The Annals of Statistics
1991, Vol. 19, No. 1, 382–393

CHAINING VIA ANNEALING[1]

BY M. EVANS

University of Toronto

Chaining, in combination with adaptive importance sampling, can provide an effective technique for the numerical evaluation of high-dimensional integrals in the context of a posterior analysis. In many statistical problems ways of applying chaining can be found which depend heavily on the structure of the problem. In this paper we consider a very general method of implementing chaining for arbitrary integrals. Also, we show that chaining can be applied to solve global optimization problems and prove several generalizations of a theorem of Pincus.

1. Introduction. We are concerned here with the numerical evaluation of integrals

$$\int_X f(x)\mu(dx), \tag{1}$$

where X is a metric space, μ is a measure on the Borel sets and $f\colon X \to \mathbb{R}$ is Borel measurable; typically $X = \mathbb{R}^k$ and μ will be Lebesgue measure. Such problems arise in many fields and often it is very difficult to obtain reliable methods of approximation. In the statistical context intractable high-dimensional integrals are a frequent occurrence in Bayesian analyses.

A very general approach to the approximation of (1) is via importance sampling. For this we suppose that I is a probability density on X, with respect to μ, which is easy to generate from and whose support contains that of f. We then generate a sample $x_1, \ldots, x_N$ from I and approximate (1) by

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i)/I(x_i). \tag{2}$$

By the strong law of large numbers (2) converges to (1) as $N \to \infty$. If a poor choice is made for I, however, then an adequate approximation may not be obtainable within practical computation times. As dimension grows it becomes increasingly difficult to make a good choice of I as this must be based on properties of f. We note that the problem here is quite often not in choosing an I which is in some sense optimal or close to optimal for the approximation of (1), but in finding any I which gives an adequate approximation within reasonable computation times.

Received September 1988; revised September 1989.
[1]Work supported in part by grant A3120 of the Natural Sciences and Engineering Research Council of Canada.
AMS 1980 *subject classifications*. Primary 62E30, 62E25; secondary 65C05, 65D30.
Key words and phrases. Adaptive importance sampling, chaining, annealing, global optimization, Pincus' Theorem.

When our criterion is minimizing the variance of (2), the optimal choice of I is

$$|f(x)|\Big/\int|f(x)|\mu(dx); \tag{3}$$

see, for example, Rubinstein (1981), pages 122 and 123. It is rarely feasible to generate from (3) but it does suggest that we try to choose I to mimic (3) as closely as possible. In the following section we discuss a method of adaptively choosing I from a family $\mathscr{I}$, as we sample. This leads to an approximation to (1) and a choice of a member of I which best fits (3) according to a criterion which we describe. Further discussion of adaptive approaches to high-dimensional integration can be found in Lepage (1978), Friedman and Wright (1981), Smith, Skene, Shaw and Naylor (1987) and Evans (1988).

The success of adaptive importance sampling depends, in part, on the initial choice of importance sampler $I_1 \in \mathscr{I}$ and it is often difficult to do this successfully. The idea behind chaining is to start the adaptive importance sampling at an integration problem (1) for which we know a good choice $I_1 \in \mathscr{I}$. Obtaining the best-fitting $I_2 \in \mathscr{I}$ for this problem, via adaptive importance sampling, we use I_2 as the starting importance sampler for a new problem obtained by making a small change in the integrand and find the best-fitting I_3, and so on. In this way we construct a chain to the problem of interest. In Section 3 we show that any integration problem can be approached in this way and in Section 4 relate this to global optimization problems. In Section 5 we discuss some practical problems and issues associated with this technique and present an example.

2. Adaptive importance sampling by matching characteristics. To get around the problems associated with importance sampling, which we have just discussed, we work instead with a class $\mathscr{I}$ of densities with respect to μ and adaptively choose $I \in \mathscr{I} = \{I_\lambda | \lambda \in \Lambda\}$ as we sample, to agree more and more closely with (3). Of course for each element of $\mathscr{I}$ we must have available an algorithm for sampling.

To adaptively fit $I_\lambda \in \mathscr{I}$ to (3), we let $\mathbf{m}(\lambda)$ be a vector of characteristics of I_λ each of which can be computed via an expectation and let $\mathbf{m}^*$ be the corresponding vector for (3); for example, $\mathbf{m}(\lambda)$ might consist of various moments and the probability contents of various subregions of X with respect to I_λ. Typically, the dimension of $\mathbf{m}$ will correspond to that of Λ so that $\mathbf{m}$ is a 1–1 function. The fitting then proceeds as follows:

1. Select $I_{\lambda_1} \in \mathscr{I}$ to start.
2. Generate $x_1, \ldots, x_N$ from I_{λ_1} and estimate $\mathbf{m}^*$ by $\hat{\mathbf{m}}_1$ using the obvious estimates; e.g., if $X = \mathbb{R}^1$ and the mean of I_λ is an element of $\mathbf{m}(\lambda)$, then estimate the mean of (3) by

$$\frac{\sum_{i=1}^N x_i |f(x_i)|/I_\lambda(x_i)}{\sum_{i=1}^N |f(x_i)|/I_\lambda(x_i)}. \tag{4}$$

3. Find $I_\lambda \in \mathscr{I}$ which minimizes $\|\hat{\mathbf{m}}_1 - \mathbf{m}(\lambda)\|^2$ and call this I_{λ_2}.
4. Generate $x_{N+1}, \ldots, x_{2N}$ from I_{λ_2} and compute a new estimate $\hat{\mathbf{m}}_2$ of $\mathbf{m}$ by combining the samples; for example, (4) becomes

$$\frac{\sum_{i=1}^{N} x_i |f(x_i)|/I_{\lambda_1}(x_i) + \sum_{i=N+1}^{2N} x_i |f(x_i)|/I_{\lambda_2}(x_i)}{\sum_{i=1}^{N} |f(x_i)|/I_{\lambda_1}(x_i) + \sum_{i=N+1}^{2N} |f(x_i)|/I_{\lambda_2}(x_i)}. \tag{5}$$

5. Repeat step 3 using $\hat{\mathbf{m}}_2$ and continue until the process stabilizes.

Provided we have chosen $\mathscr{I}$ sensibly, we will have that $\hat{\mathbf{m}}_n \to \mathbf{m}^*$ and in practice this can be much faster than with straight importance sampling. When things are reasonably stable we can begin to estimate (1) while continuing to adaptively choose I_λ and combining our estimates as above.

For a discussion of the construction and fitting of families $\mathscr{I}$ appropriate to many statistical contexts, see Smith, Skene, Shaw and Naylor (1987) and Evans (1988). Step 3 typically requires that we have constructed a table containing at least some of the components of $\mathbf{m}(\lambda)$. The fitting is then accomplished by searching the table. The construction of these tables is generally a good investment as they can be used for many different integration problems.

Even if the family $\mathscr{I}$ is chosen appropriately; that is, a reasonable approximation to (3) exists in $\mathscr{I}$, the algorithm described above may fail if I_{λ_1} is far from the best-fitting choice. This is particularly true as dimension rises and space becomes sparser as it takes more and more function evaluations to obtain meaningful information about (3). The technique of chaining described in the next section was designed to get around this problem.

3. Chaining for integration. Chaining combines with adaptive importance sampling, as described in Section 2, to ensure that our computation does not fail due to a poor starting point. In general chaining proceeds as follows; we consider the integral (1) as a member of a class of integrals

$$\int_X f_\theta(x)\mu(dx), \tag{6}$$

where $\theta \in \Omega$ and (6) converges to (1) as $\theta \to \theta_*$, where $f_{\theta_*} = f$. Further we suppose that for at least one value $\theta_1 \in \Omega$ we know a good starting choice $I_{\lambda_1} \in \mathscr{I}$ for the adaptive importance sampling of Section 2. We start at I_{λ_1} for the integration problem given by θ_1, find the best-fitting I_{λ_2} for this problem; that is, stop when things are reasonably stable, then use I_{λ_2} as the starting value for the problem given by θ_2 where $\theta_2 \in \Omega$ is chosen to be a small step away from θ_1 towards θ_*, then find the best-fitting I_{λ_3} for this problem and continue in this fashion constructing a chain from θ_1 to θ_*. If at any stage we take too large a step, the process may fail and this is typically indicated by unrealistically small fluctuations in the estimates of the fitting quantities. In such a situation we go back one step and choose a smaller step-size. Typically, we will choose the class so that, in some topology an Ω, (6) is continuous in θ.

In Evans (1988) this approach was discussed for the posterior analysis of linear models where high-dimensional integrals arise quite naturally. In that context, and in fact in many statistical problems, the class $\{f_\theta|\theta \in \Omega\}$ suggests itself very naturally and (6) will often have a statistical interpretation and use as a marginal posterior probability or likelihood; see Evans (1988). In such cases the family $\{f_\theta|\theta \in \Omega\}$ depends heavily on the structure of the problem and in general we can say that the greater this dependence the more efficient the chaining will be. When presented with an arbitrary integration problem, it is not clear how we should proceed to choose $\{f_\theta|\theta \in \Omega\}$. We present one approach here which has obvious connections with the technique of simulated annealing; see, for example, van Laarhoven and Arts (1987).

Given that we have decided on using $\mathscr{I} = \{I_\lambda|\lambda \in \Lambda\}$ for the adaptive importance sampling, we put $\Omega = (0, \infty] \times (0, \infty] \times \Lambda$ and, assuming without loss of generality that $f(x) \geq 0$ everywhere, define

$$f_{(t,u,\lambda)}(x) = f^{1/t}(x) I_\lambda^{1/u}(x). \tag{7}$$

If f is not nonnegative everywhere, then we work with $|f|$ and estimate (1) using the best-fitting $I \in \mathscr{I}$ for $|f|$. Clearly, for fixed λ_1, when t is large enough, $f_{(t,1,\lambda_1)}$ is like I_{λ_1} in that region of X, where I_{λ_1} puts mass and it is clear that the appropriate starting value for the evaluation of (6) for (7) is I_{λ_1}. Having chosen $(t_1, 1, \lambda_1)$, we then start the adaptive importance sampling of Section 2 and find the best-fitting I_{λ_2} in $\mathscr{I}$ for $f_{(t_1,1,\lambda_1)}$. If we have chosen t_1 too large, then I_{λ_2} may well be I_{λ_1} or very close to it which indicates we should lower t and restart. If we choose t_1 too low, $f_{(t_1,1,\lambda_1)}$ may not resemble I_{λ_1} in the region, where I_λ is high so that I_{λ_1} may not be the appropriate starting value for the adaptive importance sampling and the sampling may fail to produce a meaningful choice for I_{λ_2}. Some experimentation with values of t may be necessary to start and our criterion for success is a value for I_{λ_2} different than I_{λ_1} but not wildly different.

Having obtained I_{λ_2}, we now lower t to $t_2 < t_1$ and obtain the best-fitting I_{λ_3} for $f_{(t_2,1,\lambda_1)}$ starting the adaptive importance sampling at I_{λ_2}. If t_2 is too much smaller than t_1, then the process can fail to produce a meaningful choice for I_{λ_3} and we must raise t_2 towards t_1 when this happens. We discuss the issues concerning determining failure in Section 5. We continue this process constructing a chain $(t_1, 1, \lambda_1), (t_2, 1, \lambda_1), \ldots, (t_n, 1, \lambda_1)$, where $t_n = 1$ and this concludes the first part of the chaining. As t is lowered in $f_{(t,1,\lambda_1)}$ the influence of f in the product $f_{(x)}^{1/t} I_{\lambda_1}(x)$ becomes more prominent.

For the second part of the chaining, we start at $f_{(1,u_1,\lambda_1)}$, where $u_1 > 1$ and use $I_{\lambda_{n+1}}$ as the starting importance sampler. We find the best-fitting $I_{\lambda_{n+2}}$ for $f_{(1,u_1,\lambda_1)}$ and then work with $f_{(1,u_2,\lambda_1)}$ where $u_2 > u_1$, and so on. As u increases the influence of I_{λ_1} decreases in $f_{(1,u,\lambda_1)}$ provided of course we assume that the support of I_{λ_1} contains the support of f, as $I_{\lambda_1}^{1/u}(x) \to 1$ as $u \to \infty$ when $I_{\lambda_1}(x) > 0$. At the final step we obtain $I_{\lambda_{n+n^*}}$ as the best-fitting member of $\mathscr{I}$ for $f_{(1,\infty,1)}$ and of course this is best-fitting for f as well and we can begin the computation of (1).

Let I be a probability density with respect to μ. For $t \geq 1$, let ν_t denote the probability measure with density $d\nu_t/d\mu \propto f^{1/t}I$ and for $u \geq 1$ let η_u denote the probability measure with density $d\eta_u/d\mu \propto fI^{1/u}$ provided of course these functions are μ-integrable. The following supports our fitting procedure.

THEOREM 1. (i) *If $f \geq 0$, fI is μ-integrable, $\operatorname{supp} f \subseteq \operatorname{supp} I$, $g: X \to \mathbb{R}$ is such that $E_{\nu_t}[g]$ exists for all $t \geq 1$ and $\int g(x)I(x)\mu(dx)$ exists, then $E_{\nu_t}[g] \to E_{\nu_1}[g]$ as $t \downarrow 1$.*

(ii) *If f and I are as in* (i) *and $g: X \to \mathbb{R}$ is such that $E_{\eta_u}[g]$ exists for all $u \geq 1$, then $E_{\eta_u}[g] \to E_{\eta_\infty}[g]$ as $u \to \infty$, where η_∞ is the probability measure with $d\eta_\infty/d\mu \propto f$.*

PROOF. (i) We have for $t \geq 1$ that $|g(x)f^{1/t}(x)I(x)| \leq |g(x)|I(x) + |g(x)|f(x)I(x)$ and $|f^{1/t}(x)I(x)| \leq I(x) + f(x)I(x)$. Then by the dominated convergence theorem,

$$\int g(x)\, f^{1/t}(x)I(x)\mu(dx) \to \int g(x)\, f(x)\mu(dx)$$

and

$$\int f^{1/t}(x)I(x)\mu(dx) \to \int f(x)I(x)\mu(dx) > 0 \text{ as } t \downarrow 1.$$

This gives the result.

(ii) This proceeds as in (1) with slight changes. □

We immediately have the following corollary.

COROLLARY 2. (i) $\nu_t \to_w \nu_1$ *as* $t \downarrow 1$. (ii) $\eta_u \to_w \eta_\infty$ *as* $u \to \infty$.

PROOF. If $g \in C(x)$ the class of all bounded, continuous, real-valued functions, then the hypotheses of (i) and (ii) regarding g are automatically satisfied. □

In the idealistic circumstances where we could continuously lower t and for each t exactly obtain the best-fitting $I \in \mathscr{I}$, it is clear that the process will work. The above represents a discrete approximation to the ideal and it may fail if this approximation is too coarse. We discuss this further in Section 5.

4. Chaining for global optimization. We now show how chaining with adaptive importance sampling can be used for global optimization problems. In Bayesian contexts chaining can be used for the evaluation of posterior expectations and modes.

For this we let f, I, μ and ν_t be as in Section 3 but now allow t to range in $(0, \infty)$. Also let μ_I be the measure with density $d\mu_I/d\mu = I$, $f^* = \sup\{f(x)| x \in X\}$ and put $M_f = \{x|f(x) = f^*\}$. We prove several generalizations of a result due to Pincus (1968). In effect we show that, under certain conditions,

the net $\{\nu_t | t > 0\}$ of probability measures converges weakly to a probability measure concentrated on M_f, obtaining by conditioning the probability measure with density $\propto fI$ to this set.

We first consider contexts where M_f is singleton and $f(x)$ is near f^* only when x is near this point. Let S^c denote the complement of a set S, $\bar{S}$ the closure of S, $\setminus$ denote set-theoretic difference and $B_\delta(x)$ be the open ball in X of radius δ centered at x.

THEOREM 3. *Suppose $M_f = \{x_0\}$, $f \geq 0$, is continuous at x_0, there exists relatively compact open S containing x_0 and $\varepsilon_0 > 0$ such that $f(x) < f^* - \varepsilon_0$ when $x \in S^c$ and the support of μ_I contains x_0. Then if $g\colon X \to \mathbb{R}$ is continuous at x_0 and $E_{\mu_I}[g]$ exists, then $E_{\nu_t}[g] \to g(x_0)$ as $t \to 0$.*

PROOF. The boundedness of f and the existence of $E_{\mu_I}[g]$ implies the existence of all the integrals we write below. Then

$$\left|\int g(x) f^{1/t}(x)\mu_I(dx) \Big/ \int f^{1/t}(x)\mu_I(dx) - g(x_0)\right| \leq \int |g(x) - g(x_0)| f^{1/t}(x)\mu_I(dx) \Big/ \int f^{1/t}(x)\mu_I(dx). \tag{8}$$

Since g is continuous at x_0, there is a $\delta_1 > 0$ such that $|g(x) - g(x_0)| < \varepsilon_1$ when $x \in B_{\delta_1}(x_0)$. Then (8) is bounded above by

$$\frac{\varepsilon_1 \int_{B_{\delta_1}(x_0)} f^{1/t}(x)\mu_I(dx) + \int_{B^c_{\delta_1}(x_0)} |g(x) - g(x_0)| f^{1/t}(x)\mu_I(dx)}{\int f^{1/t}(x)\mu_I(dx)} \leq \varepsilon_1 + \frac{\int_{B^c_{\delta_1}(x_0)} |g(x) - g(x_0)| f^{1/t}(x)\mu_I(x)}{\int_{B_{\delta_2}(x_0)} f^{1/t}(x)\mu_I(dx)}. \tag{9}$$

Clearly, we can choose δ_1 so that $B_{\delta_1}(x_0) \subseteq S$. Thus $\bar{S} \setminus B_{\delta_1}(x_0)$ is compact and f achieves its maximum there say $f^{**} < f^*$. Let $\varepsilon_2 = \min\{\varepsilon_0, f^* - f^{**}\}$. Then $f(x) \leq f^* - \varepsilon_2$ for every $x \notin B^c_{\delta_1}(x_0)$. Now choose δ_2 so that $f(x) > f^* - \varepsilon_2/2$ when $x \in B_{\delta_2}(x_0)$. Then (9) is bounded above by

$$\varepsilon_1 + \left[\frac{f^* - \varepsilon_2}{f^* - \varepsilon_2/2}\right]^{1/t} \frac{\int |g(x) - g(x_0)| \mu_I(dx)}{\mu_I\big(B_{\delta_2}(x_0)\big)}. \tag{10}$$

Now specify $\varepsilon > 0$ and specify $\varepsilon_1 < \varepsilon/2$. Then specify δ_1, δ_2 and ε_2 as above and finally specify δ_3 so that for $t \in (0, \delta_3)$ the second term in (10) is less than $\varepsilon/2$. Then (10) is bounded by ε and we are done. □

We have the following immediate corollary.

COROLLARY 4. *The net of probability measures $\{\nu_t | t > 0\}$ converges weakly to the probability measure degenerate at x_0 as $t \downarrow 0$.*

Theorem 3 is established in Pincus (1968) for the special case that X is a compact subset of $\mathbb{R}^n$ and μ_I is the uniform probability measure on X.

For contexts where f does not have a unique global maximum in the sense of Theorem 3 we have the following result.

THEOREM 5. *Suppose that $M_f \neq \phi$ and f and μ_I are such that there exists a regular conditional probability given f, which we denote by τ_s for $s \in \mathbb{R}$, and the support of μ_I contains M_f. Then if $g\colon X \to \mathbb{R}$ is such that $E_{\mu_I}[g]$ exists and*

$$G_g(s) = \int_X g(x)\tau_s(dx) \tag{11}$$

is continuous at $s = f^*$ then $E_{\nu_t}[g] \to E_{\tau_{f^*}}[g]$ as $t \to 0$.

PROOF. Let m denote the probability measure on $\mathbb{R}$ induced by μ_I and f. Then

$$\begin{aligned}
&\left|\int g(x)\nu_t(dx) - \int g(x)\tau_{f^*}(dx)\right| \\
&\quad = \left|\frac{\int_0^{f^*}\int g(x)\, f^{1/t}(x)\tau_s(dx)\, m(ds)}{\int f^{1/t}(x)\mu_I(dx)} - \int g(x)\tau_{f^*}(dx)\right| \\
&\quad \leq \frac{\int_0^{f^*} s^{1/t}\left|G_g(s) - G_g(f^*)\right| m(ds)}{\int_0^{f^*} s^{1/t} m(ds)} \qquad (12) \\
&\quad = \left[\int_{f^*-\delta_1}^{f^*} s^{1/t}\left|G_g(s) - G_g(f^*)\right| m(ds)\right. \\
&\qquad \left. + \int_0^{f^*-\delta_1} s^{1/t}\left|G_g(s) - G_g(f^*)\right| m(ds)\right] \Big/ \int_0^{f^*} s^{1/t} m(ds).
\end{aligned}$$

Now choose δ_1 so that $|G_g(s) - G_g(f^*)| < \varepsilon_1$ when $|s - f^*| < \delta_1$. Then choose $\delta_2 < \delta_1$ and (12) is bounded above by

$$\begin{aligned}
&\varepsilon_1 + (f^* - \delta_1)^{1/t} \frac{\int_0^{f^*-\delta_1}\left|G_g(s) - G_g(f^*)\right| m(ds)}{\int_{f^*-\delta_2}^{f^*} s^{1/t} m(ds)} \qquad (13) \\
&\quad \leq \varepsilon_1 + \left[\frac{f^* - \delta_1}{f^* - \delta_2}\right]^{1/t} \frac{\int_0^{f^*-\delta_1}\left|G_g(s) - G_g(f^*)\right| m(ds)}{m([f^* - \delta_2, f^*])}
\end{aligned}$$

and, as in Theorem 3, this can be made smaller than an arbitrary $\varepsilon > 0$ for all $t \in (0, \delta_2)$. □

We note that Theorem 5 does not imply Theorem 3. The hypotheses of Theorem 5 are clearly satisfied in many practical contexts; for example, in discrete contexts G_g can generally be extended to a continuous function on all of $\mathbb{R}$ via linear interpolation between those s values where $G_g(s)$ is given by

the trivial definition of conditional expectation. The following is an immediate consequence of the proof of Theorem 5.

COROLLARY 6. *If $G_g(s)$, as defined by* (11), *is continuous at $s = f^*$ for every $g \in C(X)$, then $\nu_t \to_w \tau_{f^*}$ as $t \downarrow 0$.*

To use adaptive importance sampling with chaining to obtain M_f, we then continue the chain $(t_1, 1, \lambda_1), (t_2, 1, \lambda_1), \ldots, (1, 1, \lambda_1)$ adding $(t_{n+1}, 1, \lambda_1)$, $(t_{n+2}, 1, \lambda_1), \ldots$, where $1 > t_{n+1} > t_{n+2} > \cdots$. If the family $\mathscr{I}$ is rich enough, the I_{λ_k} should converge to a distribution concentrated on M_f. For example if Theorem 3 applies, $X = \mathbb{R}^k$, each $I_\lambda \in \mathscr{I}$ has all its second moments and $\mathscr{I}$ is closed under rescaling, then I_{λ_k} should converge to a distribution which is degenerate at $M_f = \{x_0\}$. Often the means and variances of I_λ are included in $\mathbf{m}(\lambda)$ so it is easy to monitor when this convergence is attained; see the example of the following section. For contexts where we want to find where f attains its global minimum we work instead with $1/f$.

5. Comments on applications and an example. In many specific contexts, where we are interested in integrating or calculating global optima of a function, it is clear that algorithms will exist which are more efficient than what we have described here. On the other hand, the preceding applies very generally and it is very easy to implement. Also we could consider using this approach to obtain a good starting point for an algorithm which is more efficient when we have this kind of information; for example, using an iterative improvement algorithm in optimization problems or multiple quadrature algorithms in integration problems.

In particular, the literature on integration techniques seems devoid of recommendations on how one should proceed to compute a reasonable approximation to an arbitrary high-dimensional integral; for example, Monte Carlo approaches which rely on long computation times cannot be depended on to always work. On the other hand, chaining via annealing does provide such an approach and is applicable even when we have very little information concerning f. Further we can assess the accuracy of a given approximation by rerunning the chain inserting new steps between those of the original chain and then comparing answers. A drawback of this algorithm is that it is not amenable to contexts where we want answers quickly and with little involvement with the computation. It seems best suited to situations where we have a specific integral to compute and do not mind interacting with the program to control the schedules of choices for t and u. For example, we may be assessing the accuracy of some other approximation procedure for (1).

The performance of chaining will certainly be influenced by our choice of $\mathscr{I}$. If f is unimodal with reasonably elliptical contours, then the family $\mathscr{I}$ of k-dimensional normal distributions seems reasonable. In Evans (1988) a generalization of this family is discussed which is appropriate to various statistical contexts. Other families are certainly possible; for example, in contexts where

f is multimodal with finitely many modes then mixtures of multivariate normals would be appropriate members of $\mathscr{I}$.

As with all algorithms in this area there is no guarantee that our procedures will always work and it is a considerable problem to determine diagnostics which will reliably indicate when failure has occurred. Chaining can, however, be viewed as a technique for increasing our confidence in the accuracy of our computations. In the problems where we have used chaining we have proceeded as follows; having selected θ_i, a step in the chain, we estimate the components of $\mathbf{m}^*$ via adaptive importance sampling for several iterations and when we observe that these estimates are stable we change θ. Of course, a more automated approach is also possible based on the standard errors of these estimates; that is, specifying a formal stopping rule. The failure of a step in the chain occurs when f_{θ_i} is too unlike $f_{\theta_{i-1}}$; for example, the mass of f_{θ_i} is shifted far away from where the mass of $f_{\theta_{i-1}}$ is located. Our criterion for determining when this sort of failure has occurred is again to observe the changes in the estimates of $\mathbf{m}^*$. If these changes are much smaller than sampling variability suggests they should be, we know a failure of the above kind has occurred and we replace θ_i by something closer to θ_{i-1}. This is clearly not completely satisfactory but has worked well in a number of problems.

The approach we have presented for global optimization is an alternative to simulated annealing via the Metropolis algorithm. It is difficult to make direct comparisons but we note we have replaced the problem of an appropriate choice of Markov process and acceptance criterion by the adaptive sampling and appropriate choice of $\mathscr{I}$. In many contexts this seems much easier to do. Also our approach can be used for the evaluation of (1) which is not the case with the Markov process approach.

In many statistical problems f will be the prior times the likelihood and if the likelihood dominates, then under very general conditions f can be approximated by a multivariate normal density at least locally near its maximum. If f is exactly proportional to a multivariate normal, then so is $f^{1/t}$ and if I is a multivariate normal, then $f^{1/t}I^{1/u}$ is also exactly proportional to a multivariate normal. This is a partial justification for the use of these transformations in statistical contexts. Particularly, when being used for integration, as discussed in Evans (1988), some other class of transformations can be much more useful; for example, if $\int_X g(x)\mu(dx)$ can be exactly evaluated, then $\int_X(\alpha f(x) + (1-\alpha)g(x))\mu(dx)$ for $\alpha \in (0,1)$ may be appropriate.

A slight generalization of chaining as described in Sections 3 and 4 was found to be useful. Recall that at the ith stage the best-fitting importance sampler was denoted by $I_{\lambda_{i+1}}$. Our chain then actually consisted of

$$(t_1, 1, \lambda_1), (t_2, 1, \lambda_2), \ldots, (1, 1, \lambda_{n+1}), (1, u_1, \lambda_{n+2}), \ldots;$$

that is, we replace I_{λ_1} by I_{λ_2} at the first step and I_{λ_2} by I_{λ_3} at the next step, and so on. While not, strictly speaking, having the support of the earlier theoretical development this had the effect of speeding up the convergence. Intuitively, this is supported by the fact that at as t is lowered I_{λ_i} should be

TABLE 1
Schedule for location

Step	t	u	Number of iterations
1	20	1	2
2	15	1	2
3	10	1	2
4	8	1	2
5	6	1	2
6	5	1	2
7	4	1	3
8	3	1	3
9	2	1	3
10	1	1	4
11	0.9	1	2
12	0.8	1	2
13	0.7	1	2

concentrating its mass in regions where f is high and when u is raised I_{λ_t} is becoming more diffuse.

We now consider a somewhat artificial example to illustrate the use of the algorithm. For this we put $k = 20$, $p(x) = 1 + x^2/2$ and

$$f(\mathbf{x}) = \prod_{i=1}^{k} p(x_i - 20) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - 20)^2\right\}. \tag{14}$$

The integral of f is 3325.26 and it has a unique global maximum at $20 \cdot \mathbf{1}$. For simplicity here we take $\mathscr{I}$ to be the family of multivariate normal distributions $N_{20}(\boldsymbol{\mu}, \Sigma)$, $\mathbf{m}(\boldsymbol{\mu}, \Sigma)$ to be the vector consisting of the means and covariances and take $N = 1000$. Pretending we know nothing about f, we let the starting I be the standard normal density on $\mathbb{R}^{20}$, and note that the behaviour of f near $\mathbf{0}$ contains very little information about its global behaviour. Actually a slight modification of our algorithm improves performance here. For the first part of the chaining we restrict $\mathscr{I}$ to the subclass $\{N_{20}(\boldsymbol{\mu}, I) | \boldsymbol{\mu} \in \mathbb{R}^{20}\}$ and only fit using the means.

Table 1 gives the schedule for the first part of the chaining. The number of iterations indicates how large a sample was used at each step; for example, if m iterations were used at a step, then the step used a sample of $m \times 1000$ with the estimates of the means being updated every 1000. The schedule was generated interactively by observing changes in the means from iteration to iteration. At the end of this step all 20 of the estimates of the means lay in the interval (18.82176, 20.33771).

Table 2 gives the schedule for the global maximization. For this we also fit the variances and of course we estimate the global maximum by the mean vector. For the ith step we used the variance matrix from the $(i - 1)$st step for the first three iterations while estimating the variance matrix for the ith step and started with the identity matrix. At the end of the chain all 20 estimates of

TABLE 2
Schedule for global maximization

Step	t	u	Number of iterations
1	1	1	3
2	0.8	1	3
3	0.6	1	3
4	0.5	1	3
5	0.4	1	3

TABLE 3
Schedule for the integration

Step	t	u	Number of iterations
1	1	10^2	3
2	1	10^3	3
3	1	10^5	5
4	1	10^6	23

the means lay in the interval (19.90, 20.06) and hence we know all coordinates of the global maximum with an absolute error of ± 0.1 and a relative error of $\pm 0.5\%$. This level of accuracy was maintained for several more steps.

Table 3 gives the schedule for the integration. We fit the means and variances as described for the global maximization step. Here fairly large steps sufficed but this will not be the case in general. The estimate of (1) at the final step was 3322.03 an absolute error of 3.23 and a relative error of 0.1%. For the schedule given above no failures were observed at any of the steps. It is not hard, however, to construct schedules for this example where a failure will occur. This schedule was determined with very little experimentation. No claim is made as to its optimality or its portability to other examples.

We note that the above example shows that the convergence can be slow. The significance of the algorithm, however, lies in the fact that it provides a useful tool for attacking general integration and global optimization problems in statistics. Techniques for improving the efficiency are a subject for continuing study. We also note that this provides a common algorithmic approach to maximum likelihood and Bayesian inference methods.

REFERENCES

EVANS, M. (1988). Monte Carlo computation of marginal posterior quantiles. Technical Report 11, Dept. Statist., Univ. Toronto.

FRIEDMAN, J. and WRIGHT, M. (1981). A nested partitioning procedure for numerical multiple integration. *ACM Trans. Math. Software* **7** 76–92.

LEPAGE, G. (1978). A new algorithm for adaptive multidimensional integration. *J. Comput. Phys.* **27** 192–203.

PINCUS, M. (1968). A closed form solution of certain programming problems. *Oper. Res.* **16** 690–694.

RUBINSTEIN, R. Y. (1981). *Simulation and the Monte Carlo Method*. Wiley, New York.

SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H. and NAYLOR, J. C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician* **36** 75–82.

VAN LAARHOVEN, P. J. M. and ARTS, E. H. L. (1987). *Simulated Annealing: Theory and Applications*. Reidel, Dordrecht.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA M5S 1A1