Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

A BAYESIAN APPROACH TO STOCHASTIC ROOT FINDING

Rolf Waeber
Peter I. Frazier
Shane G. Henderson

School of Operations Research and Information Engineering
Cornell University
Ithaca, NY 14850, USA

ABSTRACT

A stylized model of one-dimensional stochastic root-finding involves repeatedly querying an oracle as to whether the root lies to the left or right of a given point x. The oracle answers this question, but the received answer is incorrect with probability $1-p(x)$. A Bayesian-style algorithm for this problem that assumes knowledge of $p(\cdot)$ repeatedly updates a density giving, in some sense, one's belief about the location of the root. We demonstrate how the algorithm works, and provide some results that shed light on its performance, both when $p(\cdot)$ is constant and when $p(\cdot)$ varies with x.

1 INTRODUCTION

The deterministic *root-finding problem* aims to locate a point x^* that solves the equation $g(x^*)=0$ for some function g. One setting is where g is unknown, but an oracle will return $g(x)$ when queried at the point x. If n calls to the oracle are allowed, a pivotal decision is at which points $x_1,\ldots,x_n$ to evaluate the function g, and how to gather information about x^*, so that the final estimate $\hat{x}_n$ of x^* is accurate, i.e., $\|\hat{x}_n - x^*\|$ is small.

In the one-dimensional case that we focus on in this paper, g is assumed to be monotone on an interval containing x^*. The well-known bisection search algorithm can then locate x^*. If, in addition, g is differentiable and the oracle not only returns the value $g(x)$, but also $g'(x)$, then the Newton-Raphson algorithm provides a very efficient search method, though it is not as robust as bisection search. (See, for example, Ben-Tal and Nemirovski 2001, Ruszczyński 2006 for detailed discussions of the deterministic root-finding problem.)

In the *stochastic root-finding problem* (SRFP), the function g can only be observed with noise. Again, a critical step in solving the SRFP is to decide at which points $X_1, X_2, \ldots, X_n$ one should query the oracle and in turn observe the random sequence $Y_1(X_1), Y_2(X_2), \ldots, Y_n(X_n)$ so that a good estimate of x^* can be produced based on the n noisy function evaluations. (Here we allow the points $X_1, X_2, \ldots, X_n$ to be random since they may depend on previous function estimates.) In this noisy setting, one can construct a bisection-like algorithm based on a Bayesian viewpoint that efficiently locates x^*. This is the focus of this paper.

This algorithm sits in contrast with most existing SRFP algorithms that extend the Newton-Raphson algorithm. The class of stochastic approximation algorithms follow an iterative scheme of the form

$$X_{n+1} = X_n + a_n Y_n(X_n), \tag{1}$$

where a_n is a sequence of step lengths that decreases with n. Such recursive algorithms were first introduced in the seminal papers Robbins and Monro (1951) and Kiefer and Wolfowitz (1952). Since then an extensive literature on the SRFP has emerged. See, for example, Kushner and Yin (2003), Lai (2003), Asmussen and Glynn (2007), Broadie et al. (2011), Pasupathy and Kim (2011) and references therein for overviews and recent developments. Under certain assumptions, stochastic approximation methods generate a sequence

of points X_n that converge to x^* at an asymptotic rate that is of the order $n^{-1/2}$, i.e., $(n^{1/2}|X_n - x^*| : n \geq 1)$ is a tight sequence of random variables. Our theoretical and empirical results suggest that the probabilistic bisection-search algorithm discussed herein has a *geometric* rate of convergence, i.e., for some $c > 1$, $(c^n|X_n - x^*| : n \geq 1)$ is tight. The improvement from a polynomial to an exponential asymptotic rate of convergence is significant, but it comes about through the strong (and unrealistic) assumption that the probability of observing an incorrect sign in the outcomes $Y_n(X_n)$ is *known*. Given that we make such a strong assumption, this paper should be viewed as a call for research into practical algorithms, rather than as the introduction of a new root-finding algorithm.

Bisection search under uncertainty has been studied in the information theory and computer science literatures; see Horstein (1963), Burnashev and Zigangirov (1974), Rivest et al. (1980), Pelc (1989), Karp and Kleinberg (2007), Ben-Or and Hassidim (2008), Castro and Nowak (2008), Nowak (2008), and Nowak (2009). In this literature, the search space is usually discretized into n intervals and the goal is to determine the interval containing x^*. We are instead interested in the behavior of the (continuous) residuals $R_n := \hat{X}_n - x^*$, where $\hat{X}_n$ is the current best estimate of x^*.

The idea behind the algorithm is to sequentially update a prior density on the location of x^* according to Bayes' rule. The updated (posterior) density reflects our belief in the location of x^*. Under some assumptions on the noise model, the introduced policy, which measures at the median of the posterior, is optimal in the sense of minimizing the expected posterior entropy. The algorithm dates back to Horstein (1963), although our observation of its entropic optimality and our convergence analysis on continuous problems both appear to be new. Recently this approach has been adopted in Jedynak et al. (2011) for similar problems appearing in computer vision. Proofs of the results presented in this paper and more empirical examples are available in Waeber, Frazier, and Henderson (2011b).

2 PROBLEM STATEMENT

Let $g : (0,1) \to \mathbb{R}$ be such that there exists a unique point $x^* \in (0,1)$ with $g(x) > 0$ for all $x < x^*$ and $g(x) < 0$ for all $x > x^*$. The goal is to locate the point x^*. The function g cannot be observed directly, so we must instead learn about g via stochastic simulation where x is a control parameter of the simulator. For any $x \in (0,1)$ the simulator produces random outcomes $Z(x) = g(x) + \varepsilon(x) \in \mathbb{R}$, where $\varepsilon(x)$ represents stochastic noise. A common assumption is that $\mathbb{E}[\varepsilon(x)] = 0$. However, we assume that the median of $\varepsilon(x)$ is zero. For symmetric noise distributions, the two assumptions coincide. The assumption of zero median allows us to reformulate the problem by defining the function $r(x,x^*) := \mathbb{P}[Z(x) > 0]$, where $r(x,x^*) > 1/2$ for all $x < x^*$, and $r(x,x^*) < 1/2$ for all $x > x^*$. We include x^* as an argument to r to emphasize that the response depends on both the point queried x and the location of the root x^*. The noisy bisection algorithm we introduce in Section 3 uses only $Y(x) = \text{sign}(Z(x))$ when learning about g. In this case, the information exploited is simply whether x^* is to the left or right of x, and this "direction" may be wrong. Discarding information may seem counterproductive, because the magnitude of $Z(x)$ contains additional information about $g(x,x^*)$. As we will see, however, this makes a Bayesian update tractable, and the resulting algorithm appears to converge rapidly. Using reduced information may also provide a more robust estimator of x^* when the noise $\varepsilon(x)$ is heavy-tailed.

Define functions $p(x,x^*) := \max(r(x,x^*), 1 - r(x,x^*))$ and $q(x,x^*) := 1 - p(x,x^*)$, so that p gives the probability that the oracle provides a correct answer. By definition of the functions r and g, it follows that $p(x,x^*) > 1/2$ for $x \neq x^*$, and $p(x^*,x^*) = 1/2$. Further, we assume that after sampling at x, the value of $p(x,x^*)$ is revealed. This is unrealistic, since in practice one must estimate $p(x,x^*)$. Nevertheless, we proceed under this assumption, leaving the situation where $p(x,x^*)$ is not revealed for future research.

The goal is to select the points $\{X_1, X_2, \ldots, X_n\}$ so as to obtain an estimate $\hat{X}_n$ of x^* with small absolute error $|\hat{X}_n - x^*|$. The sampling points X_n can depend on the previously observed values X_m, $Y_m(X_m)$, and $p(X_m,x^*)$ for $m < n$ and, optionally, some other independent random source (allowing for randomized policies). A method for selecting these points is called a policy or an algorithm.

In our analysis, we consider two settings: the frequentist setting, and the Bayesian setting. In the frequentist setting, x^* is fixed but unknown. In the Bayesian setting, rather than considering a single value of x^*, we consider average case performance across a wide variety of x^*. This is done by supposing that, before sampling begins, X^* was drawn at random from a prior probability density f_0 over $(0,1)$. In the Bayesian setting, we denote the root by X^* rather than x^* to indicate that it is modeled as a random variable. Sampling then proceeds with the drawn value of X^* (we do not re-draw X^*), and our goal is to find X^*.

The prior f_0 may reflect one's initial belief regarding the location of the root. If one has no strong beliefs, then one may use a uniform (noninformative) prior $f_0(x) = \mathbb{1}\{x \in (0,1)\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function. For the empirical examples in Section 4, we use this uniform prior.

To summarize, the problem structure considered in this paper is as follows:

- In the frequentist setting, $X^* = x^* \in (0,1)$ is fixed. In the Bayesian setting, X^* is drawn once, at time $n = 0$, from the prior density f_0.
- $p : (0,1) \times (0,1) \to [1/2, 1]$ is a fixed function such that $p(X^*, X^*) = 1/2$ and $p(x, X^*) > 1/2$ for all $x \neq X^*$. In the Bayesian setting, this statement is assumed to hold f_0 almost surely, and in the frequentist setting it is assumed to hold only for $X^* = x^*$.
- There exists a stochastic simulator (oracle) that, given X_n, produces a random output $Y_n(X_n) \in \{-1, +1\}$. The distribution of $Y_n(X_n)$ is given by three cases:

$$\text{If } X_n < X^*, \qquad Y_n(X_n) = \begin{cases} +1, & \text{if } U_n < p(X_n, X^*), \\ -1, & \text{if } U_n \geq p(X_n, X^*). \end{cases}$$

$$\text{If } X_n > X^*, \qquad Y_n(X_n) = \begin{cases} +1, & \text{if } U_n \geq p(X_n, X^*), \\ -1, & \text{if } U_n < p(X_n, X^*). \end{cases}$$

$$\text{If } X_n = X^*, \qquad Y_n(X_n) = \begin{cases} +1, & \text{if } U_n \geq 1/2, \\ -1, & \text{if } U_n < 1/2. \end{cases}$$

  Here $U = (U_n : n \geq 1)$ is a sequence of iid $U(0,1)$ random variables.
- Each X_n is chosen based on the information available at time n. Formally, each X_n is a random variable satisfying $X_n \in \mathscr{F}_{n-1} = \sigma(X_m, Y_m(X_m), p(X_m, X^*), V_k : 1 \leq m \leq n-1, 0 \leq k \leq n-1)$, where $V = (V_k : k \geq 0)$ is a sequence of iid $U(0,1)$ random variables introduced to allow randomized policies. The sequences U and V are independent of each other. In the Bayesian setting, they are also independent of the random variable X^*.

Also, we devote considerable attention to "the constant p case" where $p(x, X^*)$ is equal to a constant $p \in (1/2, 1]$ for all $x \neq X^*$, due to its tractability. In the Bayesian setting, this statement is understood in the f_0 almost sure sense, and in the frequentist setting it is understood as applying to $X^* = x^*$.

3 PROBABILISTIC BISECTION ALGORITHM

Our approach to the problem described above is motivated by the bisection method. Using bisection without accounting for random noise will fail in most cases, since a single wrong answer from the oracle leads the search astray. We instead successively update a distribution giving our "belief" in the location of the root.

3.1 Bayesian Posterior

The algorithm that we consider below uses intuition derived from Bayesian analysis. In the Bayesian setting, we may calculate a posterior distribution f_n on X^* at any time n. This posterior distribution is the conditional distribution of X^* given the information available at time n. In the constant p case, the

posterior is given by the following recursive updating equations:

$$\text{If } Y=+1 \text{ and } p(X_n,X^*)\neq 1/2, \text{ then } f_{n+1}(x)\propto \begin{cases} p(X_n,X^*)f_n(x), & \text{if } x>X_n, \\ q(X_n,X^*)f_n(x), & \text{if } x\leq X_n, \end{cases} \tag{2}$$

$$\text{If } Y=-1 \text{ and } p(X_n,X^*)\neq 1/2, \text{ then } f_{n+1}(x)\propto \begin{cases} q(X_n,X^*)f_n(x), & \text{if } x>X_n, \\ p(X_n,X^*)f_n(x), & \text{if } x\leq X_n. \end{cases} \tag{3}$$

If $p(X_n,X^*)=1/2$, then $X_n=X^*$, and the posterior at time $n+1$ is a point mass at X_n. In the Bayesian setting, this last case occurs with probability 0.

Altering a density at a single point does not change the probability distribution, so we arbitrarily update f_n at the point $x=X_n$ the same way we update $x<X_n$. Appendix A gives the derivation of the updating equations and the proportionality constants. When p is not constant, the correct update is much more intricate, but the algorithm discussed below adopts these same updating equations in the spirit of a heuristic.

3.2 Algorithm Description

The probabilistic bisection algorithm is motivated by the following Bayesian optimality analysis. This analysis uses the entropy, which is a summary measure of how much information the density f_n contains about the root X^*. The entropy is defined for any density on $(0,1)$ by $H(f):=-\int_{[0,1]}\log_2(f(x))f(x)\,\mathrm{d}x$. Using this measure and given a fixed simulation budget $N\in\mathbb{N}$, the optimality analysis seeks a policy that minimizes the expected entropy of the posterior distribution at time N, $\mathbb{E}^{\pi^*}[H(f_N)|f_0]=\inf_{\pi\in\Pi}\mathbb{E}^{\pi}[H(f_N)]$. Here, a generic policy is denoted π, the space of all possible policies is denoted Π, and the expectation under policy π is denoted $\mathbb{E}^{\pi}$. Solving this optimization problem for a general function $p(\cdot,\cdot)$ appears to be difficult. However, the corresponding dynamic program can be solved explicitly in the constant p case.

Theorem 1 In the Bayesian setting with constant p, the policy that always measures at the median of f_n minimizes the expected entropy of f_N for any $N\in\mathbb{N}$.

See Waeber, Frazier, and Henderson (2011b) for a proof of Theorem 1. The density f_n is positive everywhere so has a unique median given by $F_n^{-1}(1/2)$, where F_n is the corresponding cumulative distribution function (cdf). Hence, for the constant p case the optimal policy sets $X_{n+1}=F_n^{-1}(1/2)$ for $n\geq 0$. Horstein (1963) introduced this algorithm for the constant p case for application in transmitting information over a noisy communication channel, but did not consider optimality results such as Theorem 1. This algorithm is referred to as "probabilistic bisection" in Castro and Nowak (2008) and elsewhere.

Theorem 1 motivates the following generalization of the probabilistic bisection algorithm to the case where $p(\cdot,\cdot)$ is not constant:

1. Choose prior density function f_0 that is positive on $(0,1)$.
2. for $n=1$ to $N-1$
   (a) Calculate the next measurement point: $X_n=F_{n-1}^{-1}(1/2)$, where F_{n-1} is the cdf of f_{n-1}.
   (b) Call the simulator at point X_n, to obtain the random variables $Y_n(X_n)\in\{-1,+1\}$ and $p(X_n,X^*)$.
   (c) If $p(X_n,X^*)=1/2$ then $X_n=X^*$ and we can stop sampling. Otherwise, continue.
   (d) Calculate the density f_{n+1} from f_n, X_n, $Y_n(X_n)$, and $p(X_n,X^*)$ using (2) and (3).
3. Return $\hat{X}_N=F_N^{-1}(1/2)$ as the estimate for x^*.

In the constant p case, the densities f_n are the posterior densities. Otherwise, the posterior is different from f_n, but we use f_n anyway as an approximation to the posterior in the spirit of a heuristic. Although the algorithm is derived using a Bayesian analysis, it can be studied in the frequentist setting. In doing so, the density f_0 used by the algorithm is simply viewed as a parameter of the algorithm.

The final estimate $\hat{X}_N$ is the median of f_N, which is optimal if we want to minimize the expected absolute error under the probability density f_N, i.e., $\mathbb{E}[|X-\hat{X}_N|]$ where X is a random variable with density f_N. In this case, each median X_n is the best estimate of X^* after $n-1$ calls to the oracle. Therefore, we can drop the cumbersome notation $\hat{X}_n$ and focus directly on the sequence $(X_n)_{n\in\mathbb{N}}$ from here onwards.

Figure 1 shows a sample path of the density f_n after $n = 0, 1, 2, 3, 50, 100$ calls to the oracle in the constant p case where $p = 0.6$ and $x^* = 0.25$. The prior density f_0 is that of a $U(0,1)$ random variable, i.e., $f_0(u) = \mathbb{1}\{u \in (0,1)\}$. The vertical lines depict x^* (dashed, green) and the current median X_n (solid, red). The piecewise constant line (solid, blue) depicts the posterior density f_n. Above every plot the (noisy) answer of the oracle is given. For time step $n = 1$ the oracle is wrong. The posterior density appears to converge to a point mass at x^*, and Theorem 2 confirms that this will happen.

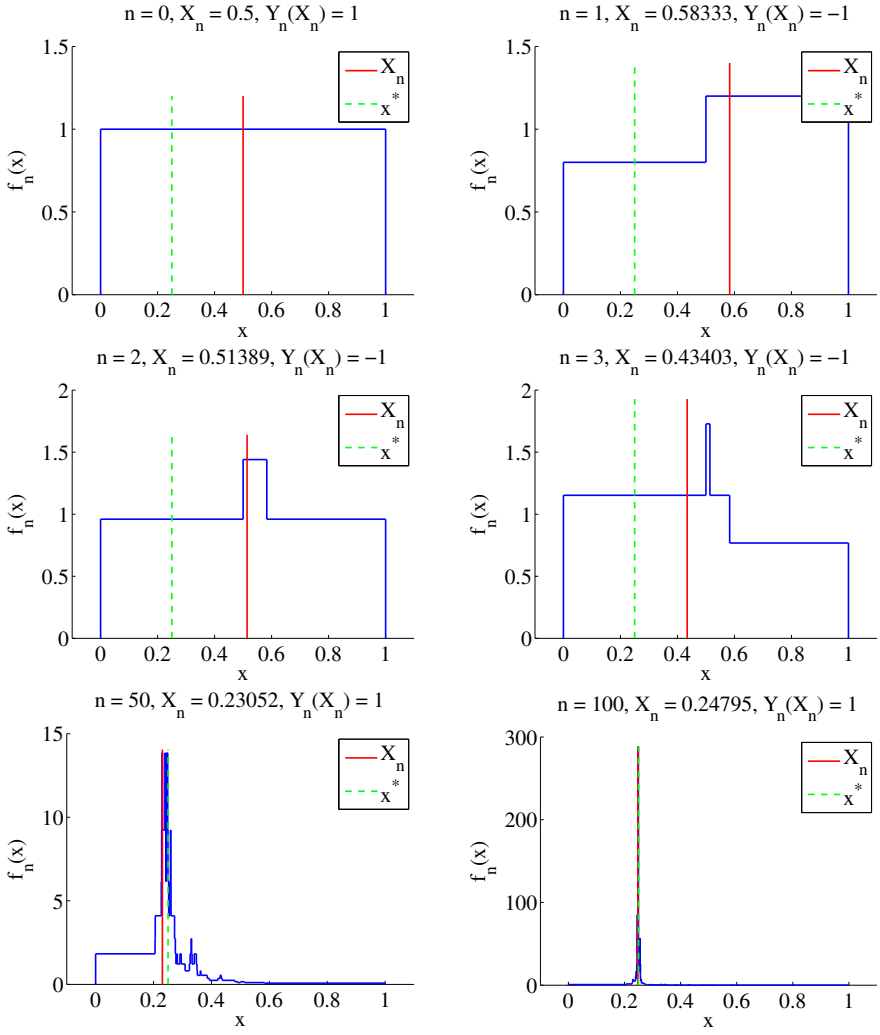

Figure 1: The density f_n at time points $n = 0, 1, 2, 3, 50, 100$ on a sample path.

3.3 Algorithm Analysis

We now turn to convergence properties of the algorithm. Roughly speaking, the first three results give conditions under which (a) the sequence of medians converges to x^* (Theorem 2), (b) the density f_n converges to that of a point mass at x^* (Theorem 3), and (c) the density f_n fails to converge to that of a point mass (Proposition 4). The remaining results relate to the rate of convergence one might expect. For proofs, see Waeber, Frazier, and Henderson (2011b).

Theorem 2 In the frequentist setting, suppose $p(\cdot, x^*)$ is bounded away from 1/2 outside any neighborhood of the point x^*. Then the sequence of medians converges to x^* almost surely, i.e., $\mathbb{P}(\lim_{n\to\infty} X_n = x^*) = 1$.

The condition on $p(\cdot,x^*)$ is quite weak, and should hold in practice. As a consequence of Theorem 2, the sequence of medians converges to x^* in the L^1-norm, i.e., $\lim_{n\to\infty}\mathbb{E}[|X_n - x^*|] = 0$, as follows from bounded convergence since $|X_n - x^*| \leq 1$ for all n. In the Bayesian setting, recall that the root X^* is a random realization from the prior density f_0. Then almost sure convergence holds, i.e., $\mathbb{P}_{f_0}(\lim_{n\to\infty} X_n = X^*) = \mathbb{E}_{f_0}(\mathbb{P}(\lim_{n\to\infty} X_n = X^*|X^*)) = 1$, and convergence in L^1 again follows from bounded convergence.

Theorem 3 In the Bayesian setting with constant $p > 1/2$, $\lim_{n\to\infty} F_n(x) = \mathbb{1}(x \geq X^*)$ for all $x \neq X^*$ almost surely.

Theorem 3 shows that, in the Bayesian setting with constant p, the posterior distribution converges to a point mass at X^* as $n \to \infty$ a.s. This immediately yields convergence in the frequentist setting where the root x^* is fixed, at least for f_0 almost all x^* in $[0,1]$, by conditioning on $X^* = x^*$. We believe that the imperfection (where the convergence has not been proven for *all* $x^* \in [0,1]$) is an artifact of the current proof rather than a property of the algorithm.

In the general (nonconstant p) case, simple and natural examples exist where the posterior distribution fails to converge to a point mass. Henceforth we often write $p(\cdot)$ for $p(\cdot,x^*)$ and $p(\cdot,X^*)$ for convenience.

Proposition 4 Consider the frequentist setting. Suppose that the algorithm begins with a uniform prior, i.e., $f_0 = \mathbb{1}\{x \in (0,1)\}$, $p(\cdot)$ is bounded away from $1/2$ outside any neighborhood of x^*, and $0 < x^* < 1/2$. Suppose further that $p(x)/x$ is non-increasing for $x \geq x^*$ with strict decrease at x^*, i.e., $p(x)/x < p(x^*)/x^* = 1/2x^*$. Then the sequence of medians converges to x^* from the right, and the posterior does not converge to a point mass.

Although Proposition 4 shows that the posterior may not converge to a point mass in the nonconstant p case, recall that Theorem 2 shows that, irrespective of whether the posterior converges to a point mass or not, the sequence of medians converges to the root. We turn next to the *rate* of convergence. Here we have only limited results, but those results suggest that the rate of convergence is exponential. Theorem 5 below provides an upper bound on the expected first hitting time of a neighborhood of x^*.

Theorem 5 In the frequentist setting, let $\delta > 0$ be such that $A = [x^* - \delta, x^* + \delta] \subseteq (0,1)$. Assume $p(\cdot) = p > 1/2$ outside A. Let $\tau(A) = \inf\{n \geq 0 : X_n \in A\}$, be the first time that the sequence of medians hits the set A. Then, starting from the uniform prior on $(0,1)$,

$$\mathbb{E}[\tau(A)] \leq -\log_2(2\delta)/r(p), \tag{4}$$

where $r(p) = p\log_2(2p) + (1-p)\log_2(2(1-p))$.

Theorem 5 establishes that the sequence of medians hits a neighborhood of the root in expected time that is logarithmic in the size of the neighborhood when $p(\cdot)$ is constant. When $p = 1$ the result reduces to the usual time required for deterministic bisection to return an interval of width 2δ.

The next result shows that this exponential rate of convergence persists on sample paths, at least for the minimum distance to the root seen to date.

Theorem 6 Consider the frequentist setting with constant p. Define $M_n := \min_{m=0,1,\ldots,n} |X_m - x^*|$, the minimum distance to x^* seen by time n. Then $c^n M_n \to 0$ a.s. as $n \to \infty$, for any nonnegative constant $c < 2^{r(p)} = 2p^p(1-p)^{1-p}$.

The upper bound on c in Theorem 6 is strictly greater than 1, so the closest median to date approaches x^* at an exponential rate. We have not been able to establish that the full sequence $(X_n : n \geq 0)$ of medians shares this property, although empirical results in Section 4 certainly suggest that this is the case. Such a fast rate of convergence (for an admittedly stylized setting where $p(\cdot)$ is constant and known) is enticing, since the best known rate of convergence in the setting of stochastic root-finding is $n^{-1/2}$ rather than c^{-n}.

Rate-of-convergence results for non-constant $p(\cdot)$ appear to be more difficult to establish. We turn next to empirical results to gain some insight.

4 EMPIRICAL RESULTS

The relative performance of different algorithms depends on both the specific problem at hand and the simulation analyst's risk tolerance (see Waeber et al. 2011a for a detailed discussion in the setting of discrete simulation-optimization problems). Therefore, measuring the performance of a stochastic root-finding algorithm is a challenging task.

To analyze the performance of the probabilistic bisection algorithm, we focus on the behavior of the residuals $R_n = X_n - x^*$. More specifically, we compare the means, the 95%-quantiles, and histograms of R_n for three different cases in this section. The first case has constant p and $X^* \sim U(0,1)$. The second case has nonconstant p and $X^* \sim U(0,1)$. The third case has nonconstant p but $x^* \in (0,1)$ is fixed.

4.1 Constant p, $X^* \sim U(0,1)$

Consider the constant p case. We are interested in the rate of convergence of the sequence X_n to X^*. The right plot in Figure 2 shows a semilog plot of the estimated expected absolute residuals $\mathbb{E}|R_n|$ after n iterations of the probabilistic bisection algorithm under a uniform prior. The logarithmic scale on the y-axis means that a straight line in the plot would suggest a geometric rate of convergence, i.e., the existence of some $c > 1$ such that $c^n \mathbb{E}[|X_n - X^*|]$ is bounded in n. Here the line appears to be slightly curved upwards, so a slower-than-geometric rate of convergence under the expectation operator might hold. Interestingly however, we often observe very fast convergence for individual sample paths. For example, the left plot of Figure 2 shows four arbitrary sample paths of the absolute residuals $|X_n - X^*|$, again in a semilog plot. These figures suggest that a geometric rate of convergence might hold pathwise.

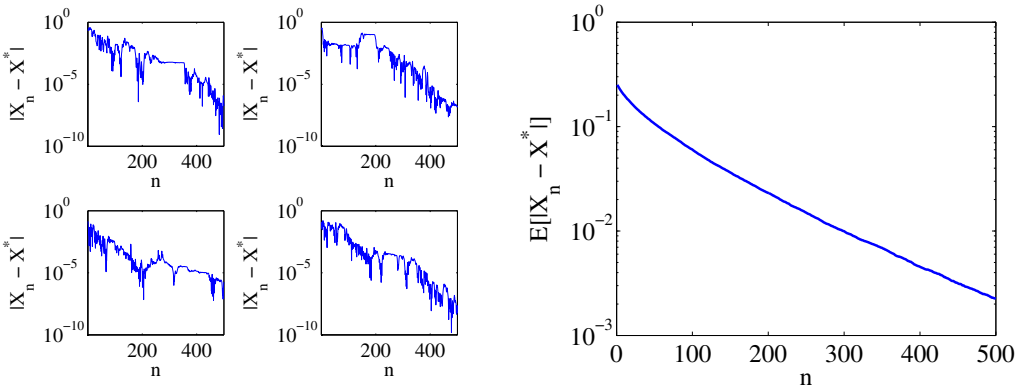

Figure 2: *Left:* Four sample paths of the absolute residuals $|R_n| = |X_n - X^*|$.
Right: Estimated mean absolute residuals, $\mathbb{E}[|X_n - X^*|]$, as a function of n with constant $p(\cdot) = p = 0.6$ under the uniform prior. The plot is the average of 10,000 runs and the maximal estimated relative error, i.e., estimated standard deviation divided by the estimated mean, over all n is less than 0.1.

We conjecture that the true rate of convergence of the expected absolute residuals is subexponential, perhaps due to the existence of sample paths that converge very slowly. The left plot in Figure 3 shows a histogram of the residuals $R_n = X_n - X^*$ after $n = 300$ iterations of the probabilistic bisection algorithm. The high peak at 0 indicates that after 300 iterations about 85% of the sample paths are very close to the point X^*. However, a few outliers are still very far away from X^*, making the overall distribution of R_n extremely heavy-tailed, and certainly not Gaussian. The right plot in Figure 3 gives the 95%-quantile of the random variable $|R_n| = |X_n - X^*|$ for various n. These quantiles show a geometric rate of convergence of the quantile to zero. So we observe an extremely fast rate of convergence for the vast majority of sample paths, but very bad performance on a few paths.

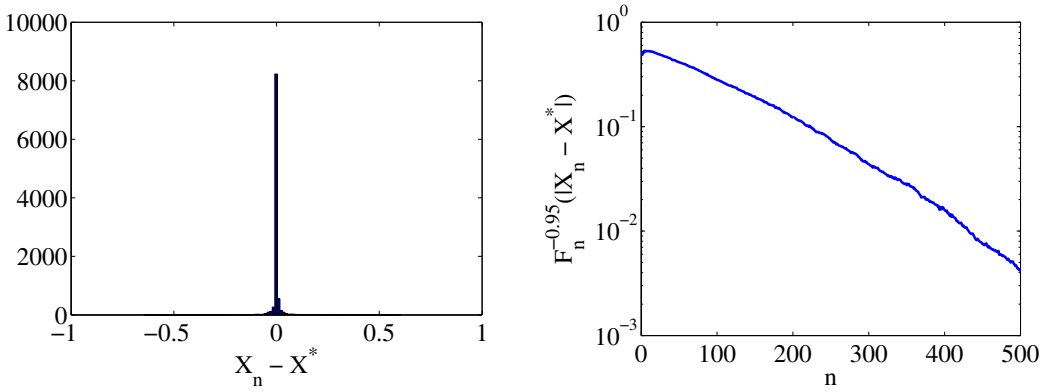

Figure 3: *Left:* The histogram of the residuals $X_n - X^*$ after $n = 300$ iterations for constant $p(\cdot) = p = 0.6$. *Right:* The estimated 95%-quantile of $|R_n| = |X_n - X^*|$ for constant $p(\cdot) = p = 0.6$. The quantiles are estimated by 10,000 simulated sample paths and the maximal estimated relative error over all n is less than 0.1. The standard deviation of the quantiles are estimated using bootstrapping.

4.2 Nonconstant p, $X^* \sim U(0,1)$

We now conduct the same analysis for the nonconstant p case. Assume that for any sequence $z_n \to x^*$ and any x^*, the function $p(z_n, x^*)$ approaches 1/2 from above. (This is a natural assumption in the setting of stochastic root-finding problems where the function g is continuous around x^*.) In particular, we consider functions $p(\cdot, x^*)$ of the form $p(x, x^*) = \min\{1/2 + \beta|x - x^*|, 1\}$ for some $\beta > 0$. In this case, the sequence of stepsizes $(|X_{n+1} - X_n| : n \geq 0)$ appears to converge to zero much faster than in the constant p case. This may cause the rate of convergence for non-constant p to be slower than for the constant p case. The rapid decrease in stepsize appears to occur for the following reason: when X_n is close to X^* then $p(X_n, X^*)$ is close to 1/2. Hence the updating equations cause very little change in the posterior density and the new median X_{n+1} will be very close to X_n. On one hand, this effect is desirable because the sequence of estimators X_n therefore automatically stabilizes as we get closer to X^*. On the other hand, it can prevent the posterior from converging to a point mass (see Proposition 4). In addition, the posterior entropy no longer necessarily converges to $-\infty$ and sampling at some point other than the median may better minimize the expected posterior entropy when X_n is close to X^*. (Recall, that sampling at the median is only optimal for the constant p case.) Nevertheless, the probabilistic bisection algorithm that samples at the median performs quite well in terms of minimizing the residuals R_n.

Although the rate of convergence for non-constant p is expected to be slower than for constant p, similar qualitative behavior are observed in both cases. Figure 4 shows the estimated mean absolute residuals, $\mathbb{E}[|X_n - X^*|]$, on semilog (left plot) and log-log (right plot) scales for $\beta = 1/2$ and $X^* \sim U(0,1)$. The semilog plot suggests that the rate of convergence is slower than geometric, while the log-log plot suggests that the rate of convergence is faster than polynomial.

Again, it is useful to look at the distribution and quantiles of the residuals. In Figure 5, the histogram of residuals shows extremely heavy tails. Furthermore, in contrast to the previous results for constant $p(\cdot)$, the plot of the 95%-quantiles of the absolute residuals $|R_n|$ does not support a geometric rate of convergence. This suggests that a geometric rate of convergence of the quantiles might not hold uniformly for all root locations $x^* \in (0,1)$, but the quantiles might still converge at a geometric rate for a fixed value $x^* \in (0,1)$, as considered in the next section.

4.3 Nonconstant p, Fixed $x^* \in (0,1)$

Now consider the behavior of the sequence X_n when $x^* \in (0,1)$ assumes a fixed value. We take x^* to be 0.2655 (an arbitrary value) and the same function $p(x, x^*) = \min\{1/2 + \beta|x - x^*|, 1\}$ with $\beta = 1/2$ as in

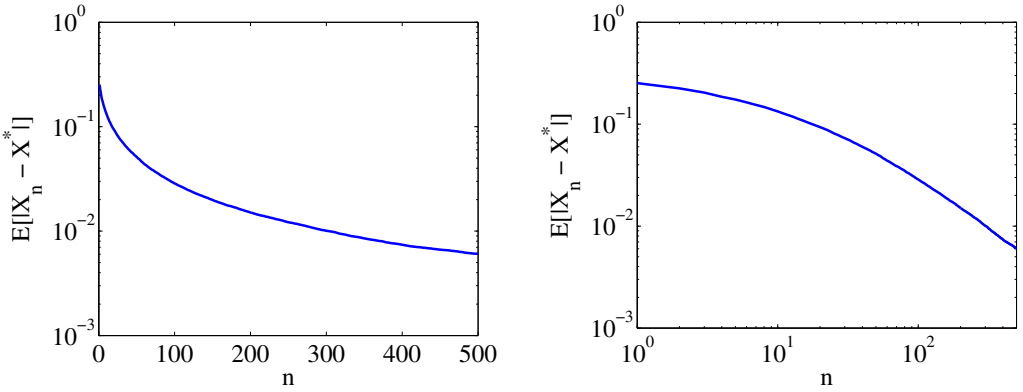

Figure 4: The estimated expected absolute residuals $\mathbb{E}[|X_n - X^*|]$ when $p(\cdot)$ varies, from 10,000 independent sample paths. The maximal estimated relative error over all n is less than 0.05.

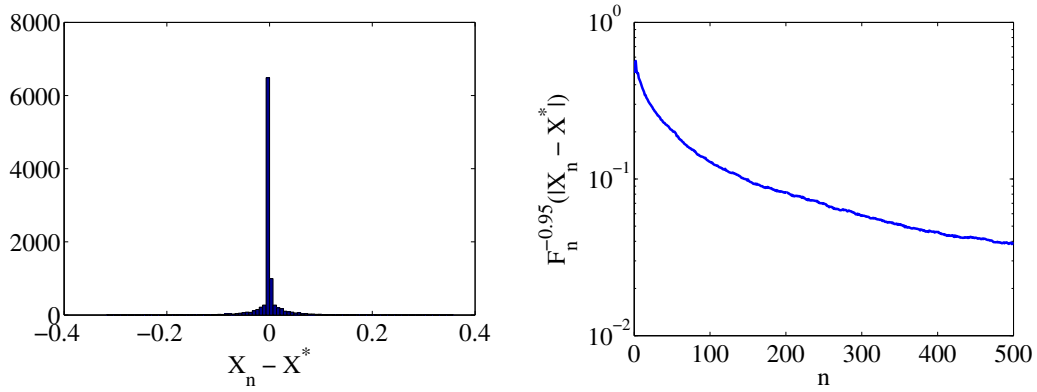

Figure 5: *Left:* Histogram of the residuals for $n = 300$. The residuals are very heavy tailed.
Right: Convergence of the 95% quantile of the absolute residuals. A geometric rate does not appear to hold. The quantiles are estimated by 10,000 simulated sample paths and the maximal estimated relative error over all n is less than 0.05. The standard deviation of the quantiles are estimated using bootstrapping.

Section 4.2. Figure 6 provides four sample paths of the absolute residuals $|R_n| = |X_n - x^*|$. The first three sample paths are chosen arbitrarily, while the fourth (bottom right) one is intentionally chosen to exhibit bad finite-time behavior. Notice that although a geometric rate of convergence might hold for most individual paths, sometimes it can take a very long time until the algorithm starts to converge to x^*. This observation is further supported by the right plot in Figure 6, which shows the estimated mean absolute residuals $\mathbb{E}[|X_n - x^*|]$. The plot suggests a slower-than-geometric rate of convergence of the expected residuals, again most likely due to the heavy-tailed behavior of the residuals shown in the histogram in Figure 7. The right plot in Figure 7 shows the convergence of the 95%-quantile for a fixed x^*. Note, that for these quantiles we do observe a geometric (or faster than geometric) rate of convergence. These results suggest that sample paths converge at a slower-than-geometric rate with very low probability, or perhaps sample paths converge at a random geometric rate, where the support of the random rate distribution includes 0.

As a side note, consistent with Proposition 4, the sequence of medians X_n approaches x^* from the right, so that all residuals $R_n = X_n - x^*$ are positive, as can be seen in the histogram in Figure 7.

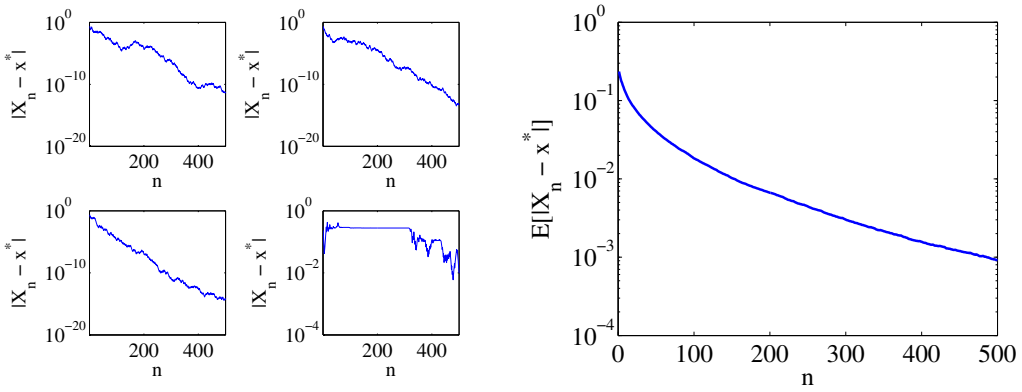

Figure 6: *Left:* Semilog plots of four sample paths of the probabilistic bisection algorithm with nonconstant p and $x^* = 0.2655$. Sometimes it takes a long time until a geometric convergence rate is observed. *Right:* Semilog plot of the estimated mean residuals. This plot is based on 10,000 independent sample paths and the maximal estimated relative error over all n is less than 0.1.

5 CONCLUSIONS AND ONGOING RESEARCH

We introduced a bisection-search method for stochastic root-finding problems based on updating a probability density function according to Bayes' rule and sampling at the median of the posterior. The algorithm provides a consistent estimator of the root x^*. Theoretical and empirical results when the probability $p(\cdot)$ is constant suggest an exponential rate of convergence. Empirical results suggest that the expected absolute residuals $\mathbb{E}[|X_n - x^*|]$ converge to zero at a rate between polynomial and exponential, and that the pathwise rate is geometric in most cases, but might be significantly slower for a small subset of sample paths. Compared to other existing stochastic root-finding methods, which typically have a rate of convergence $O(n^{-1/2})$, a geometric rate of convergence $O(c^{-n})$ for some $c > 1$ would be a tremendous improvement.

The probabilistic bisection algorithm is a novel approach to the stochastic root-finding problem and further investigation is necessary to see whether a practical variant can be developed that retains rapid convergence properties. The following is a subset of our ongoing research.

- We would like to prove or disprove the geometric rate of convergence suggested by our experiments.
- Why are the residuals $R_n = X_n - x^*$ of the medians so heavy-tailed?
- In reality, the probability of achieving the correct sign, $p(\cdot)$, is unknown. Estimating $p(X_n)$ at each X_n is therefore necessary and will inevitably slow down convergence, but to what rate is unclear.
- The policy of sampling at the median minimizes the expected posterior entropy in the constant p case. In the general case, sampling at other points, for example another quantile of F_n, might improve the rate of convergence of X_n to x^* or ensure the posterior entropy converges to $-\infty$.

ACKNOWLEDGMENTS

Shane Henderson was supported, in part, by National Science Foundation grant number CMMI-0800688. Peter Frazier and Rolf Waeber were supported, in part, by AFOSR YIP FA9550-11-1-0083.

A UPDATING EQUATIONS

Lemma 1 The following statements hold in the Bayesian setting in the constant p case, on the event $p(X_m, X^*) \neq 1/2$ for all $m \leq n$: (i) The posterior density satisfies (2) and (3). (ii) The multiplicative proportionality constant for (2) is γ^{-1}, and for (3) is $(1-\gamma)^{-1}$, where $\gamma = (1 - F_n(X_n))p(X_n, X^*) + F_n(X_n)q(X_n, X^*)$ and F_n denotes the cdf of the density f_n. (iii) If X_n is the median of f_n, then $\gamma = 1/2$ and the multiplicative constant for both (2) and (3) is 2.

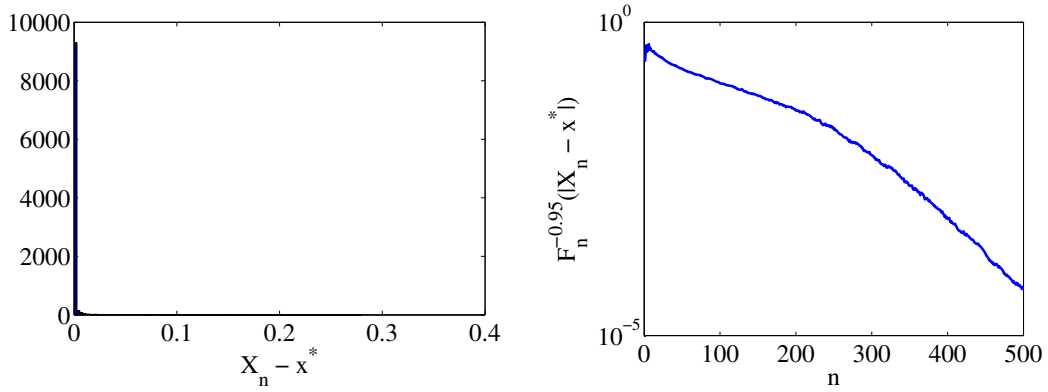

Figure 7: *Left:* The histogram of residuals for nonconstant p and fixed $x^* = 0.2655$ when $n = 300$. The sequence of medians X_n approaches the point x^* from the right, so that all the residuals are positive. *Right:* The estimated 95%-quantile of the absolute residuals as a function of n, estimated from 10,000 paths. The maximal estimated relative error over all n is less than 0.05. The standard deviation of the quantiles are estimated using bootstrapping.

Proof. Define 2 events, $A = \{Y_n(X_n) = +1, p(X_n, X^*) \neq 1/2\}$ and $B = \{Y_n(X_n) = -1, p(X_n, X^*) \neq 1/2\}$. These events have the following conditional probabilities.

$$\mathbb{P}(A|X_n < X^*, \mathscr{F}_{n-1}) = p(X_n, X^*), \quad \mathbb{P}(A|X_n > X^*, \mathscr{F}_{n-1}) = q(X_n, X^*), \quad \mathbb{P}(A|X_n = X^*, \mathscr{F}_{n-1}) = 0,$$
$$\mathbb{P}(B|X_n < X^*, \mathscr{F}_{n-1}) = q(X_n, X^*), \quad \mathbb{P}(B|X_n > X^*, \mathscr{F}_{n-1}) = p(X_n, X^*), \quad \mathbb{P}(B|X_n = X^*, \mathscr{F}_{n-1}) = 0.$$

This allows us to compute the conditional distribution of A given $\mathscr{F}_{n-1}$ as

$$\begin{aligned}\mathbb{P}(A|\mathscr{F}_{n-1}) &= \mathbb{P}(X_n < X^*|\mathscr{F}_{n-1})\mathbb{P}(A|X_n < X^*, \mathscr{F}_{n-1}) + \mathbb{P}(X_n > X^*|\mathscr{F}_{n-1})\mathbb{P}(A|X_n > X^*, \mathscr{F}_{n-1}) \\ &= (1 - F_n(X_n))p(X_n, X^*) + F_n(X_n)q(X_n, X^*) = \gamma,\end{aligned}$$

where the first equation follows from the law of total probability and $\mathbb{P}(X^* = X_n|\mathscr{F}_{n-1}) = 0$. Similarly,

$$\mathbb{P}(B|\mathscr{F}_{n-1}) = (1 - F_n(X_n))q(X_n, X^*) + F_n(X_n)p(X_n, X^*) = 1 - \gamma.$$

The result now follows from Bayes' rule. That is, on the event A, which corresponds to (2), we have

$$f_n(x) = \frac{\mathbb{P}(A|\mathscr{F}_{n-1}, X^* = x)f_{n-1}(x)}{\mathbb{P}(A|\mathscr{F}_{n-1})} = \begin{cases} \gamma^{-1}p(X_n, X^*)f_{n-1}(x), & \text{if } x > X_n, \\ \gamma^{-1}q(X_n, X^*)f_{n-1}(x), & \text{if } x < X_n, \\ 0. & \text{if } x = X_n. \end{cases}$$

For compactness, we may then alter the density at $x = X_n$ without altering the posterior distribution it implies, obtaining (2). The expression (3) for $f_n(x)$, that holds on the event B, is derived similarly. □

REFERENCES

Asmussen, S., and P. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*, Volume 57. Springer Verlag.

Ben-Or, M., and A. Hassidim. 2008. "The Bayesian learner is optimal for noisy binary search". *Annual IEEE Symposium on Foundations of Computer Science*:221–230.

Ben-Tal, A., and A. Nemirovski. 2001. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, Volume 2. SIAM.

Broadie, M., D. Cicek, and A. Zeevi. 2011. "General bounds and finite-time improvement for the Kiefer-Wolfowitz stochastic approximation algorithm". *Operations Research*:To appear.

Burnashev, M., and K. Zigangirov. 1974. "An interval estimation problem for controlled observations". *Problemy Peredachi Informatsii* 10 (3): 51–61.

Castro, R., and R. Nowak. 2008. "Active learning and sampling". In *Foundations and Applications of Sensor Management*, edited by A. O. Hero, D. A. Castañón, D. Cochran, and K. Kastella, 177–200. Springer.

Horstein, M. 1963. "Sequential transmission using noiseless feedback". *IEEE Transactions on Information Theory* 9 (3): 136–143.

Jedynak, B., P. I. Frazier, and R. Sznitman. 2011. "Questions with noise: Bayes optimal policies for entropy loss". *Journal of Applied Probability* to appear.

Karp, R., and R. Kleinberg. 2007. "Noisy binary search and its applications". In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 881–890. Society for Industrial and Applied Mathematics.

Kiefer, J., and J. Wolfowitz. 1952. "Stochastic estimation of the maximum of a regression function". *The Annals of Mathematical Statistics* 23 (3): 462–466.

Kushner, H., and G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Verlag.

Lai, T. 2003. "Stochastic approximation". *The Annals of Statistics* 31 (2): 391–406.

Nowak, R. 2008. "Generalized binary search". In *46th Annual Allerton Conference on Communication, Control, and Computing*, 568–574.

Nowak, R. 2009. "Noisy generalized binary search". In *Advances in neural information processing systems*, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Volume 22, 1366–1374.

Pasupathy, R., and S. Kim. 2011. "The stochastic root-finding problem: overview, solutions, and open questions". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 21 (3): 19.

Pelc, A. 1989. "Searching with known error probability". *Theoretical Computer Science* 63 (2): 185–202.

Rivest, R., A. Meyer, D. Kleitman, K. Winklmann, and J. Spencer. 1980. "Coping with errors in binary search procedures". *Journal of Computer and System Sciences* 20 (3): 396–404.

Robbins, H., and S. Monro. 1951. "A stochastic approximation method". *The Annals of Mathematical Statistics* 22 (3): 400–407.

Ruszczyński, A. 2006. *Nonlinear Optimization*. Princeton University Press.

Waeber, R., P. I. Frazier, and S. G. Henderson. 2011a. "A framework for selecting a selection procedure". *ACM Transactions on Modeling and Computer Simulation (TOMACS)*:under revision.

Waeber, R., P. I. Frazier, and S. G. Henderson. 2011b. "Optimal entropic bisection search in a noisy environment". Technical report, Cornell University.

AUTHOR BIOGRAPHIES

ROLF WAEBER is a Ph.D. student in the School of Operations Research and Information Engineering at Cornell University. His research interest is in simulation optimization algorithms and in quantitative risk management. His e-mail address can be found via www.orie.cornell.edu.

PETER I. FRAZIER is an assistant professor in the School of Operations Research and Information Engineering at Cornell University. He received a Ph.D. in Operations Research and Financial Engineering from Princeton University in 2009. In 2010 he received the AFOSR Young Investigator Award. His research interest is in the optimal acquisition of information, with applications in simulation, medicine and operations management. His web page can be found via www.orie.cornell.edu.

SHANE G. HENDERSON is a professor in the School of Operations Research and Information Engineering at Cornell University. He is the simulation area editor at Operations Research, and an associate editor

for Management Science. He co-edited the Proceedings of the 2007 Winter Simulation Conference. His research interests include discrete-event simulation and simulation optimization, and he has worked for some time with emergency services. His web page can be found via www.orie.cornell.edu.