Consistency of a Myopic Bayesian Algorithm for One-Dimensional Global Optimization

JAMES M. CALVIN
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, U.S.A.

(Received: 1 August 1991; accepted: 19 May 1992)

Abstract. A sequential Bayesian method for finding the maximum of a function based on myopically minimizing the expected dispersion of conditional probabilities is described. It is shown by example that an algorithm that generates a dense set of observations need not converge to the correct answer for some priors on continuous functions on the unit interval. For the Brownian motion prior the myopic algorithm is consistent; for any continuous function, the conditional probabilities converge weakly to a point mass at the true maximum.

Key words. Bayesian optimization, consistency.

Suppose we wish to locate the maximum of a real-valued function defined on a set by observing the value of the function at sequentially selected observation points. Assume that we have some prior knowledge about the relative likelihood of various functions, and that we can formalize this knowledge in the form of a probability distribution on the functions (i.e., view the function as a sample path of a stochastic process). As we sequentially observe the value of the function at various sites we can update the prior probability distribution. As the number of observations increases, one might hope that the conditional probability distribution for the maximum would become concentrated around the true maximum. In this paper we consider myopic algorithms that at each stage choose the next observation to minimize the expected variance of the posterior distribution of the maximum. Our primary purpose is to address the question of consistency of the myopic algorithm.

Optimization algorithms based on Bayesian methods and applications are surveyed in Mockus (1989), Törn and Žilinskas (1989), and Betrò (1991).

The next section introduces the problem and the notation. In Section 2 we describe the myopic optimization algorithm. The convergence properties of the myopic algorithm are investigated in Section 3, and Section 4 gives a worst-case analysis of the algorithm for Brownian motion.

1. Notation

Given a real-valued function f defined on the unit interval $[0, 1]$, let f^* denote the global maximum of the function. We consider the problem of locating f^* by

Journal of Global Optimization **3**: 223–232, 1993.

sequential observation; that is, we choose

$$t_1, t_2(f(t_1)), t_3(f(t_1), f(t_2)), \ldots,$$

where $t_n \in [0, 1]$ is the nth site at which we choose to observe the value $f(t_n)$.

Suppose we are given a probability on functions defined on the unit interval (i.e., we view $\{f(t): t \in [0, 1]\}$ as a stochastic process). By an algorithm we mean a rule for determining the sequence $\{t_k\}$. An algorithm can also be thought of as inducing a sequence of probability measures $\{P_n\}$, where

$$P_n(A) = P\{f^* \in A \mid \text{observations up to time } n\}.$$

We take the performance criterion of algorithms to be the minimization of the variance of the $\{P_n\}$.

Consider a probability space $(\Omega = C([0, 1]), \mathcal{F}, \mu)$, where $C([0, 1])$ is the set of continuous real-valued functions defined on the unit interval and $\mathcal{F}$ is the Borel σ-field of $C([0, 1])$ with the uniform topology (Billingsley, 1968). Unless otherwise noted, expectations will always be with respect to μ. For $\omega \in \Omega$ set

$$f(t) = f(t; \omega) = \omega(t),$$

and

$$f^* = f^*(\omega) = \sup\{f(t): t \in [0, 1]\}.$$

An algorithm is a rule for choosing the next site at which to observe the function based on past observations. Consider maps of the form

$$A: \Omega \to [0, 1]^{\infty},$$

where we write

$$A(\omega) = (t_1(f), t_2(f), t_3(f), \ldots).$$

Let

$$\mathcal{F}_n = \sigma(t_k, f(t_k); k = 1, 2, \ldots, n\}$$

for $n = 1, 2, \ldots$, be the σ-field generated by the observations up to time n, and

$$\mathcal{F}_\infty = \sigma\{\cup_n \mathcal{F}_n\}.$$

We can think of $\mathcal{F}_n$ as representing the information available after n observations.

We consider as algorithms the subset of maps A that satisfy the requirement that

$$t_n \in \mathcal{F}_{n-1}, \quad n = 1, 2, \ldots,$$

where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the trivial σ-field. In words, the nth site at which to observe is a function of the first $(n-1)$ observations; i.e., the algorithm is deterministic and uses only the information that has been acquired by time n.

Let P_n be a regular conditional probability distribution of f^* given $\mathcal{F}_n$ (for

existence, see Breiman (1968));

$$P_n(B) = \mu\{\omega\colon f^*(\omega) \in B \mid \mathcal{F}_n\}$$

for Borel sets B. Note that P_n is a random measure.

Let

$$M_n = \max\{f(t_1), f(t_2), \ldots, f(t_n)\}$$

be the maximum value observed by time n.

We will be mainly interested in the Brownian motion prior on $C([0, 1])$. The following formula (from Shepp (1979)) for the conditional distribution of the maximum of a Brownian motion given its value at two endpoints will be used in the following sections:

$$\mu(\sup_{0\leq s\leq t} \omega(s) > y \mid \omega(0) = 0, \omega(t) = x) = \exp\left(-\frac{2y(y-x)}{t}\right). \tag{1}$$

2. Myopic Algorithms

Let

$$V_n = \mathrm{Var}(f^* \mid \mathcal{F}_n) = E((f^*)^2 \mid \mathcal{F}_n) - E^2(f^* \mid \mathcal{F}_n).$$

THEOREM 1. *For any algorithm,* $\{(V_n, \mathcal{F}_n)\colon n = 0, 1, 2, \ldots\}$ *is a positive supermartingale.*

Proof. We need to show that

$$E(V_{n+1} \mid \mathcal{F}_n) \leq V_n.$$

Jensen's inequality implies that

$$E(E^2(f^* \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n) \geq E^2(f^* \mid \mathcal{F}_n), \tag{2}$$

and since $\mathcal{F}_n \subset \mathcal{F}_{n+1}$,

$$E(E((f^*)^2 \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n) = E((f^*)^2 \mid \mathcal{F}_n). \tag{3}$$

Using these facts

$$\begin{aligned} V_n - E(V_{n+1} \mid \mathcal{F}_n) &= E((f^*)^2 \mid \mathcal{F}_n) - E^2(f^* \mid \mathcal{F}_n) - E(E((f^*)^2 \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n) \\ &\quad + E(E^2(f^* \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n) \\ &= E(E^2(f^* \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n) - E^2(f^* \mid \mathcal{F}_n) \geq 0, \end{aligned}$$

where the second equality follows from (3) and the final inequality is a result of (2). ∎

If $\mathrm{Var}(f^* \mid f(t)) = E(f^* - E(f^* \mid f(t))^2 \mid f(t))$, then

$$\mathrm{Var}(f^*) = E(\mathrm{Var}(f^* \mid f(t))) + \mathrm{Var}(E(f^* \mid f(t))) . \tag{4}$$

Therefore the expected gain (the current variance minus the expected posterior variance) from making an observation at t is $\mathrm{Var}(E(f^* \mid f(t)))$.

COROLLARY 2. *There is a random variable V such that*

$$V_n \to V \text{ a.s. } (\mu) .$$

Proof. This follows from the supermartingale convergence theorem (Theorem II-2-9, Neveu, 1975), since $\{(V_n, \mathcal{F}_n): n = 0, 1, 2, \ldots\}$ is a positive supermartingale. ■

We are now in a position to pose the one stage (myopic) optimization problem: choose t to

$$\text{minimize } E(\mathrm{Var}(f^* \mid f(t))) , \tag{5}$$

or, equivalently,

$$\text{maximize } \mathrm{Var}(E(f^* \mid f(t))) , \tag{6}$$

where the expectations are with respect to the current probability distribution. Clearly the expected decrease in variance given by (6) is non-negative. Furthermore, if f^* is positive and $f^* \in \mathcal{F}_\infty$ then with probability one,

$$E(f^* \mid \mathcal{F}_n) \to E(f^* \mid \mathcal{F}_\infty) = f^*$$

outside $\{E(f^* \mid \mathcal{F}_n) = \infty \text{ for all } n\}$ (Corollary II-2-9 of Neveu, 1975), in which case $V_n \to 0$ with probability one. Therefore, if the prior is supported by $C([0, 1])$ and the algorithm generates a dense set of observations for any sample path, the Bayesian optimizer will convince himself that he has found the maximum. In the next section we consider the question of whether he is necessarily correct or if it is possible that he is convinced of the wrong answer.

3. Consistency of Myopic Algorithms

The following theorem shows that under the myopic algorithm for Brownian motion, the set of observation sites becomes dense.

THEOREM 3. *Let f be a continuous real-valued function on $[0, 1]$. If μ is Brownian motion, then the set of observation sites $\{t_1, t_2, \ldots\}$ generated by the myopic algorithm applied to f is dense in $[0, 1]$.*

Proof. Let μ_n^f and P_n^f be regular conditional probability distributions given $\mathcal{F}_n$ (since $C[0, 1]$ is a complete separable metric space, the existence of regular

conditional probabilities is guaranteed (Breiman, 1968)). Note that the μ_n^f and P_n^f are not random measures, since f is fixed and algorithms are deterministic.

Let $0 \leq s_1 < s_2 \leq 1$ and set $t = (s_1 + s_2)/2$. We will show a contradiction results if we assume that no observation will ever be made in the interval (s_1, s_2).

First we show that the expected gain from observing at t is bounded below by a positive number. If we observe at t, then for $A > 0$ we have by Chebyshev's inequality (all expectations are with respect to the current probability distribution) that

$$\begin{aligned}\operatorname{Var}(E(f^* \mid f(t))) &\geq A^2 P(|E(f^* \mid f(t)) - E(f^*)| \geq A) \\ &\geq A^2 P(E(f^* \mid f(t)) - E(f^*) \geq A) \geq A^2 P(f(t) - E(f^*) \geq A)\,.\end{aligned}$$

The last inequality follows from the fact that $E(f^* \mid f(t)) \geq f(t)$. Now since f is a continuous function on a compact set,

$$\sup_n \int y P_n^f(dy) \leq B$$

for some constant B. Therefore,

$$\operatorname{Var}(E(f^* \mid f(t))) \geq A^2 P(f(t) - B \geq A) > 0\,,$$

since $f(t)$ is normally distributed with variance bounded below by $(s_2 - s_1)/4$ and mean bounded below. Then for any n,

$$\begin{aligned}E(V_0) - E(V_n) &= E((V_0 - V_1) + (V_1 - V_2) + \cdots + (V_{n-1} - V_n)) \\ &\geq A^2 P(\omega(t) - B \geq A) \to \infty\,,\end{aligned}$$

contradicting the boundedness of $E(V_n)$. Therefore, any subinterval will eventually have a new observation placed in it, and the result follows. ■

We now turn to the question of consistency of algorithms. Suppose we have a prior on continuous functions and a set of observations that becomes dense in the domain. Does that imply that the conditional distribution of the maximum converges weakly to a point mass at the true maximum? That is, if a Bayesian analyst succeeds in convincing himself that he has found the answer, is he necessarily correct? The following example (derived from Example 4 in Diaconis and Freedman (1983)), shows that the answer is in general negative even for functions in the support of the prior.

EXAMPLE 1. Define sequences of functions g_n, h_n in $C([0, 1])$ as shown in Figure 1.

The function g_n takes the values 3/4 at 0, 0 at 1/4 and $1-2^{-n}$, 2^{-n} at 1, and interpolates linearly in between. The function h_n is defined similarly but takes the value 1 at 1.

Define a prior π on $C([0, 1])$ by putting mass $\frac{c}{n^3}$ on g_n and $\frac{c}{n^2}$ on h_n, where c is a

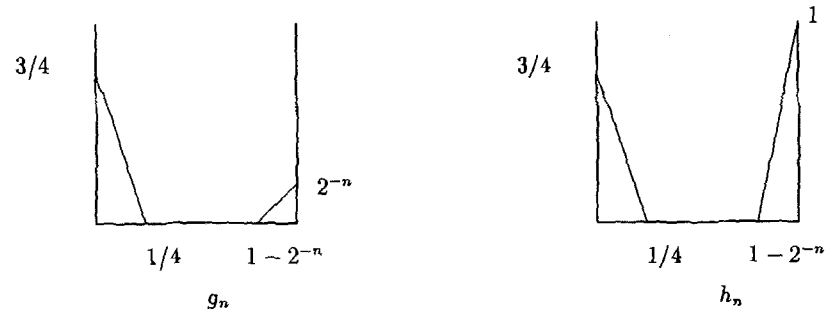

Fig. 1.

normalizing constant. Let $g_\infty = \lim_{n\to\infty} g_n$ (so g_∞ is in the support of π, the smallest closed set of probability one). Note that all functions in the support of π have a unique maximum: 3/4 for the g_n's (and g_∞), and 1 for the h_n's. Suppose we observe the values at the grid of mesh 2^{-n}; at stage n, we observe the value of the function at $\{k\cdot 2^{-n}: k=0,1,\ldots,2^n-1\}$. If the function we are observing is g_∞, then after observing the grid of mesh 2^{-n} we can rule out $\{g_k, h_k; k<n\}$, and by choice of the weights as we make more observations we become more and more convinced that the true function is one of the h_n, and therefore that the maximum is 1:

$$P_n(\{1\}) = \frac{\sum_{k=n}^{\infty} \frac{1}{k^2}}{\sum_{k=n}^{\infty} \frac{1}{k^2} + \sum_{k=n}^{\infty} \frac{1}{k^3}} \to 1,$$

and so the conditional distribution of the maximum converges weakly to a point mass at 1, while the true maximum of g_∞ is 3/4. Thus a Bayesian will convince himself of the wrong answer.

It is easy to see that the myopic algorithm would terminate successfully after the first observation, so this example does not show that the myopic algorithm is not consistent. ■

Positive results are available for the myopic algorithm corresponding to Brownian motion. First we present a technical lemma that will be used in the proof of the subsequent theorem.

LEMMA 4. *For $a>0$, consider the optimization problem: choose n and $t_1, t_2, \ldots, t_n$ to*

$$\text{minimize} \prod_{k=1}^{n} (1-\exp(-a/t_k))$$

subject to

$$\sum_{k=1}^{n} t_k = 1 \quad \text{and} \quad 0 \leq t_k \leq \Delta, \quad k=1,\ldots,n.$$

(*Take* $e^{(-a/0)} = 0$). *For* Δ *sufficiently small, a solution is*

$$t_k^* = \Delta,\ k = 1, 2, \ldots, n^* - 1;\ t_{n^*}^* = 1 - \sum_{k=1}^{n^*-1} t_k^*,$$

where $n^* = \lceil \Delta^{-1} \rceil$ *is the smallest integer greater than or equal to* Δ^{-1}.

Proof. Suppose that there is a solution $t_1, \ldots, t_n$ with two components positive and less than Δ; say $0 < t_1 < \Delta$ and $0 < t_2 < \Delta$. We will show that replacing t_1, t_2 with $t_1' = \Delta$, $t_2' = t_1 + t_2 - \Delta$ (or $t_1' = t_1 + t_2$, $t_2' = 0$ if $t_1 + t_2 < \Delta$) gives a strictly smaller value to the function to be minimized. Consider the function

$$\varphi(t) = (1 - e^{-\frac{a}{t}})(1 - e^{-\frac{a}{t_1 + t_2 - t}}),\quad 0 \leq t \leq t_1 + t_2,$$

which, for $t_1 + t_2$ sufficiently small ($t_1 + t_2 \leq a/2$ suffices) is concave, achieves its maximum at $\hat{t} = (t_1 + t_2)/2$, and is symmetric about $\hat{t}$. Since we want to minimize, we are best off choosing $t_1' = \Delta$ and $t_2' = t_1 + t_2 - \Delta$, which establishes the lemma for the case $t_1 + t_2 > \Delta$. If $t_1 + t_2 < \Delta$ the objective function is minimized by combining the two intervals ($t_1' = t_1 + t_2$, $t_2' = 0$), since $\varphi(t)$ is minimized at $t = 0$. ■

We are now ready for the main result on consistency.

THEOREM 5. *If the prior on* $C([0, 1])$ *is Brownian motion and* $f \in C([0, 1])$ *with* $f(0) = 0$, *then* $P_n^f \Rightarrow \delta_{f^*}$, *where* $\Rightarrow$ *denotes weak convergence of probability measures and* δ_x *is the distribution degenerate at* x.

Note that, as in Theorem 3, the P_n^f are not random measures since f is fixed and algorithms are deterministic.

Proof. Since f is continuous and the observations become dense by Theorem 3, we have $M_n \uparrow f^*$ as $n \to \infty$. We will show that for any $\epsilon > 0$,

$$P_n^f(f^* > M_n + \epsilon) \to 0.$$

By renumbering if necessary, we can assume that the n observation sites are

$$0 = t_0 < t_1 < t_2 < \cdots < t_n \leq 1.$$

Then

$$P_n^f(f^* > M_n + \epsilon) = 1 - \prod_{k=1}^{n} P_n^f(f^* \leq M_n + \epsilon),$$

where f_k^* is the supremum over $[t_{k-1}, t_k]$. Let $Y(t)$ be the maximum of a Brownian bridge of length t (a Brownian motion conditioned to be 0 at t, see Shepp (1979)), so that

$$P(Y(t) > \epsilon) = \exp(-2\epsilon^2/t),$$

where we have used (1). Let $\Delta_k = t_k - t_{k-1}$, and $\Delta = \max \Delta_k$. Then

$$P_n^f(f_k^* \leq M_n + \epsilon) \geq P(Y(\Delta_k) \leq \epsilon),$$

and therefore

$$\begin{aligned} P_n^f(f^* > M_n + \epsilon) &= 1 - \prod_{k=1}^{n} P_n^f(f_k^* \leq M_n + \epsilon) \\ &\leq 1 - \prod_{k=1}^{n} P(\hat{Y}(\Delta_k) \leq \varepsilon) = 1 - \prod_{k=1}^{n} [1 - \exp(-2\epsilon^2/\Delta_k)] \\ &\leq 1 - (1 - \exp(-2\epsilon^2/\Delta))^{[1/\Delta]+1} \end{aligned}$$

where the last inequality follows from Lemma 4. Since $\Delta \to 0$ (as the observations become dense), Δ will eventually be small enough for the Lemma to apply. As the diameter of the partition $\Delta \to 0$, the last expression converges to 0.

Now we have that M_n is non-decreasing, converges to f^*, P_n is supported on $[M_n, \infty)$ and $P_n((M_n + \epsilon, \infty)) \to 0$ as $n \to \infty$ for any $\epsilon > 0$, and so we can conclude that

$$P_n \Rightarrow \delta_{f^*},$$

as was to be shown. ■

The proof of consistency uses the formula (1) for the distribution of the maximum of a tied-down Brownian motion. Extending the result to the multivariate case would seem to require a similar formula for the maximum of a tied-down random field.

Finally, we give an example to show that the myopic algorithm need not be optimal.

EXAMPLE 2. For this example we will take the domain to be the discrete set $\{0, 1, 2, 3, 4\}$ and take the prior to a simple random walk conditioned to take the value 0 at 4.

Let Z_1, Z_2, Z_3, Z_4 be independent, identically distributed random variables with

$$P(Z_i = 1) = \frac{1}{2} = P(Z_i = -1).$$

Let $S_0 = 0$ and for $1 \leq n \leq 4$ set $S_n = Z_1 + \cdots + Z_n$. Define a stochastic process $\{X_n : 0 \leq n \leq 4\}$ with distribution

$$P(X_{n_1} = k_1, \ldots, X_{n_j} = k_j) = P(S_{n_1} = k_1, \ldots, S_{n_j} = k_j \mid S_4 = 0), \quad 0 \leq n \leq 4.$$

The myopic algorithm for locating $Y = \max X_k$ chooses $t_0 = 2$, and then if $X_2 = 0$, $t_1 = 1$ ($t_1 = 3$ gives the same value). Under the myopic algorithm, $E(V_1) = \frac{1}{8}$ and $E(V_2) = \frac{1}{12}$.

Now consider an alternative algorithm. Take $t_0 = 1$ and $t_1 = 2$ if $X_1 = 1$ and

$t_1 = 3$ if $X_1 = -1$. Under this algorithm, $E(V_1) = \frac{2}{9}$ and $E(V_2) = 0$. This algorithm is two-step optimal and is distinct from the myopic (one-step optimal) algorithm. ■

4. Worst Case Bound

From Lemma 4 we can see that the variance converges to zero most slowly for the function that is identically zero. In this section we obtain a bound on the rate of convergence.

We take the function $f(x) \equiv 0$, $0 \leq x \leq 1$ and the prior to be Brownian bridge. The myopic algorithm then produces a uniform grid of observations. That is, after $2^n - 1$ observations the sites are $k \cdot 2^{-n}$ for $k = 1, 2, \ldots, 2^n - 1$. For $n = 2^k - 1$ for some k,

$$P_n(f^* \leq y) = [P(\hat{Y}(1/n) \leq y)]^n = [1 - \exp(-2ny^2)]^n . \tag{7}$$

From this we can derive a lower bound on the performance of the myopic algorithm.

THEOREM 6. *Let Y_n be a random variable with distribution P_n; i.e., Y_n represents the maximum random variable after n observations when the observed function is $f \equiv 0$. Then*

$$\sqrt{8n \log(n)} \left(Y_n - \sqrt{\frac{\log(n)}{2n}} \right) \Rightarrow Y^* ,$$

where

$$P(Y^* \leq y) = \exp(-\exp(-y)) .$$

Proof. Using Equation 7,

$$P\left(\sqrt{8n \log(n)} \left(Y_n - \sqrt{\frac{\log(n)}{2n}} \right) \leq y \right) = P\left(Y_n \leq \frac{y}{\sqrt{8n \log(n)}} + \sqrt{\frac{\log(n)}{2n}} \right)$$

$$= \left[1 - \exp\left(-2n \left(\frac{y^2}{8n \log(n)} + \frac{\log(n)}{2n} + \frac{2y}{4n} \right) \right) \right]^n$$

$$= \left[1 - \exp\left(-\left[\frac{y^2}{4 \log(n)} + \log(n) + y \right] \right) \right]^n$$

$$\rightarrow \exp(-\exp(-y))$$

as $n \rightarrow \infty$. ■

5. Conclusions

We have formulated and analyzed a myopic algorithm for finding the maximum of an unknown function based on squared error loss. Our primary concern has been

the consistency of the algorithm. If the optimizer assumes a prior probability on continuous functions on the unit interval and follows the myopic algorithm for that prior, will he become more and more confident that he has found maximum? And if he does become convinced, is he necessarily right?

The answer to the first question is affirmative under fairly general conditions. If the prior is Brownian motion, then the answer to the second question is also affirmative. There exist priors for which an algorithm that generates a dense set of observations need not converge to the right answer. However, it is unknown if there are priors for which the myopic algorithm is inconsistent.

Acknowledgement

This research was supported by the National Science Foundation under grant DDM-9010770.

References

1. Betrò, B. (1991), Bayesian Methods in Global Optimization, *Journal of Global Optimization* **1**, 1–14.
2. Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.
3. Breiman, L. (1968), *Probability*, Addison-Wesley, Reading, Mass.
4. Diaconis, P. and D. Freedman (1983), Frequency Properties of Bayes' Rules, in G. E. P. Box, T. Leonard, and C. F. Wu (eds.), *Scientific Inference, Data Analysis, and Robustness*, Academic, New York.
5. Diaconis, P. and D. Freedman (1986), On the Consistency of Bayes Estimates, *Ann. Statist.* **14**, 1–26.
6. Mockus, J. (1989), *Bayesian Approach to Global Optimization: Theory and Applications*, Kluwer, Dordrecht.
7. Törn, A. and A. Žilinskas (1989), *Global Optimization*, Springer-Verlag, Berlin.
8. Neveu, J. (1975), *Discrete-Parameter Martingales*, North-Holland, Amsterdam.
9. Shepp, L. A. (1979), The Joint Density of the Maximum and Its Location for a Wiener Process with Drift, *J. Appl. Prob.* **16**, 423–427.