ON THE APPROXIMATION OF CERTAIN PROBABILITY MEASURES BY A SET OF POINTS

L.A. BREYER

Abstract. In this paper we describe a framework for the comparison of a finite set of points with a probability measure if the latter belongs to a simple class. The measure of closeness chosen quantifies the degree of agreement obtained when a prescribed collection of test functions is simultaneously integrated with the respect to the given probability measure, and the set of points (identified with a set of point masses). No specific assumptions are made about the provenance of the point set, although our results have a clear application to certain Markov chain Monte Carlo integration problems.

1. Introduction

Numerical integration problems often require the sampling of some function at a selected set of point locations in space. In simple problems, the points are chosen along a regular grid, as with the technique of Riemann integration. More sophisticated methods are often needed in more complicated problems, which are often high dimensional. At this point Monte Carlo techniques are commonly applied, which requires the generation of a set of points as a stationary random sequence, often representing the path traveled by a specially designed Markov chain. The dependence between the successive points is often hard to assess, although convergence theorems guarantee that integration is successful when the set of points is sufficiently large.

In this note, we shall describe certain properties shared by *all sets of points*, regardless of the way they were generated. Our focus will be on comparing an arbitrary set of points with an arbitrary probability measure in terms of the agreement obtained when integrating a given collection of test functions. When applied to

Date: May, 2000.

1991 *Mathematics Subject Classification.* Primary 60J, Secondary 60F.

Key words and phrases. Markov chain Monte Carlo, Hypothesis Testing, Ergodic Theorem.

certain types of integration problems with a hierarchical structure (see the example in Section 4), this allows the analysis of integration algorithms which are far more complicated than those which afford a probabilistic study.

Briefly, suppose we are given a probability density on a high dimensional space. The normalizing constant may not be known, but it is often the object of at least part of the study. Assume a satisfactory approximation to the normalizing constant is available. This computation often requires a specialized method. We shall show in Theorem 3 how to utilize this constant to compare an arbitrary point set, represented by a suitably close probability density, with the target density of interest. Since the point set above is arbitrary, it might represent the output of a possibly very sophisticated algorithm. Moreover, an important benefit arising from the ability to assess the closeness to target of our point set is the ability to gauge the impact of removing an arbitrary collection of points from the set. This allows the design of point sets of minimal size, which is crucial if those points are to approximate the target economically.

The structure of this document is as follows: First, we discuss point sets as a way of approximating probability measures in a formal framework. Next, in Section 3 we define a limited class of tractable probability densities on $\mathbb{R}^d$ and show how the required properties can be used to gauge the distance to an arbitrary point set (represented by a collection of point masses). Section 4 illustrates these results with an example, where the target distribution is of an hierarchical type. The final section offers possible simple extensions to more general situations where the normalizing constant is not strictly required.

2. Approximating probability measures by point sets

Consider a probability measure $\pi(dx)$ on some measurable space E. Modern integration methods are often based upon some generalization of the classical Law of Large Numbers, which gave us Monte Carlo integration: Given a collection $\boldsymbol{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_t\}$ of points in E, where each point can be interpreted independently as the realized value of some random variable with distribution π, we can approximately integrate any function $f : E \to \mathbb{R}$ such that $\int |f| \, d\pi < \infty$, as

$$\frac{1}{t} \sum_{s=1}^{t} f(\mathfrak{x}_s) = \int f d\pi + \text{error}.$$

The error vanishes in the limit as t tends to infinity. Moreover, when $\int |f|^2 \, d\pi < \infty$, the error is $O(t^{-1/2})$, with a constant which is usually estimated through the Central Limit Theorem.

In the advanced literature, the point set $\boldsymbol{X}$ can be generated from a Markov chain, a semi Markov process, etc. Once a sufficient number of points is generated, the set $\boldsymbol{X}$ is used for all subsequent integration requirements as a replacement for μ. The number of points constituting $\boldsymbol{X}$ is chosen to satisfy several constraints, including storage limitations (generating a completely new set $\boldsymbol{X}$ for each required integral may be expensive, so $\boldsymbol{X}$ is usually saved) and convergence requirements (all functions of interest should integrate to within a small percentage error). The latter constraint is the hardest, since convergence analyses often require deep understanding of the generation method and sophisticated mathematical analyses. It is noteworthy that such analyses often ignore any features, such as smoothness or symmetry, exhibited by the integrands.

In contrast, below we shall ignore completely the provenance of a given set $\boldsymbol{X}$, but use properties of the integrand instead. Probabilistic techniques only enter in the next section. The following familiar example encapsulates all that we shall say in the rest of this section.

Example 1: Riemann Integration. Let $E = [0, 1]$ and consider the point set $\boldsymbol{X} = \{1/t, 2/t, \ldots, (t-1)/t, 1\}$. Thus the typical point is $\mathfrak{x}_s = s/t$. If f is a differentiable function, we can write

$$\int_0^1 f(x)dx = \frac{1}{t}\sum_{s=1}^{t} f(\mathfrak{x}_s) + \text{error}, \quad |\text{error}| \le \sup_{0\le y\le 1} |f'(y)| \sup_{1\le s\le t} \left|\mathfrak{x}_s - \mathfrak{x}_{s-1}\right|.$$

Ordinarily, we choose t (hence $\boldsymbol{X}$) large enough to ensure a small error for a given function f.

Suppose now that $\boldsymbol{X}$ (hence t) is fixed. For any given choice of t, there will always exist differentiable functions which give an arbitrarily large error. We ask which functions f can be integrated to within a specified accuracy. Write $\|f\| = \sup_{0\le y\le 1} |f(y)|$, and set

$$\mathcal{F} = \{f : \sup_{0\le y\le 1} |f'(y)| \le \|f\|\}.$$

For $f \in \mathcal{F}$, we have $|\text{error}| \le \epsilon \|f\|$ where $\epsilon = 1/t$, thus any function in $\mathcal{F}$ is integrated by $\boldsymbol{X}$ to within an error that is a percentage of its size. □

In the example above, it is easier to integrate a function if it varies little in between the points $\mathfrak{x}_s \in \boldsymbol{X}$. The set $\mathcal{F}$ is designed to encode such an assumption. We give now some simple definitions which will be used througout the paper.

Definition 1.

- *Let $\boldsymbol{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_t\}$ be a collection of points in E. The empirical distribution associated with $\boldsymbol{X}$ is the measure*
$$L_{\boldsymbol{X}}(dy) = \frac{1}{t}\sum_{s=1}^{t} \delta_{\mathfrak{x}_s}(dy).$$
- *Let $\mathcal{F}$ be a collection of normed functions f on E, with norm $\|\cdot\|$. Two measures μ and ν are called substitutable (for integration on $\mathcal{F}$ to within accuracy $\epsilon > 0$) if*
$$\left|\int f d\mu - \int f d\nu\right| < \epsilon \|f\| \quad \text{for all } f \in \mathcal{F}.$$
- *Given a point set $\boldsymbol{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_t\}$ and a probability density function $\dot{\pi}(x)$ on $E = \mathbb{R}^n$, the* **Approximate Integration Problem** *consists in deciding if*
$$\left|\int f dL_{\boldsymbol{X}} - \int f(x)\dot{\pi}(x)dx\right| < \epsilon \|f\| \ \ \forall f \in \mathcal{F}, \tag{1}$$
or in words: can $L_{\boldsymbol{X}}$ be substituted for $\dot{\pi}(x)dx$?

The first definition above simply establishes a correspondence between point sets and measures. The empirical distribution is easy to deal with computationally, but hard to manipulate analytically. In the second definition, if we take $\mu = L_{\boldsymbol{X}}$, we seek another distribution ν which can be substituted for $L_{\boldsymbol{X}}$ and is more tractable. Having made the substitution, we can then ask how close ν is to some specified target $\dot{\pi}(x)dx$, and attempt to answer the Approximate Integration Problem. A formal definition of tractability shall be given in the next section. Here we note simply that substitution of probability measures can potentially simplify the task when combined with the triangle inequality. More precisely, if ν is substitutable for $L_{\boldsymbol{X}}$, then we have
$$\left|\int f dL_{\boldsymbol{X}} - \int f d\pi\right| \leq \epsilon \|f\| + \left|\int f d\nu - \int f d\pi\right|,$$
and only the second term on the right need be estimated. This strategy is used in the next section. We end this section with several examples, none of which are crucial for the sequel.

Example 2: $\mathcal{F}$ too big. Let $E = \mathbb{R}$ and take $\mathcal{F}$ as the set of all bounded measurable functions, with norm $\|f\| = \sup_{x\in E} |f(x)|$. There exist no absolutely continuous probability measures ν on E which are substitutable for $L_{\boldsymbol{X}}$, for any $\epsilon < 1$. To see this by a contradiction, suppose such a measure existed, with $\nu(dx) = \pi(x)dx$ say. Then if we take $f(\mathfrak{x}_s) = 0$ for all $\mathfrak{x}_s \in \boldsymbol{X}$, and $f(x) = 1$ otherwise, we see that

$$\left|\int f dL_{\boldsymbol{X}} - \int f d\nu\right| = \left|\frac{1}{t}\sum_{s=1}^{t} f(\mathfrak{x}_s) - \int f(y)\pi(y)dy\right| = 1 = \|f\| > \epsilon \|f\|.$$

This simple example shows that if the set $\mathcal{F}$ is too large, many interesting measures may not be substitutable for $L_{\boldsymbol{X}}$ under our definition. □

Example 3: $\mathcal{F}$ too small. Let $\mathcal{F}$ be the set of constant functions on E, with norm $\|f\| = \sup_{x\in E} |f(x)|$ as before. Clearly, all probability measures ν give

$$\left|\int f dL_{\boldsymbol{X}} - \int f d\nu\right| = |\|f\| - \|f\|| = 0, \quad \text{for all } f \in \mathcal{F},$$

thus showing that ν is a substitute of $L_{\boldsymbol{X}}$ for any $\epsilon \geq 0$. This example shows that if $\mathcal{F}$ is too small, all possible measures will be substitutes of $L_{\boldsymbol{X}}$. □

Example 4: Histogram. Let $E = [0, 1]$, and take $\mathcal{F}$ to be the bounded Lipschitz functions, with supremum norm $\|f\| = \sup_{x\in E} |f(x)|$, and Lipschitz constant $\mathrm{Lip}(f) = \sup_{x,y} |f(x) - f(y)| / (\|f\| |x - y|)$. We set

$$\mathcal{F} = \Big\{ f : \|f\| < \infty \quad \text{and } \mathrm{Lip}(f) < \infty \}.$$

Let $\mathfrak{P} = \{Q_i = \epsilon[i, i + 1], i = 0, \ldots, [1/\epsilon] - 1\}$ be a partition of E into intervals of length $|Q_i| = \epsilon$, and set

$$\mathrm{hist}_{\boldsymbol{X}}(y) = \sum_{Q_i\in\mathfrak{P}} \epsilon^{-1} 1_{Q_i}(y)\big(\#\{\mathfrak{x}_s \in Q_i\}/t\big),$$

then $\nu(dy) := \mathrm{hist}_{\boldsymbol{X}}(y)dy$ is a probability distribution, and

$$\left|\int f dL_{\boldsymbol{X}} - \int f d\nu\right| = \left|\frac{1}{t}\sum_{s=1}^{t} f(\mathfrak{x}_s) - \int f(y)\mathrm{hist}_{\boldsymbol{X}}(y)dy\right| = \left|\frac{1}{t}\sum_{i}\sum_{\mathfrak{x}_s\in Q_i} f(\mathfrak{x}_s) - \sum_i \epsilon^{-1} 1_{Q_i}(y)\big(\#\{\mathfrak{x}_s \in Q_i\}/t\big)\int_{Q_i} f(y)dy\right|$$

$$
\begin{aligned}
&\leq \frac{1}{t}\sum_i \left| \sum_{\mathfrak{x}_s \in Q_i} \epsilon^{-1} \int_{Q_i} f(\mathfrak{x}_s)dy - \sum_{\mathfrak{x}_s \in Q_i} \epsilon^{-1} \int_{Q_i} f(y)dy \right| \\
&\leq \frac{1}{t}\sum_i \sum_{\mathfrak{x}_s \in Q_i} \epsilon^{-1} \int_{Q_i} |f(\mathfrak{x}_s) - f(y)|\, dy \\
&\leq \frac{1}{t}\sum_i \sum_{\mathfrak{x}_s \in Q_i} \epsilon^{-1} \|f\| \operatorname{Lip}(f) \int_{Q_i} \epsilon dy \\
&\leq \epsilon \|f\| \operatorname{Lip}(f).
\end{aligned}
$$

Here we have an example of a nontrivial absolutely continuous measure ν which is substitutable for $L_{\boldsymbol{X}}$ to within accuracy ϵ. □

Example 5: Gaussian kernel estimator. Let $\varphi_\epsilon(x) = (2\pi\epsilon)^{-p/2} \exp\bigl(-x^2/2\epsilon\bigr)$ be the Gaussian density function on $\mathbb{R}^p$. For any $\boldsymbol{X}$, the measure

$$\nu(dy) = \frac{1}{t}\sum_{s=1}^{t} \varphi_\epsilon(y - \mathfrak{x}_s)dy$$

is absolutely continuous, and if we set $\mathcal{F} = \{f : |||f||| < \infty\}$, where

$$|||f||| = \sup_x |\nabla f(x)|/(1 + |x|),$$

then a straightforward calculation using the fundamental theorem of calculus gives

$$\left| \int f dL_{\boldsymbol{X}} - \int f d\nu \right| \leq \epsilon |||f||| \Bigl(\epsilon + \frac{\alpha}{t}\sum_{s=1}^{t} |\mathfrak{x}_s| + 1\Bigr), \quad f \in \mathcal{F},$$

where $\alpha = \int |y|\, \varphi_1(y)dy$. Thus ν is substitutable for $L_{\boldsymbol{X}}$ to within an accuracy which now depends on $\boldsymbol{X}$. □

3. Normalized target densities with tractable components

In this section, we consider a restricted class of *tractable* distributions on $\mathbb{R}^d$, whose normalizing constant is known. We shall apply the concepts of the previous section to these distributions, which are formally defined below. In the next section, we explore some available options when the normalization is unknown.

Definition 2. *Let $\dot{\pi}(x^1, \ldots, x^d)$ be a* normalized *probability density on $\mathbb{R}^d$. We say that $\dot{\pi}$ has* tractable components *if for each $i = 1, \ldots, d$, the conditional density*

$$\lambda_i(x^i \,|\, x^j, j \neq i) := \dot{\pi}(x^1, \ldots, x^d) \Big/ \int \dot{\pi}(x^1, \ldots, x^d)dx^i$$

is completely known (including the normalization constant, which is dependent on $x^j, j \neq i$), and affords exact simulation.

Consider a fixed component i. We aim to discuss in this section the Approximate Integration Problem for the i-th marginal density

$$\dot{\pi}_i(x^i) := \int \cdots \int \dot{\pi}(x^1, \ldots, x^d) \prod_{j \neq i} dx^j.$$

As the integral above is potentially very difficult, we cannot (and shall not) assume that $\dot{\pi}_i(x^i)$ is known explicitly.

We state the following assumption:

Assumption (A): Let b be a nonzero even integer. For $a = 1, 2, 3, \ldots, b$ and every $x \in \mathbb{R}^d$, the function

$$\mu_i(a, x) := \frac{1}{a!} \int (x^i - z)^a \lambda_i(z \,|\, x^j, j \neq i) dz$$

is finite and exactly computable.

Our full analysis is split into several steps.

Substitution of $L_{\boldsymbol{X}}$. Take $\boldsymbol{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_t\} \subset \mathbb{R}^n$ to be a fixed point set, which we use to define the substitute probability density $\nu_{\boldsymbol{X}}(z)$ on $\mathbb{R}$ by

$$\nu_{\boldsymbol{X}}(z) := \frac{1}{t} \sum_{s=1}^{t} \lambda_i(z \,|\, \mathfrak{x}_s^j, j \neq i), \quad z \in \mathbb{R}. \tag{2}$$

Since λ_i is completely known and allows exact simulation, so does $\nu_{\boldsymbol{X}}$. Note also that $\nu_{\boldsymbol{X}}$ is already normalized. When the set $\boldsymbol{X}$ is chosen to approximate $\dot{\pi}(x^1, \ldots, x^d)$, the density $\nu_{\boldsymbol{X}}(x^i)$ can be expected to approximate the marginal $\dot{\pi}_i(x^i)$.

Next, if we are to substitute $\nu_{\boldsymbol{X}}$ for $L_{\boldsymbol{X}}$, we need an error estimate. To this end, consider a function f belonging to the set

$$\mathcal{F} := \big\{ f : \mathbb{R} \to \mathbb{R},\ f \text{ is differentiable } b \text{ times with } \|f^{(b)}\| < \infty \big\}, \tag{3}$$

where $\big\|f^{(b)}\big\| = \sup_z \big|f^{(b)}(z)\big|$. From Taylor's theorem,

$$\left| \int f(x^i) dL_{\boldsymbol{X}}(x^1, \ldots, x^d) - \int f(x^i) \nu_{\boldsymbol{X}}(x^i) dx^i \right|$$
$$\leq \left| \int \frac{1}{t} \sum_{s=1}^{t} (f(\mathfrak{x}_s^i) - f(y)) \lambda_i(y \,|\, \mathfrak{x}_s^j : j \neq i) dy \right|$$

$$\leq \left| \frac{1}{t} \sum_{s=1}^{t} \sum_{a=1}^{b-1} f^{(a)}(\mathfrak{x}_s^i) \mu_i(a, \mathfrak{x}_s) \right| + \frac{\|f^{(b)}\|}{t} \sum_{s=1}^{t} \mu_i(b, \mathfrak{x}_s).$$

We shall denote the right hand side above simply by $e_i(f, \boldsymbol{X})$; it is straightforward to compute. An application of the triangle inequality thus gives

$$\left| \int f(x^i) dL_{\boldsymbol{X}}(x^1, \ldots, x^d) - \int f(x^i) \dot{\pi}_i(x^i) dx^i \right| \leq e_i(f, \boldsymbol{X}) + \left| \int f(x^i) \big(\nu_X(x^i) - \dot{\pi}_i(x^i) \big) dx^i \right|. \tag{4}$$

We now discuss the second term on the right.

Estimating the distance to target I. Recall that $\nu_{\boldsymbol{X}}(x^i)$ is meant as an approximation of the unknown marginal density $\dot{\pi}_i(x^i)$. Similar approximations to the full target density $\dot{\pi}(x^1, \ldots, x^d)$ are also possible; among them, the following,

$$\widehat{\pi}_{\boldsymbol{X}}(x^1, \ldots, x^d) := \frac{1}{t} \sum_{s=1}^{t} \lambda_i(x^i \,|\, \mathfrak{x}_s^j, j \neq i) \prod_{\substack{k=1 \\ k \neq i}}^{d} \lambda_k(x^k \,|\, x^i; x^l, l < k; \mathfrak{x}_s^m, m > k), \tag{5}$$

deserves particular attention for the following reasons: firstly, it is explicitly computable and aready normalized, and secondly, its i-th marginal is $\nu_{\boldsymbol{X}}(x^i)$. We may therefore estimate

$$\begin{aligned} &\left| \int f(x^i) \Big[\nu_{\boldsymbol{X}}(x^i) - \dot{\pi}_i(x^i) \Big] dx^i \right| \\ &\qquad = \left| \int \cdots \int f(x^i) \Big[\widehat{\pi}_{\boldsymbol{X}}(x^1, \ldots, x^d) - \dot{\pi}(x^1, \ldots, x^d) \Big] \prod_{j=1}^{d} dx^j \right| \\ &\qquad \leq \|f\| \int \cdots \int \Big| \widehat{\pi}_{\boldsymbol{X}}(x^1, \ldots, x^d) - \dot{\pi}(x^1, \ldots, x^d) \Big| \prod_{j=1}^{d} dx^j. \end{aligned}$$

In this bound, the integrand is now explicitly calculable for any choice of configuration $x = (x^1, \ldots, x^d)$. It is also straightforward to simulate from $\widehat{\pi}$: choosing an initial configuration uniformly from $\boldsymbol{X}$, we update the components by Gibbs sampling, starting with the i-th component, and then taking the others in order.

A useful identity. The final ingredient in our analysis is based upon a useful alternative description of the L^1 distance, first reported in (Brooks et al., 1996). Specifically, for any two densities $p(x)$ and $q(x)$ on $\mathbb{R}^d$, we have

$$\int |p(x) - q(x)| \, dx = 2 - 2 \int \big(p(x) \wedge q(x) \big) dx$$

$$= 2 - 2\int \left[1 \wedge \frac{q(x)}{p(x)}\right] p(x)dx,$$

as can be easily checked on a diagram. Substituting q with $\dot{\pi}(x^1, \ldots, x^d)$ and p with $\widehat{\pi}(x^1, \ldots, x^d)$ respectively, we obtain the fundamental inequality

$$\int \left|\nu_{\boldsymbol{X}}(x^i) - \dot{\pi}_i(x^i)\right| dx^i \leq 2\int \left[1 - \frac{\dot{\pi}(x^1, \ldots, x^d)}{\widehat{\pi}_{\boldsymbol{X}}(x^1, \ldots, x^d)}\right]_+ \widehat{\pi}_{\boldsymbol{X}}(x^1, \ldots, x^d) \prod_{j=1}^{d} dx^j. \tag{6}$$

Estimating the distance to target II. At this point, it is no longer possible to continue analytically. It is obvious however that we may proceed using probabilistic methods and Monte Carlo approximation in particular. This may seem circular at first, since we propose to estimate the Monte Carlo error, which depends upon the distance from $\nu_{\boldsymbol{X}}(x^i)$ to $\dot{\pi}_i(x^i)$, by another Monte Carlo calculation, with its own inherent error. However, we are in fact in a much better position to analyse the latter.

Firstly, we know how to simulate from $\widehat{\pi}_{\boldsymbol{X}}$ exactly, which gives us a supply of independent random variables, and secondly, the integrand (in square brackets) is bounded. This latter fact is crucial, for it allows an *exact* probabilistic error analysis, originally due to Hoeffding (1963). In the following theorem, we therefore propose a probabilistic resolution of the Approximate Integration Problem.

Theorem 3. *Let $\dot{\pi}(x^1, \ldots, x^d)$ be a properly normalized probability density on $\mathbb{R}^d$ with tractable components and unknown i-th marginal $\dot{\pi}_i(x^i) = \int \dot{\pi}(x^1, \ldots, x^d) \prod_{j \neq i} dx^j$. For any $\boldsymbol{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_t\} \subset \mathbb{R}^d$, we define a distribution $\widehat{\pi}_{\boldsymbol{X}}$ by* (5) *on $\mathbb{R}^d$. Choose a finite i.i.d. sequence $Z_1, \ldots, Z_n$ from $\widehat{\pi}_{\boldsymbol{X}}$, and $\epsilon > 0$. Then*

$$\mathbb{P}\left[\int \left|\nu_{\boldsymbol{X}}(x) - \dot{\pi}_i(x)\right| dx > \epsilon + \frac{2}{n}\sum_{k=1}^{n}\left(1 - \frac{\dot{\pi}(Z_k)}{\widehat{\pi}_{\boldsymbol{X}}(Z_k)}\right)_+\right] \leq \exp(-2n\epsilon^2). \tag{7}$$

Consequently, if f is any bounded function belonging to the family $\mathcal{F}$ defined in (3)*, and provided Assumption (A) holds, we have*

$$\mathbb{P}\left[\left|\frac{1}{t}\sum_{s=1}^{t} f(\mathfrak{x}_s^i) - \int f(x)\dot{\pi}_i(x)dx\right| \right.$$
$$\left. \leq \|f\|\left(\dot{e}_i(f, \boldsymbol{X}) + 2\epsilon + \frac{2}{n}\sum_{k=1}^{n}\left(1 - \frac{\dot{\pi}(Z_k)}{\widehat{\pi}_{\boldsymbol{X}}(Z_k)}\right)_+\right)\right] \geq 1 - e^{-2\epsilon^2 n},$$

where

$$\dot{e}_i(f,\boldsymbol{X}) = \|f\|^{-1}\left[\left|(1/t)\sum_{s=1}^{t}\sum_{a=1}^{b-1} f^{(a)}(\mathfrak{x}_s^i)\mu_i(a,\mathfrak{x}_s)\right| + (\|f^{(b)}\|/t)\sum_{s=1}^{t}\mu_i(b,\mathfrak{x}_s)\right].$$

Proof. Let $Y_k = \big(1-\dot{\pi}(Z_k)/\widehat{\pi}_{\boldsymbol{X}}(Z_k)\big)_+$; since these variables are IID and bounded, Hoeffding's (1963) estimate applies, to the effect that

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} Y_k - \mathbb{E}Y_1 > t\right] \leq e^{-2t^2 n}.$$

Combining this with (6) gives the claimed results. □

Remark 1. The one sided estimate (7) of the L^1 distance in Theorem 3 can be easily turned into a two sided estimate, which would give both upper and lower confidence bounds on the distance from $\widehat{\pi}_{\boldsymbol{X}}$ to the target $\dot{\pi}$.

Remark 2. Suppose that the target $\dot{\pi}$ is not properly normalized, in the sense that only a funcion $\pi(x) \propto \dot{\pi}(x)$ is available. If the normalizing constant can be estimated from above, i.e. there exists a known constant $a \geq \int \pi(x)dx$, we obtain straightaway the conservative estimate

$$\mathbb{P}\left[\int |\nu_{\boldsymbol{X}}(x) - \dot{\pi}_i(x)|\, dx > \epsilon + \frac{2}{n}\sum_{k=1}^{n}\left(1 - \frac{a^{-1}\pi(Z_k)}{\widehat{\pi}_{\boldsymbol{X}}(Z_k)}\right)_+\right] \leq \exp(-2n\epsilon^2). \tag{8}$$

Note however that this estimate does not tend to zero. As $\widehat{\pi}_{\boldsymbol{X}}$ approximates $\dot{\pi}$, the bound above converges to the value $2\big(1 - a^{-1}\int \pi(x)dx\big)$, which is only small when the bounding constant a is sharp.

An upper bound a on the normalizing constant will be obtained in the example in the next section, for which the following assumption holds:

Assumption (B): The target distribution $\dot{\pi}(x)$ is proportional to a function

$$\pi(x^1,\ldots,x^d) = \exp\big(\sum_{k=1}^{N}\varphi_k(x^1,\ldots,x^d)\big)m(x^1,\ldots,x^d),$$

where for each k, $\sup_z \varphi_k(z) \equiv \bar{\varphi}_k < \infty$ is bounded above, and $m(x^1,\ldots,x^d)$ is a probability density which affords exact simulation. ++++++++++++

Remark 3. In Theorem 3, the boundedness assumption on f is required to be able to use the L^1 distance estimate. When f is unbounded but integrable, it must in fact be *effectively bounded*, that is there exists a set of the form $C = \{f \leq M\}$ which contributes overwhelmingly to the integral. Call M an effective bound, writing $\|f\|_{\text{eff}} = M$. The original unbounded integrand f can then be considered

bounded for the purposes of Theorem 3, with a small loss in accuracy. We do not pursue the details here, save to note that a plot of $\nu_{\boldsymbol{X}}$ together with an analysis of the tail behaviour of $\dot{\pi}_i(x^i)$ may help decide on a suitable constant.

Remark 4. The computational cost of the test is clearly at most proportional to the size t of the point set $\boldsymbol{X}$ and thus scales well. For example, to get an estimate of the L^1 distance accurate from above to within 0.1, with an error probability of 0.01 requires from (7) the generation of $n = 922$ test random variables Z_k, each of which is generated at a cost proportional to t, followed by the evaluation of the sample average, again at a cost proportional to t. Subsequently calculating the substitution error $\dot{e}(f, \boldsymbol{X})$ for any function of interest also has a cost proportional to t.

4. Example

In this section, we discuss the 'Rats' hierarchical model of Gelfand et al. (1990), which may also be found in the BUGS software distribution as example 1. Setting $N = 30$ and $T = 5$, the target distribution factorizes as

$$\begin{aligned}(9)\quad \dot{\pi}(\alpha_1,\dots,\alpha_N,\beta_1,\dots,\beta_N,\tau_\alpha,\tau_\beta,\tau_c,\alpha_c,\beta_c) &= C^{-1}\prod_{i=1}^{N}\prod_{j=1}^{T}\dot{\pi}(y_{ij}\,|\,x_j,\alpha_i,\beta_i,\tau_c)\\ &\quad\times\dot{\pi}(\alpha_i\,|\,\alpha_c,\tau_\alpha)\dot{\pi}(\beta_i\,|\,\beta_c,\tau_\beta)\dot{\pi}(\tau_c)\dot{\pi}(\alpha_c)\dot{\pi}(\tau_\alpha)\dot{\pi}(\beta_c)\dot{\pi}(\tau_\beta),\end{aligned}$$

where

$$\dot{\pi}(y_{ij}\,|\,x_j,\alpha_i,\beta_i,\tau_c) = \sqrt{\frac{\tau_c}{2\pi}}\exp-\frac{\tau_c}{2}\big(y_{ij}-\alpha_i-\beta_i(x_j-\bar{x})\big)^2,$$

$$\dot{\pi}(\alpha_i\,|\,\alpha_c,\tau_\alpha) = \sqrt{\frac{\tau_\alpha}{2\pi}}\exp-\frac{\tau_\alpha}{2}(\alpha_i-\alpha_c)^2,\qquad \dot{\pi}(\beta_i\,|\,\beta_c,\tau_\beta) = \sqrt{\frac{\tau_\beta}{2\pi}}\exp-\frac{\tau_\beta}{2}(\beta_i-\beta_c)^2,$$

$$\dot{\pi}(\tau_c) = \Gamma(a)^{-1}b^a\tau_c^{a-1}e^{-b\tau_c}1_{(\tau_c>0)},\qquad \dot{\pi}(\tau_\alpha) = \Gamma(a)^{-1}b^a\tau_\alpha^{a-1}e^{-b\tau_\alpha}1_{(\tau_\alpha>0)},$$

$$\dot{\pi}(\tau_\beta) = \Gamma(a)^{-1}b^a\tau_\beta^{a-1}e^{-b\tau_\beta}1_{(\tau_\beta>0)},\qquad \dot{\pi}(\alpha_c) = \sqrt{\frac{\delta}{2\pi}}\exp-\frac{\delta}{2}\alpha_c^2,$$

$$\dot{\pi}(\beta_c) = \sqrt{\frac{\delta}{2\pi}}\exp-\frac{\delta}{2}\beta_c^2,$$

and $a = b = 0.001$, $\delta = 0.000001$, $\bar{x} = (x_1 + \cdots + x_T)/T$. The variables y_{ij} and x_j are constants which characterize the normalization factor C.

Conditional distributions. From (9), we can also read off the full conditionals, viz.

$$\lambda_{\alpha_i}(z \,|\, \alpha_c, \tau_\alpha, \tau_c) = \sqrt{\frac{\sigma_\alpha}{2\pi}} \exp -\frac{\sigma_\alpha}{2}(z - \mu_{\alpha,i})^2,$$

where $\sigma_\alpha = \tau_\alpha + T\tau_c$ and $\mu_{\alpha,i} = \Big(\alpha_c\tau_\alpha + \tau_c \sum_{j=1}^T y_{ij}\Big)/\sigma_\alpha$,

$$\lambda_{\beta_i}(z \,|\, \beta_c, \tau_\beta, \tau_c) = \sqrt{\frac{\sigma_\beta}{2\pi}} \exp -\frac{\sigma_\beta}{2}(z - \mu_{\beta,i})^2,$$

where $\sigma_\beta = \tau_\beta + \tau_c \sum_{j=1}^T (x_j - \bar{x})^2$ and $\mu_{\beta,i} = \Big(\beta_c\tau_\beta + \tau_c \sum_{j=1}^T y_{ij}(x_j - \bar{x})\Big)/\sigma_\beta$,

$$\lambda_{\tau_\alpha}(z \,|\, \alpha_1, \ldots, \alpha_N, \alpha_c) = \Gamma(\tfrac{N}{2} + a)^{-1} b_\alpha^{N/2+a} z^{N/2+a-1} e^{-b_\alpha z} 1_{(z>0)},$$

where $b_\alpha = b + \frac{1}{2}\sum_{i=1}^N (\alpha_i - \alpha_c)^2$,

$$\lambda_{\tau_\beta}(z \,|\, \beta_1, \ldots, \beta_N, \beta_c) = \Gamma(\tfrac{N}{2} + a)^{-1} b_\beta^{N/2+a} z^{N/2+a-1} e^{-b_\beta z} 1_{(z>0)},$$

where $b_\beta = b + \frac{1}{2}\sum_{i=1}^N (\beta_i - \beta_c)^2$,

$$\lambda_{\tau_c}(z \,|\, \alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_N) = \Gamma(\tfrac{NT}{2} + a)^{-1} b_c^{NT/2+a} z^{NT/2+a-1} e^{-b_c z} 1_{(z>0)},$$

where $b_c = b + \frac{1}{2}\sum_{j=1}^T \sum_{i=1}^N \big(y_{ij} - \alpha_i - \beta_i(x_j - \bar{x})\big)^2$,

$$\lambda_{\alpha_c}(z \,|\, \alpha_1, \ldots, \alpha_N, \tau_\alpha) = \sqrt{\frac{\sigma_{c,\alpha}}{2\pi}} \exp -\frac{\sigma_{c,\alpha}}{2}(z - \mu_{c,\alpha})^2,$$

with $\sigma_{c,\alpha} = \delta + N\tau_\alpha$ and $\mu_{c,\alpha} = \sum_{i=1}^N \alpha_i/\sigma_{c,\alpha}$, and finally

$$\lambda_{\beta_c}(z \,|\, \beta_1, \ldots, \beta_N, \tau_\beta) = \sqrt{\frac{\sigma_{c,\beta}}{2\pi}} \exp -\frac{\sigma_{c,\beta}}{2}(z - \mu_{c,\beta})^2,$$

with $\sigma_{c,\beta} = \delta + N\tau_\beta$ and $\mu_{c,\beta} = \sum_{i=1}^N \beta_i/\sigma_{c,\beta}$.

Normalizing constant. To calculate the normalizing constant, we proceed as follows: write for $n = 0, \ldots, N$,

$$\varphi_n(\alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_N, \tau_c) = \exp -\frac{\tau_c}{2} \sum_{i=1}^n \sum_{j=1}^T \big(y_{ij} - \alpha_i - \beta_i(x_j - \bar{x})\big)^2,$$

and define also the unnormalized probability densities

$$m_n(\alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_N, \tau_c) = \varphi_{n-1}(\alpha_1, \ldots, \alpha_N, \beta_1, \ldots, \beta_N, \tau_c) \cdot \prod_{i=1}^n \pi(\alpha_i \,|\, \alpha_c, \tau_\alpha)\pi(\beta_i \,|\, \beta_c, \tau_\beta)\dot{h}(\tau_c)\dot{\pi}(\alpha_c)\dot{\pi}(\tau_\alpha)\dot{\pi}(\beta_c)\dot{\pi}(\tau_\beta),$$

where

$$\dot{h}(\tau_c) = \Gamma(a+\tfrac{1}{2})^{-1} b^{a+1/2} \tau_c^{a+1/2-1} e^{-b\tau_\beta} 1_{\{\tau_c>0\}},$$

and as per our convention $\dot{m}_k(x) = m_k(x)/\int m_k(x)dx$. Then we have

$$\begin{aligned}\log C &= -\frac{1}{2}(NT\log 2\pi - \log b) + \log\frac{\Gamma(a+1/2)}{\Gamma(a)} + \log\int \varphi_N(x)\dot{m}_1(x)dx \\ &= -\frac{1}{2}(NT\log 2\pi - \log b) + \log\frac{\Gamma(a+1/2)}{\Gamma(a)} + \sum_{k=1}^{N}\log\int e^{\psi_k(x)}\dot{m}_k(x)dx,\end{aligned}$$

where

$$\psi_k(\alpha_1,\dots,\alpha_N,\beta_1,\dots,\beta_N,\tau_c) = -\frac{\tau_c}{2}\sum_{j=1}^{T}\big(y_{kj}-\alpha_k-\beta_k(x_j-\bar{x})\big)^2.$$

It is impractical to compute the integral $\int \varphi_N(x)\dot{m}_1(x)dx$ directly, for the integrand $\varphi_N(x) = \exp\sum_{k=1}^{N}\psi_k(x)$ is vanishingly small. Since the functions ψ_k are much smaller in magnitude than the full sum $\sum_{k=1}^{N}\psi_k$, the series of integrals with respect to $\dot{m}_k$ is easier to calculate, and yields numbers of manageable size, for which a simple error approximation based on Hoeffding's bound is feasible. For each k, suppose we generate $Z_{k1},\dots,Z_{kr}$ IID variables with density m_k. Since we have $\psi_k \le 0$, we must have

$$\mathbb{P}\left[\left|\int e^{\psi_k(x)}m_k(x)dx - \frac{1}{r}\sum_{s=1}^{r}e^{\psi_k(Z_{ks})}\right| < \delta\right] \ge 1-2e^{-2r\delta^2},$$

and this gives a simple procedure for choosing the number r of samples according to the accuracy required. In the sequel, we shall take r large enough so that $\delta \le \epsilon\big(\frac{1}{r}\sum_{s=1}^{r}e^{\psi_k(Z_{ks})}\big)$ and $2r\delta^2 \ge -\log\epsilon$, for $\epsilon = 0.1$. Note again that this strategy is only practical (i.e. r is of a reasonable magnitude) provided that the typical values of $e^{\psi_k(Z_{ks})}$ are not vanishingly small, which is the whole point of splitting the original integral. It remains to discuss how to generate IID random variables $Z_{ks} \sim \dot{m}_k$.

Assume that for each $k = 1,\dots,N$, the probability density $\dot{m}_k$ exists on its own copy $E_k = \mathbb{R}^{65}$ of the state space. We define a probability distribution $\Pi(dy) = (1/N)\sum_{k=1}^{N}1_{E_k}(dy)\dot{m}_k(y)dy$ on the N-fold disjoint union $E = E_1 \cup \cdots \cup E_N$, noting that $\dot{m}_1(x)$ can be sampled from exactly. Define a Markov chain X_t with stationary distribution Π on E as follows: If t is even and $X_t \in E_k$, $k > 1$, then X_{t+1} is obtained from X_t by one sweep of the Gibbs sampler corresponding to $\dot{m}_k$.

If $k = 1$, then $X_{t+1} \sim m_1$ exactly, and independently of $X_t \in E_1$. If t is odd and $X_t \in E_k$, then X_{t+1} is projected onto $E_{k\pm1}$ with probability $\frac{1}{2} \cdot 1 \wedge \exp \pm\psi_k(X_t)$ (if k falls outside the set $\{1, \ldots, N\}$ we do nothing). This is analogous to simulated tempering (see Moller and Nichols, 1999). Since the chain X_t regenerates whenever $X_t \in E_1$, we may use the ROCFTP method (Wilson, 1999) to identify random times when X_t is in equilibrium. This produces a stream of independent random variables from all the distributions m_k.

More precisely, let $q \geq 2N$ denote a fixed number of iterations, and partition the sample path X_t into the sections $X_{[0,q)} = \{X_1, \ldots, X_q\}$, $X_{[q,2q)} = \{X_{q+1}, \ldots, X_{2q}\}$, etc. For each such section, the path $X_{[tq,(t+1)q)}$ is merely one among the collection of paths starting from all possible initial states in E at time tq and utilizing the same realized source of randomness as $X_{[tq,(t+1)q)}$. With probability $1/2$, when $s \in \{tq+1, \ldots, (t+1)q\}$ is odd, all such paths are projected from E_k to E_{k-1} since $\psi_k \leq 0$ for all k. Alternatively, with probability $1/2$, some paths are projected from E_k to E_{k+1} while others reject the change. If all possible paths reach E_1 before time $(t+1)q$, we say that coalescence occurs in the interval $[tq, (t+1)q)$, since all possible paths are identical from then on. It coalescence occurs in $[tq, (t+1)q)$, then $X_{tq} \in E_k$ is an IID sample from m_k provided simply that this is not the very first coalescence since the beginning of the simulation.

5. Unnormalized target densities with tractable components

In this section, we propose various extensions of Theorem 3 to the case of target densities $\dot{\pi}(x)$ whose normalization constant is not computable. Instead, we shall use an unnormalized version $\pi(x)$, which satisfies $\dot{\pi}(x) = \pi(x)/\int \pi(x)dx$. This requires a tradeoff which translates into a loosening of the bound on L^1 distance developed in (7), and a possible increase in the complexity of the required calculations.

Exact Simulation. We state here without proof a simple modification of Theorem 3 which applies whenever random variables can be simulated from $\dot{\pi}$ exactly. It is based on the following inequality

$$\int |\widehat{\pi}(x) - \dot{\pi}(x)|\, dx \leq \iint |\widehat{\pi}(x)\dot{\pi}(y) - \dot{\pi}(x)\widehat{\pi}(y)|\, dxdy \tag{10}$$

$$= 2 \iint \left[1 - \frac{\pi(y)\widehat{\pi}(x)}{\pi(x)\widehat{\pi}(y)} \right]_+ \dot{\pi}(x)dx\, \widehat{\pi}(y)dy.$$

Note that the integrand in square brackets only involves the unnormalized density $\pi(x)$. The L^1 distance from $\widehat{\pi}(x)$ to $\dot{\pi}(x)$ bounds the distance from $\nu_{\boldsymbol{X}}(x^i)$ to $\dot{\pi}_i(x^i)$ as in (6), so we have

Theorem 4. *Let $\pi(x^1, \ldots, x^d)$ be any unnormalized probability density on $\mathbb{R}^d$ with tractable components. For any $\boldsymbol{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_t\} \subset \mathbb{R}^d$, we define a distribution $\widehat{\pi}_{\boldsymbol{X}}$ by* (5) *on $\mathbb{R}^d$. Choose a finite i.i.d. sequence $Z_1, \ldots, Z_n$ from $\widehat{\pi}_{\boldsymbol{X}}$, an i.i.d. sequence $Z'_1, \ldots, Z'_n$ from $\dot{\pi}$, and $\epsilon > 0$. Then*

$$\mathbb{P}\left[\int |\widehat{\pi}_{\boldsymbol{X}}(x) - \dot{\pi}_i(x)|\, dx < 2\epsilon + \frac{2}{n} \sum_{k=1}^{n} \left(1 - \frac{\pi(Z_k)\widehat{\pi}_{\boldsymbol{X}}(Z'_k)}{\pi(Z'_k)\widehat{\pi}_{\boldsymbol{X}}(Z_k)} \right)_+ \right] \geq 1 - e^{-2\epsilon^2 n}. \tag{11}$$

When exact simulation from $\dot{\pi}(x)$ is feasible, several well established alternatives to Theorem 4 are possible. In particular, the classical Central Limit Theorem can be invoked on any test function f as discussed at the beginning of Section 2. We note however that in so doing, the variance $\int |f|^2\, d\pi$ must itself be estimated, while no such secondary estimate is required in Theorem 4. Moreover, it is arguable that (11) is more informative than a combination of estimated means and variances. The assumption that $\dot{\pi}$ affords exact simulation is however highly specialized, thus undesirable, and we consider next a strategy which does not require it.

Minorization. An interesting connection appears upon inspection of (10), namely that the bound on the right is twice the average rejection probability for an Independence Sampler with target $\dot{\pi}(x)$ and proposal distribution $\widehat{\pi}(x)$. It is well known (Mengersen and Tweedie, 1996) that this sampler is uniformly ergodic if and only if

$$\beta_{\boldsymbol{X}} = \inf_{x \in \mathbb{R}^d} \left\{ \frac{\widehat{\pi}_{\boldsymbol{X}}(x)}{\pi(x)} \right\} > 0. \tag{12}$$

Whenever this holds, we can use the simple bound

$$2 \iint \left[1 - \frac{\pi(y)\widehat{\pi}(x)}{\pi(x)\widehat{\pi}(y)} \right]_+ \dot{\pi}(x)dx\, \widehat{\pi}(y)dy \leq 2 \int \left[1 - \beta_{\boldsymbol{X}} \frac{\pi(y)}{\widehat{\pi}(y)} \right]_+ \widehat{\pi}(y)dy,$$

and derive a statement similar to (8).

The situation is not so simple however, as can be seen in a particular case, when $\widehat{\pi}_{\boldsymbol{X}}$ is replaced by its defining expression (5), and $i = 1$, $d = 3$. We then have

$$\widehat{\pi}_{\boldsymbol{X}}(x^1,x^2,x^3)/\pi(x^1,x^2,x^3)$$
$$= \frac{1}{t}\sum_{s=1}^{t} \lambda_1(x^1 \,|\, \mathfrak{x}_s^2,\mathfrak{x}_s^3)\lambda_2(x^2 \,|\, x^1,\mathfrak{x}_s^3)\lambda_1(x^3 \,|\, x^1,x^2)/\pi(x^1,x^2,x^3)$$
$$= \frac{1}{t}\sum_{s=1}^{t} \lambda_2(\mathfrak{x}_s^2 \,|\, x^1,\mathfrak{x}_s^3)\lambda_3(\mathfrak{x}_s^3 \,|\, x^1,x^2)/\pi(\mathfrak{x}_s^2,\mathfrak{x}_s^3).$$

From this, it appears that (12) can hold only if both λ_2 and λ_3 are bounded away from zero as we let x^1 and x^2 vary over the plane, which severely limits the dependence of λ_2 and λ_3 upon their respective parameters.

Acknowledgement. This research was supported by the European Union TMR network ERB-FMRX-CT96-0096 on "Spatial and Computational Statistics".

References

[1] Brooks, S.P., Dellaprotas, P. and Roberts, G.O. (1996) An approach to diagnosing total variation convergence of MCMC algorithms.

[2] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables, *Amer. Statist. Assoc. J.* **58**, pp. 13-30.

[3] Mengersen, K.L. and Tweedie, R.L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**.

Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7E, DK-9220 Aalborg, Denmark

E-mail address: lbreyer@math.auc.dk