)^{1/2} (\hat{\alpha}(c) - \alpha(\theta^*)) \Rightarrow \sigma(\theta^*)\sqrt{2r\left(\frac{4+d}{4}\right)\left(q - \frac{d}{d+4}\right)}$$

as $c \to \infty$.

PROOF. For part (i), observe that

$$\begin{aligned}
& m^{2/d} |\max_{1\le i\le m} \alpha(\theta_i) - \hat{\alpha}(c)| \\
={}& m^{2/d} |\max_{1\le i\le m} (\alpha(\theta_i) - \hat{\alpha}(c))| \\
={}& m^{2/d} n^{-1/2} |\max_{1\le i\le m} n^{1/2}(\alpha(\theta_i) - \hat{\alpha}(c))| \\
\le{}& m^{2/d} n^{-1/2} \max_{1\le i\le m} |\tilde{\alpha}_n(\theta_i)| \\
={}& r^{2/d+1/2} c^{-(1/2)(1-q(4+d)/d)} \max_{1\le i\le m} |\tilde{\alpha}_n(\theta_i)|(1+o(1)),
\end{aligned}$$

where $\tilde{\alpha}(\cdot)$ is again defined, as in the proof of Theorem 3.2, to be $\tilde{\alpha}_n(\theta) = n^{1/2}(\alpha_n(\theta) - \alpha(\theta))$ and the $o(1)$ term absorbs the difference between m and n and their asymptotic values rc^q and $r^{-1}c^{1-q}$, respectively.

Put $p_0 = (1/2)\big(1 - q(4+d)/d\big)$. For $\varepsilon > 0$, the union bound and Markov's inequality implies

$$\begin{aligned}
& P(c^{-p_0} \max_{1\le i\le m} |\tilde{\alpha}_n(\theta_i)| > \varepsilon) \\
\le{}& mP(|\tilde{\alpha}_n(\theta_1)| > \varepsilon c^{p_0}) \\
\le{}& C(p) r \varepsilon^{-p} c^{q-pp_0} \int_\Lambda E|X_1(\theta) - \alpha(\theta)|^p g(\theta)\, d\theta; \qquad (21)
\end{aligned}$$

see (12) for the last inequality. If we use Assumption 2, it follows that (21) converges to zero as $c \to \infty$ for p sufficiently large, and consequently

$$m^{2/d}(\hat{\alpha}(c) - \max_{1\le i\le m} \alpha(\theta_i)) \xrightarrow{P} 0$$

as $c \to \infty$. Theorem 3.1 therefore implies that

$$\begin{aligned}
& m^{2/d}(\alpha(\theta^*) - \hat{\alpha}(c)) \\
={}& m^{2/d}\{\alpha(\theta^*) - \max_{1\le i\le m} \alpha(\theta_i)\} + m^{2/d}\{\max_{1\le i\le m} \alpha(\theta_i) - \hat{\alpha}(c)\} \\
\Rightarrow{}& \text{Weibull}(a, d/2) + 0 = \text{Weibull}(a, d/2)
\end{aligned}$$

as $c \to \infty$, proving part (i).

For part (ii), observe that, for $x > 0$,

$$\begin{aligned}
& P\left(\hat{\alpha}(c) \le \alpha(\theta^*) + x\sqrt{\frac{\log c}{c^{1-q}}}\right) \\
={}& \exp\left[m \log\left\{1 - \int_{\mathbb{R}^d} g(\theta) P\left(\alpha_n(\theta) > \alpha(\theta^*) + x\sqrt{\frac{\log c}{c^{1-q}}}\right) d\theta\right\}\right]. \qquad (22)
\end{aligned}$$

This integral is lower bounded for c sufficiently large, by

$$
\begin{aligned}
&\int_{B(1)} \left[\frac{g(\theta^* + n^{-1/4}A^T y)}{n^{d/4}} \right. \\
&\qquad \left. \times P\left(\tilde{\alpha}_n(\theta^* + n^{-1/4}A^T y) > \frac{1}{2} y^T D y (1+o(1)) + x\sqrt{\frac{n \log c}{c^{1-q}}}\right)\right] dy \\
&\geq \frac{\inf_{\theta \in \Lambda} g(\theta)}{n^{d/4}} \cdot \int_{B(1)} P\left(\tilde{\alpha}_n(\theta^* + n^{-1/4}A^T y) > \lambda_d + x\sqrt{\frac{\log c}{r}}\right) dy \\
&\geq \frac{\inf_{\theta \in \Lambda} g(\theta) \cdot \mathrm{vol}(B(1))}{n^{d/4}} \cdot \inf_{\|\theta - \theta^*\| \leq n^{-1/4}} P\left(\tilde{\alpha}_n(\theta) \geq \lambda_d + x\sqrt{\frac{\log c}{r}}\right) \\
&= \frac{\inf_{\theta \in \Lambda} g(\theta) \cdot \mathrm{vol}(B(1))}{n^{d/4}} \inf_{\|\theta - \theta^*\| \leq n^{-1/4}} P\left(N(0,1) \geq \frac{\lambda_d + x\sqrt{r^{-1}\log c}}{\sigma(\theta)}\right)(1+o(1)),
\end{aligned}
$$

where the $o(1)$ terms are uniform over their respective domains. This final equality follows from a careful consideration of the argument of p. 548–552 of Feller [1971], in support of the Corollary stated there on p. 552; such considerations make clear that the $o(1)$ term can indeed be taken to be uniform, in view of the bounded exponential moments implied by Assumption 2. Using a well-known asymptotic for the normal tail probability (see, for example, p. 175 of Feller [1968]) then establishes the following lower bound on the integral in (22), namely

$$
\begin{aligned}
&= \inf_{\theta \in \Lambda} g(\theta) \cdot n^{-d/4} \cdot \mathrm{vol}(B(1)) \\
&\quad \times \inf_{\|\theta - \theta^*\| \leq n^{-1/4}} \frac{\sigma(\theta) \exp\left\{-(\lambda_d + x\sqrt{r^{-1}\log c})^2/[2\sigma^2(\theta)]\right\}}{\sqrt{2\pi}(\lambda_d + x\sqrt{r^{-1}\log c})}(1+o(1)).
\end{aligned}
$$

When multiplied by m, the lower bound can therefore be represented as

$$
\begin{aligned}
&\exp\left[\log c \left(q - (1-q)\frac{d}{4} - \frac{x^2}{2r\sigma^2(\theta^*)}\right) + o(\log c)\right] \\
&= \exp\left[\log c \left(\frac{4+d}{4}\left\{q - \frac{d}{d+4}\right\} - \frac{x^2}{2r\sigma^2(\theta^*)}\right) + o(\log c)\right] \qquad (23)
\end{aligned}
$$

by the continuity of $\sigma^2(\theta)$ in $B_0(n^{-1/4})$.

As a consequence, if $x < \sqrt{2r\sigma^2(\theta^*)(1+d/4)(q-d/(d+4))}$, this quantity converges to infinity as $c \to \infty$. This implies that $P(\hat{\alpha}(c) \leq \alpha(\theta^*) + x\sqrt{\log c / c^{1-q}}) \to 0$ as $c \to \infty$.

We now need to show that if $x > \sqrt{2r\sigma^2(\theta^*)(1+d/4)(q-d/(d+4))}$, then $P(\hat{\alpha}(c) \leq \alpha(\theta^*) + x\sqrt{\log c/c^{1-q}}) \to 1$ as $c \to \infty$. The integral appearing in (22) can be written as

$$\int_{B(k\sqrt{\log c})} \left[\frac{g(\theta^* + n^{-1/4}A^T y)}{n^{d/4}} \right.$$
$$\left. \times\ P\left(\tilde{\alpha}_n(\theta^* + n^{-1/4}A^T y) > \frac{1}{2} y^T D y(1+o(1)) + x\sqrt{\frac{n \log c}{c^{1-q}}} \right) \right] dy$$
$$+ \int_{[B(k\sqrt{\log c})]^c} \left[\frac{g(\theta^* + n^{-1/4}A^T y)}{n^{d/4}} \right.$$
$$\left. \times\ P\left(\tilde{\alpha}_n(\theta^* + n^{-1/4}A^T y) > n^{1/2}\{\alpha(\theta^*) - \alpha(\theta^* + n^{-1/4}A^T y)\} + x\sqrt{\frac{n \log c}{c^{1-q}}} \right) \right] dy. \quad (24)$$

For the first integral, we can upper bound it via

$$(k\sqrt{\log c})^d \mathrm{vol}(B(1)) \cdot n^{-d/4} \sup_{\|\theta-\theta^*\| \leq n^{-1/4}} g(\theta) \cdot \sup_{\|\theta-\theta^*\| \leq n^{-1/4}} P\left(\tilde{\alpha}_n(\theta) \geq \sqrt{\frac{n \log c}{c^{1-q}}} \right).$$

Applying the same arguments as leading to (23) establishes that this upper bound, when multiplied by m, tends to zero for x in the above range.

For the second integral, we note that $n^{1/2}\{\alpha(\theta^*) - \alpha(\theta^* + n^{-1/4}A^T y)\} \geq (\lambda_1/2)k^2 \log c \cdot (1+o(1))$ in $[B(k\sqrt{\log c})]^c$, so $n^{1/2}\{\alpha(\theta^*) - \alpha(\theta^* + n^{-1/4}A^T y)\} \geq (\lambda_1/4)k^2 \log c$ for c sufficiently large. So, the second integral can be upper bounded by

$$\int_{[B(k\sqrt{\log c})]^c} \frac{g(\theta^* + n^{-1/4}A^T y)}{n^{d/4}} P\left(\tilde{\alpha}_n(\theta^* + n^{-1/4}A^T y) > \frac{\lambda_1}{4} k^2 \log c \right) dy$$
$$\leq \int_{\mathbb{R}^d} \frac{g(\theta^* + n^{-1/4}A^T y)}{n^{d/4}} \cdot \sup_{\theta \in \Lambda} P\left(\tilde{\alpha}_n(\theta) > \frac{\lambda_1}{4} k^2 \log c \right) dy$$
$$= \sup_{\theta \in \Lambda} P\left(\tilde{\alpha}_n(\theta) > \frac{\lambda_1}{4} k^2 \log c \right)$$
$$\leq \sup_{\theta \in \Lambda} P\left(\tilde{\alpha}_n(\theta) > \frac{\lambda_1}{4} k^2 \sqrt{\log c} \right).$$

We again apply the argument leading to (23), thereby establishing that this upper bound, when multiplied by m, takes the following form.

$$\exp\left[\log c \left(q - \frac{\lambda_1^2 k^4}{32 \sup_{\theta \in \Lambda} \sigma^2(\theta)} \right) + o(\log c) \right] \to 0$$

if k is chosen sufficiently large. As a consequence, both integrals in (24), when multiplied by m, converge to zero for $x > \sqrt{2r\sigma^2(\theta^*)(1+d/4)(q-d/(d+4))}$, thereby proving that $P(\hat{\alpha}(c) \leq \alpha(\theta^*) + x\sqrt{\log c/c^{1-q}} \to 1$ for such x. □

Theorem 3.3 shows that when m and n do not satisfy the optimal trade-off condition described by (7), then $\hat{\alpha}(c)$ converges to $\alpha(\theta^*)$ at a rate slower than $c^{-2/(d+4)}$. In addition, the result describes the precise limit distributions that appear in this setting.

Note that both Theorems 3.2 and 3.3 imply that $\hat{\alpha}(c)$ is consistent as an estimator of $\alpha(\theta^*)$, a conclusion that is in agreement with our consistency results of Section 2 (since it is evident that when m and n are powers of c that $\log(m)/n$ necessarily converges to zero).

4. CONCLUSION

As indicated in the introduction, the main contribution of this article is the development of large-sample limit theory for simple (nonadaptive) simulation-based random search for the optimizer of an objective function that can be expressed as an expectation. This large-sample theory provides insight into the rates of convergence that arise as a consequence of the trade-off between exploration (m) and estimation (n), as well as the associated limit distributions that describe the random error associated with such search methods. From a practical standpoint, it should be noted that (by far) the most important such limit distribution is that arising in the context of Theorem 3.2. In particular, observe that when a simulator sets the values of m and n, the presence of the parameter r always permits one to fit such a (m, n) combination into the setting of Theorem 3.2 (by, for example, putting $c = mn$ and $r = \sqrt{m/n}c^{(4-d)/(4+d)}$. Thus, the limit distribution of Theorem 3.2 can, in principle, be used to develop large-sample confidence intervals for the class of simple random search algorithms described in this article. However, one is then faced with the difficulty of needing to estimate, from the sample, the ratio of $g(\theta^*)$ to $\sqrt{|\det H(\theta^*)|}$. While $g(\theta^*)$ can be easily estimated from the simulated data (via g evaluated at the sample minimizer), $|\det H(\theta^*)|$ is much more challenging to estimate. In particular, since our method does not take advantage of estimated derivatives, computing an approximation to $H(\theta^*)$ would be difficult.

ACKNOWLEDGMENTS

The authors would like to thank the Editors-in-Chief, the Associate Editor, and the referees for their careful reading of this article and for their suggestions. This led to a significant improvement in both the exposition and the quality of the contribution. In particular, we thank the former Editor-in-Chief for catching an error in the statement of Theorem 3.3, and improving the proof.

REFERENCES

Archetti, F., Betrò, B., and Steffè, S. 1977. A theoretical framework for global optimization via random sampling. Tech. rep., 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. *J. Amer. Statis. Assoc. 76,* 374, 467–469.

Devroye, L. 1978. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Trans. Inf. Theory IT-24,* 2, 142–151.

Edwards, C. H. 1995. *Advanced Calculus of Several Variables*. Dover Publications.

Ensor, K. B. and Glynn, P. W. 1997. Stochastic optimization via grid search. Lectures in Applied Mathematics, Mathematics of Stochastic Manufacturing Systems. American Mathematical Society, G. G. Yin and Q. Zhang Eds., vol. 33, 89–100.

Feller, W. 1968. *An Introduction to Probability and its Applications*. Vol. 1. Wiley.

Feller, W. 1971. *An Introduction to Probability and its Applications*. Vol. 2. Wiley.

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

Kiefer, J. and Wolfowitz, J. 1952. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat. 23,* 462.

L'Ecuyer, P. and Perron, G. 1994. On the convergence rates of IPA and FDC derviative estimators. *Oper. Res. 42,* 4, 643–656.

Petrov, V. V. 1995. *Limit Theorems of Probability Theory*. Oxford Science Publications.

Robbins, H. and Monro, S. 1951. A stochastic approximation method. *Ann. Math. Stat. 22,* 400.

Strang, G. 1986. *Linear Algebra ad its Applications*. Saunders College Publishing.

Yakowitz, S., L'Ecuyer, P., and Vázquez-Abad, F. 2000. Global stochastic optimization with low-dispersion points sets. *Oper. Res. 48,* 6, 939–950.

Zhigljavsky, A. A. 1991. *Theory of Global Random Search*. Kluwer Academic Publishers.

Received November 2010; revised May 2012, March 2013; accepted March 2013