Applied Mathematics and Computation 216 (2010) 2869–2880

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

On solving integral equations using Markov chain Monte Carlo methods

Arnaud Doucet [a], Adam M. Johansen [b,*], Vladislav B. Tadić [b]

[a] Departments of Statistics and Computer Science, University of British Columbia, Vancouver, BC, Canada
[b] Department of Mathematics, University of Bristol, Bristol, UK

ARTICLE INFO

Keywords:
Fredholm equation
Trans-dimensional Markov chain Monte Carlo
Sequential Importance Sampling
Sequential Monte Carlo

ABSTRACT

In this paper, we propose an original approach to the solution of Fredholm equations of the second kind. We interpret the standard Von Neumann expansion of the solution as an expectation with respect to a probability distribution defined on a union of subspaces of variable dimension. Based on this representation, it is possible to use trans-dimensional Markov chain Monte Carlo (MCMC) methods such as Reversible Jump MCMC to approximate the solution numerically. This can be an attractive alternative to standard Sequential Importance Sampling (SIS) methods routinely used in this context. To motivate our approach, we sketch an application to value function estimation for a Markov decision process. Two computational examples are also provided.

1. Fredholm equations and Von Neumann's Expansion

Fredholm equations of the second kind and their variants appear in many scientific fields including optimal control [1], molecular population genetics [2] and physics [3]. Formally, we are interested in solving the integral equation

$$f(x_0) = \int_E K(x_0, x_1) f(x_1) dx_1 + g(x_0), \tag{1}$$

where $g : E \to \mathbb{R}$ and $K : E \times E \to \mathbb{R}$ are known and $f : E \to \mathbb{R}$ is unknown.

Let us define $K^0(x, y) \triangleq 1$, $K^1(x, y) \triangleq K(x, y)$ and

$$K^n(x, y) \triangleq \int K(x, z) K^{n-1}(z, y) dz.$$

If

$$\sum_{n=0}^{\infty} \int_E |K^n(x_0, x_n) g(x_n)| dx_n < \infty, \tag{2}$$

then the solution of the Fredholm equation (1) admits the following Von Neumann series representation; see [3,4] for details:

$$\begin{aligned} f(x_0) &= \int_E K(x_0, x_1) f(x_1) dx_1 + g(x_0) = \int_E K(x_0, x_1) \left[\int_E K(x_1, x_2) f(x_2) dx_2 + g(x_1) \right] dx_1 + g(x_0) \\ &= \int_E \int_E K(x_0, x_1) K(x_1, x_2) f(x_2) dx_1 dx_2 + \int_E K(x_0, x_1) g(x_1) dy + g(x_0), \end{aligned}$$

[*] Corresponding author. Present address: Department of Statistics, University of Warwick, Coventry, UK.
E-mail addresses: arnaud@stat.ubc.ca (A. Doucet), a.m.johansen@warwick.ac.uk (A.M. Johansen), v.b.tadic@bristol.ac.uk (V.B. Tadić).

0096-3003/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2010.03.138

and, by iterating, one obtains

$$f(x_0) = g(x_0) + \sum_{n=1}^{\infty} \int_{E^n} \left(\prod_{k=1}^{n} K(x_{k-1}, x_k) \right) g(x_n) dx_{1:n}, \tag{3}$$

where $x_{i:j} \triangleq (x_i, \ldots, x_j)$ for $i \leqslant j$.

Introducing the notation

$$f_0(x_0) = g(x_0), \tag{4}$$

and, for $n \geqslant 1$,

$$f_n(x_{0:n}) = g(x_n) \prod_{k=1}^{n} K(x_{k-1}, x_k), \tag{5}$$

it is possible to rewrite (3) as

$$f(x_0) = f_0(x_0) + \sum_{n=1}^{\infty} \int_{E^n} f_n(x_{0:n}) dx_{1:n}. \tag{6}$$

We will address two problems in this paper: how to estimate the function $f(x_0)$ over the set E and how to estimate this function point-wise.

There are few scenarios in which a Fredholm equation of the second kind admits a closed-form analytic solution. A great deal of effort has been expended in the development of numerical techniques for the approximate solution of such systems. These fall into two broad categories: deterministic techniques and Monte Carlo techniques. Deterministic techniques typically depend upon quadrature or explicitly obtaining a finite-dimensional representation of the system (by discretisation or projection onto a suitable basis, for example) and then solving that system using numerical techniques. Although good performance can be obtained by these methods, they typically rely upon obtaining a good finite dimensional characterisation of the solution. This remains an active area of research, see [5,6] and references within. Finding such a representation is somewhat problem-specific and is unlikely to be practical for problems in high dimensions or in which the support of the function of interest is not compact. For this reason, we concentrate on Monte Carlo approaches in the remainder of this paper.

2. Monte Carlo methods to solve Fredholm equations

Computing (3) is challenging as it involves an infinite sum of integrals of increasing dimension. Monte Carlo methods provide a mechanism for dealing with such integrals. A sequential importance sampling strategy arises as a natural approach to this problem and that is the approach which has been taken most often in the literature. Section 2.1 summarises this approach and provides a path-space interpretation of the importance sampling which motivates the development of a novel approach in Section 2.2.

2.1. Sequential Importance Sampling

Section 2.1.1 presents the algorithm most commonly presented in the literature; Section 2.1.2 sketches some techniques for reducing the variance of the estimator provided by this algorithm and a path-space interpretation illustrating the unbiasedness of these techniques is given in Section 2.1.3. This interpretation leads naturally to a different approach to the problem which is summarised in the next section.

2.1.1. Algorithm

The use of Monte Carlo methods to solve problems of this type can be traced back 50 years. The standard approach consists of using Sequential Importance Sampling (SIS) to numerically approximate (3); see for example [3,4]. Consider a Markov chain with initial probability distribution/density $\mu(x)$ on E and a transition kernel $M(x, y)$ which gives the probability or probability density of moving to state y when the current state is x. We select μ and M such that $\mu(x) > 0$ over E and $M(x, y) > 0$ if $K(x, y) \neq 0$. Moreover, M is chosen to have an absorbing/cemetery state, say $\dagger \notin E$, such that $M(x, \{\dagger\}) = P_d$ for any $x \in E$.

The algorithm which approximates the function f proceeds as follows:

- Simulate N independent Markov chain paths $\{X^{(i)}_{0:k^{(i)}+1}\}_{i=1}^{N}$ until absorption (i.e. $X^{(i)}_{k^{(i)}+1} = \dagger$).
- Calculate the associated importance weights

$$W_1\left(X^{(i)}_{0:k^{(i)}}\right) = \begin{cases} \frac{1}{\mu\left(X^{(i)}_0\right)} \left(\prod_{k=1}^{k^{(i)}} \frac{K\left(X^{(i)}_{k-1}, X^{(i)}_k\right)}{M\left(X^{(i)}_{k-1}, X^{(i)}_k\right)} \right) \frac{g\left(X^{(i)}_{k^{(i)}}\right)}{P_d} & \text{if } k^{(i)} \geqslant 1, \\ \frac{g\left(X^{(i)}_0\right)}{\mu\left(X^{(i)}_0\right) P_d} & \text{if } k^{(i)} = 0. \end{cases} \tag{7}$$

- The empirical measure

$$\hat{f}(x_0) = \frac{1}{N}\sum_{i=1}^{N} W_1\left(X^{(i)}_{0:k^{(i)}}\right)\delta\left(x_0 - X^{(i)}_0\right) \tag{8}$$

is an unbiased Monte Carlo approximation of the function f (i.e. for any set A, $\mathbb{E}[\int_A \hat{f}(x_0)dx_0] = \int_A f(x_0)dx_0$).

If the objective is the estimation of the function $f(x_0)$ at a given point say $x_0 = x$, then, by simulating paths $\{X^{(i)}_{0:k^{(i)}+1}\}_{i=1}^{N}$ starting from $X^{(i)}_0 = x$ according to M until absorption/death and using the importance weights

$$W_2\left(X^{(i)}_{0:k^{(i)}}\right) = \begin{cases} \left(\prod_{k=1}^{k^{(i)}} \frac{K\left(X^{(i)}_{k-1}, X^{(i)}_k\right)}{M\left(X^{(i)}_{k-1}, X^{(i)}_k\right)}\right) \frac{g\left(X^{(i)}_{k^{(i)}}\right)}{P_d} & \text{if } k^{(i)} \geqslant 1, \\ \frac{g(x)}{P_d} & \text{if } k^{(i)} = 0, \end{cases} \tag{9}$$

we obtain the following unbiased estimate of $f(x)$

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} W_2\left(x, X^{(i)}_{1:k^{(i)}}\right). \tag{10}$$

2.1.2. Variance reduction

The following technique applies to both of the algorithms introduced in the previous section. Notice that, as the probability of death at a given iteration is independent of the path sampled, it is possible to use all paths of length *at least* k to estimate the k-fold integral over E^k. That is, the variance is potentially reduced and the estimator remains unbiased if we replace (8) with:

$$\tilde{f}(x_0) = \frac{1}{N}\sum_{i=1}^{N} \widetilde{W}_1\left(X^{(i)}_{0:k^{(i)}}\right)\delta\left(x_0 - X^{(i)}_0\right), \tag{11}$$

where

$$\widetilde{W}_1(x_{0:k}) = P_d \sum_{i=0}^{k} W_1(x_{0:i}). \tag{12}$$

The same considerations lead us to the conclusion that we should replace (10) with:

$$\tilde{f}(x) = \frac{1}{N}\sum_{i=1}^{N} \widetilde{W}_2\left(x, X^{(i)}_{1:k^{(i)}}\right), \tag{13}$$

where

$$\widetilde{W}_2(x_{0:k}) = P_d \sum_{i=0}^{k} W_2(x_{0:i}). \tag{14}$$

Notice that this approach leads to a reduction in the weight associated with each path by a factor of P_d, but each sample now contributes to $k^{(i)}$ (with expectation $1/P_d$) trajectories rather than one. A related idea is used in the field of reinforcement learning [7].

Another approach can be used to further reduce the variance of estimator (10). Here the first term in the Von Neumann expansion is known deterministically: it is $g(x)$. As such, there is no benefit in estimating it via Monte Carlo approximation: it will reduce the variance if one instead estimates the difference $f(x) - g(x)$ using these techniques. In order to do this, one samples $X^{(i)}_1$ from the restriction of M to E: $X^{(i)}_1 \sim M(x_0, \cdot)\mathbb{I}_E(\cdot)/(1 - P_d)$, and subsequent states from M as before until the chain enters † at a time $\geqslant 2$. This leads to a collection of samples $X^{(i)}_{1:k^{(i)}}$ with $k^{(i)} \geqslant 1$, and allows the use of the modified estimator:

$$\bar{f}(x) = g(x) + \frac{1}{N}\sum_{i=1}^{N} \overline{W}_2\left(x, X^{(i)}_{1:k^{(i)}}\right), \tag{15}$$

where

$$\overline{W}_2\left(x, X^{(i)}_{1:k^{(i)}}\right) = (1 - P_d)\widetilde{W}_2\left(x, X^{(i)}_{1:k^{(i)}}\right). \tag{16}$$

Note that both of these techniques can be employed simultaneously—and indeed, both should be used in any real implementation of these algorithms.

2.1.3. Importance sampling on path space

To check the unbiasedness of the estimates (8) and (10), we use a slightly non-standard argument which will later prove useful.

The first method to estimate the function f through (8) can be interpreted as an importance sampling technique using an importance distribution $\pi_1(n,x_{0:n})$ defined on the path space $F_1 \triangleq \biguplus_{k=0}^{\infty}\{k\}\times E^{k+1}$ where

$$\pi_1(n,x_{0:n}) = p_{1,n}\pi_{1,n}(x_{0:n}), \tag{17}$$

with $p_{1,n}$ the probability that the simulated path is of length $n+1$ (i.e. $X_{0:n}\in E^{n+1}$ and $X_{n+1} = \dagger$) and $\pi_{1,n}(x_{0:n})$ the probability or probability density of a path conditional upon this path being of length $n+1$. We have

$$p_{1,n} = \Pr\left(X_{0:n}\in E^{n+1}, X_{n+1} = \dagger\right) = (1-P_d)^n P_d, \tag{18}$$

and

$$\pi_{1,n}(x_{0:n}) = \frac{\mu(x_0)\prod_{k=1}^{n}M(x_{k-1},x_k)}{(1-P_d)^n}. \tag{19}$$

Now using (6) and importance sampling, this yields

$$f(x_0) = \frac{f_0(x_0)}{\pi_1(0,x_0)}\pi_1(0,x_0) + \sum_{n=1}^{\infty}\int_{E^n}\frac{f_n(x_{0:n})}{\pi_1(n,x_{0:n})}\pi_1(n,x_{0:n})dx_{1:n} = \mathbb{E}_{\pi_1}\left[\frac{f_k(X_{0:k})}{\pi_1(k,X_{0:k})}\right], \tag{20}$$

where the expectation is over both k and $X_{1:k}$ which are jointly distributed according to π_1.

By sampling $\{k^{(i)}, X^{(i)}_{0:k^{(i)}}\}$ $(i = 1,\ldots,N)$ according to π_1, we can obtain the following approximation

$$\hat{f}(x_0) = \frac{1}{N}\sum_{i=1}^{N}\frac{f_{k^{(i)}}\left(X^{(i)}_{0:k^{(i)}}\right)}{\pi_1\left(k^{(i)},X^{(i)}_{0:k^{(i)}}\right)}\delta\left(X_0^{(i)} - x_0\right). \tag{21}$$

It is straightforward to check using (4), (5), (7), (17), (18) and (19) that

$$\frac{f_{k^{(i)}}\left(X^{(i)}_{0:k^{(i)}}\right)}{\pi_1\left(k^{(i)},X^{(i)}_{0:k^{(i)}}\right)} = W_1\left(X^{(i)}_{0:k^{(i)}}\right),$$

thus establishing the unbiasedness of (8).

Similarly, the second method (that which estimates $f(x)$ point-wise using (10)) corresponds to an importance sampling method on the space $F_2 \triangleq \biguplus_{k=0}^{\infty}\{k\}\times E^{k}$. The importance distribution is given by $\pi_2(0,x_{1:0}) \triangleq \pi_2(0) = P_d$ and for $n \geqslant 1$

$$\pi_2(n,x_{1:n}) = p_{2,n}\pi_{2,n}(x_{1:n}),$$

with

$$p_{2,n} = \Pr(X_{1:n}\in E^n, X_{n+1} = \dagger) = (1-P_d)^n P_d, \tag{22}$$

and

$$\pi_{2,n}(x_{1:n}) = \frac{M(x,x_1)\prod_{k=2}^{n}M(x_{k-1},x_k)}{(1-P_d)^n}. \tag{23}$$

Using the importance sampling identity

$$f(x) = \frac{f_0(x)}{\pi_2(0)}\pi_2(0) + \sum_{n=1}^{\infty}\int_{E^n}\frac{f_n(x,x_{1:n})}{\pi_2(n,x_{1:n})}\pi_2(n,x_{1:n})dx_{1:n} = \mathbb{E}_{\pi_2}\left[\frac{f_k(x,X_{1::k})}{\pi_2(k,X_{1:k})}\right], \tag{24}$$

then sampling $\{k^{(i)}, X^{(i)}_{1:k^{(i)}}\}$ $(i = 1,\ldots,N)$ according to π_2, we obtain the following approximation

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{f_{k^{(i)}}\left(x,X^{(i)}_{1:k^{(i)}}\right)}{\pi_2\left(k^{(i)},X^{(i)}_{1:k^{(i)}}\right)}. \tag{25}$$

Using (4), (5), (9), (22) and (23), we have

$$\frac{f_n\left(x,X^{(i)}_{1:k^{(i)}}\right)}{\pi_2\left(k^{(i)},X^{(i)}_{1:k^{(i)}}\right)} = W_2\left(x,X^{(i)}_{1:k^{(i)}}\right),$$

thus establishing the unbiasedness of (10).

Essentially identical arguments hold for (11), (13) and (15) if one considers estimating the integral of each f_n individually, using all available samples and then takes a linear combination of these estimators.

2.1.4. Limitations of SIS

The estimates (21) and (25) will have a reasonable Monte Carlo variance if the variance of the absolute value of the weights is small. However, this can be difficult to ensure using the standard SIS approach. First, it imposes an arbitrary geometric distribution for the simulated paths length (18), (22) which might be inappropriate. Second, a product of terms $K(X_{k-1}^{(i)}, X_k^{(i)})/M(X_{k-1}^{(i)}, X_k^{(i)})$ appears in the expression of the weights if $M \neq K$ [1]; its variance typically increases approximately exponentially fast with the length of the paths. Third, if we are interested in estimating the function on E using (21), the initial distribution μ appears in the denominator of (7). This distribution has to be selected very carefully to ensure that the variance of the resulting weights will be finite. We note that in those rare instances in which one can obtain a reasonable approximation to a time-reversal kernel associated with K one could imagine sampling the sequence backwards according to this kernel and using a distributional approximation of g to initialise each chain.

The performance of SIS algorithms can usually be dramatically improved by introducing a resampling step [8,9]. The basic idea is to monitor the variance of importance weights over time and, when it becomes too large, to discard those paths with small weights and multiply those with high weights, while setting all of the weights to the same value in a principled way which retains the expectation of the estimator.

However, even with an incorporated resampling step, SIS might still be inefficient in the integral-equation context as we are interested in estimating a function which depends upon the beginning of each trajectory. Each time it is used, the resampling step decreases the diversity in the number of paths left from time 0 to the current time index. In contrast to many other applications in which SIS-type algorithms are employed, the present application is most interested in the *initial* rather than final element of the path: due to this elimination of trajectories, resampling is an effective technique only when it is the final location(s) that are of interest.

2.2. Importance sampling using trans-dimensional MCMC

In this paper, we propose an alternative approach in which we do not limit ourselves to simulating paths sequentially. The importance sampling identity (20) is valid for any distribution π_1 such that $\int_{E^n} f_n(x_{0:n})dx_{1:n} \neq 0 \Rightarrow p_{1,n} > 0$ and $f_n(x_{0:n}) \neq 0 \Rightarrow \pi_{1,n}(x_{0:n}) \neq 0$. Similarly (24) is valid when $\int_{E^n} f_n(x, x_{1:n})dx_{1:n} \neq 0 \Rightarrow p_{2,n} > 0$ and $f_n(x, x_{1:n}) \neq 0 \Rightarrow \pi_{2,n}(x_{1:n}) \neq 0$. We now show how it is possible to construct efficient importance distributions which can be sampled from using trans-dimensional MCMC methods.

2.2.1. Optimal importance distributions

When doing importance sampling in settings in which one is interested in approximating the probability distribution itself, rather than the expectation of a single function with respect to that distribution, it is usual to define the optimal proposal as that which minimises the variance of the importance weights. As our "target measure" is signed, we consider minimising the variance of the absolute value of the importance weights.

In detail, we propose selecting importance distributions $\pi_1(n, x_{0:n})$ [resp. $\pi_2(n, x_{1:n})$] which minimize the variance of the absolute value of the importance weights in (21) [resp. (25)] in order to reduce the Monte Carlo variance of these estimates.

Let us first consider case (21). We define $\pi_1(n, x_{0:n})$ on F_1 as follows. The renormalized version of the absolute value of $f_n(x_{0:n})$ is given by

$$\pi_{1,n}(x_{0:n}) = c_{1,n}^{-1}|f_n(x_{0:n})|, \tag{26}$$

with

$$c_{1,n} = \int_{E^{n+1}} |f_n(x_{0:n})|dx_{0:n}.$$

Note that if $g(x) \geqslant 0$ and $K(x, y) \geqslant 0$ for any $x, y \in E$, then assumption (2) ensures that $c_{1,n} < \infty$. However, in the more general case, we need to make the additional assumption that $c_{1,n} < \infty$ for any n. We also consider

$$p_{1,n} = c_1^{-1} c_{1,n}, \tag{27}$$

where

$$c_1 = \sum_{n=0}^{\infty} c_{1,n}. \tag{28}$$

It is assumed here that $c_1 < \infty$; this is true if (2) holds. In this case,

[1] In many applications K is not a Markov kernel and it is impossible to select $M = K$.

$$
\begin{aligned}
f(x_0) &= c_{1,0}\ \mathrm{sgn}\,(f_0(x_0))\pi_{1,0}(x_0) + \sum_{n=1}^{\infty} c_{1,n} \int_{E^n} \mathrm{sgn}(f_n(x_{0:n}))\pi_{1,n}(x_{0:n})dx_{1:n} \\
&= c_1\ \mathrm{sgn}(f_0(x_0))\pi_1(0,x_0) + c_1 \sum_{n=1}^{\infty} \int_{E^n} \mathrm{sgn}(f_n(x_{0:n}))\pi_1(n,x_{0:n})dx_{1:n},
\end{aligned}
$$

where

$$
\mathrm{sgn}(u) = \begin{cases} 1 & \text{if } u \geqslant 0, \\ -1 & \text{if } u < 0. \end{cases}
$$

Given $N \gg 1$ random samples $\{k^{(i)}, X^{(i)}_{0:k^{(i)}}\}$ distributed according to π_1, it is possible to approximate (3) by

$$
\hat{f}(x_0) = \frac{c_1}{N} \sum_{i=1}^{N} \mathrm{sgn}\Big(f_{k^{(i)}}\Big(X^{(i)}_{0:k^{(i)}}\Big)\Big)\delta\Big(x_0 - X^{(i)}_0\Big). \tag{29}
$$

This is clearly the optimal importance distribution as the variance of the absolute values of the importance weights is equal to zero. However, it is usually impossible to sample from $\pi_1(n, x_{0:n})$ exactly and to compute c_1 in closed-form.

We claim that these two problems can be satisfactorily solved in most cases using trans-dimensional MCMC. To sample from π_1, which is a distribution defined on a union of subspaces of different dimensions, we can use any trans-dimensional MCMC method such as the popular Reversible Jump MCMC (RJMCMC) algorithm [10,11]. This idea involves building an F_1-valued ergodic Markov chain $\{k^{(i)}, X^{(i)}_{0:k^{(i)}}\}_{i \geqslant 1}$ which admits π_1 as an invariant distribution. This is a generalization of the standard Metropolis–Hastings algorithm. As $i \to \infty$, one obtains (correlated) samples distributed according to π_1. Moreover, under the standard and realistic assumption that

$$
c_{1,0} = \int_E |g(x)|dx
$$

is known or can be estimated numerically we can obtain the following estimate of c_1 namely

$$
\hat{c}_1 = \frac{c_{1,0}}{\hat{p}_{1,0}},
$$

where $\widehat{p}_{1,0}$ is the proportion of random samples such that $k^{(i)} = 0$; i.e.

$$
\hat{p}_{1,0} = \frac{1}{N} \sum_{i=1}^{N} \delta_0\Big(k^{(i)}\Big). \tag{30}
$$

Now consider the case (25). The importance distribution is defined on $F'_2 = \biguplus_{k=1}^{\infty} \{k\} \times E^k$ with

$$
\pi_2(n, x_{1:n}) = p_{2,n}\pi_{2,n}(x_{1:n}), \tag{31}
$$

where

$$
\begin{aligned}
&\pi_{2,n}(x_{1:n}) = c_{2,n}^{-1}|f_n(x, x_{1:n})|, \\
&c_{2,n} = \int_{E^n} |f_n(x, x_{1:n})|dx_{1:n}
\end{aligned} \tag{32}
$$

and

$$
p_{2,n} = c_2^{-1} c_{2,n}, \tag{33}
$$

$$
c_2 = \sum_{n=1}^{\infty} c_{2,n}. \tag{34}
$$

It is assumed that $c_2 < \infty$; this is true if (2) holds. In this case,

$$
f(x) = f_0(x) + \sum_{n=1}^{\infty} c_{2,n} \int_{E^n} \mathrm{sgn}(f_n(x, x_{1:n}))\pi_n(x_{1:n})dx_{1:n} = f_0(x) + c_2 \sum_{n=1}^{\infty} \int_{E^n} \mathrm{sgn}(f_n(x, x_{1:n}))\pi(n, x_{1:n})dx_{1:n}.
$$

Given $N \gg 1$ random samples $\{(k^{(i)}, X^{(i)}_{1:k^{(i)}})\}_{i=1}^{N}$ distributed according to π_2, it is possible to approximate (3) with

$$
\hat{f}(x) = f_0(x) + \frac{c_2}{N} \sum_{i=1}^{N} \mathrm{sgn}\Big(f_{k^{(i)}}\Big(x, X^{(i)}_{1:k^{(i)}}\Big)\Big). \tag{35}
$$

To sample from π_2, we can use trans-dimensional MCMC. To estimate c_2, we use the fact that if

$$
c_{2,1} = \int_E |f_1(x, x_1)|dx_1 = \int_E |g(x_1)K(x, x_1)|dx_1
$$

is known or can be estimated numerically then we can obtain the following estimate of c_2

$$\hat{c}_2 = \frac{c_{2,1}}{\hat{p}_{2,1}},$$

where $\widehat{p}_{2,1}$ is the proportion of random samples such that $k^{(i)} = 1$; i.e.

$$\hat{p}_{2,1} = \frac{1}{N}\sum_{i=1}^{N} \delta_1\left(k^{(i)}\right). \tag{36}$$

2.2.2. A Reversible Jump Markov chain Monte Carlo algorithm

For the sake of completeness, we describe here a simple RJMCMC algorithm to sample from π_1 as defined by (26)–(28). A very similar algorithm could be proposed to sample from π_2 as defined by (31)–(34). More elaborate algorithms are discussed in [11].

This algorithm is based on update, birth and death moves. Each move is selected with probability $u_{k^{(i)}}, b_{k^{(i)}}$ or $d_{k^{(i)}}$, respectively, with $u_{k^{(i)}} + b_{k^{(i)}} + d_{k^{(i)}} = 1$, at iteration i. We also introduce two proposal distributions on E denoted by $q_u(x,\cdot)$ and $q_b(\cdot)$. We denote the uniform distribution on A by $\mathcal{U}(A)$.

Initialization.

- Set $(k^{(1)}, X_{0:k^{(1)}}^{(1)})$ randomly or deterministically.

Iteration $i \geqslant 2$.

- Sample $U \sim \mathcal{U}[0, 1]$.

  If $U \leqslant u_{k^{(i-1)}}$

  Update move

  · Set $k^{(i)} = k^{(i-1)}$, sample $J \sim \mathcal{U}(\{0, 1, \ldots, k^{(i)}\})$ and $X_J^* \sim q_u(X_J^{(i-1)}, \cdot)$.

  · With probability

$$\min\left\{1, \frac{\pi_1\left(k^{(i)}, \left(X_{0:J-1}^{(i-1)}, X_J^*, X_{J+1:k^{(i)}}^{(i-1)}\right)\right) q_u\left(X_J^*, X_J^{(i-1)}\right)}{\pi_1\left(k^{(i)}, X_{0:k^{(i)}}^{(i-1)}\right) q_u\left(X_J^{(i-1)}, X_J^*\right)}\right\} \tag{37}$$

  set $X_{0:k^{(i)}}^{(i)} = (X_{0:J-1}^{(i-1)}, X_J^*, X_{J+1:k^{(i)}}^{(i-1)})$, otherwise set $X_{0:k^{(i)}}^{(i)} = X_{0:k^{(i-1)}}^{(i-1)}$.

  Else If $U \leqslant u_{k^{(i-1)}} + b_{k^{(i-1)}}$,

  Birth move

  · Sample $J \sim \mathcal{U}\{0, 1, \ldots, k^{(i-1)} + 1\}$, sample $X_J^* \sim q_b(\cdot)$.

  · With probability

$$\min\left\{1, \frac{\pi_1\left(k^{(i-1)} + 1, \left(X_{0:J-1}^{(i-1)}, X_J^*, X_{J:k^{(i-1)}}^{(i-1)}\right)\right) d_{k^{(i-1)}+1}}{\pi_1\left(k^{(i-1)}, X_{0:k^{(i-1)}}^{(i-1)}\right) q_b(X_J^*) b_{k^{(i-1)}}}\right\} \tag{38}$$

  set $k^{(i)} = k^{(i-1)} + 1, X_{0:k}^{(i)} = (X_{0:J-1}^{(i-1)}, X_J^*, X_{J:k^{(i-1)}}^{(i-1)})$, otherwise set $k^{(i)} = k^{(i-1)}, X_{0:k^{(i)}}^{(i)} = X_{0:k^{(i-1)}}^{(i-1)}$.

  Else

  Death move

  · Sample $J \sim \mathcal{U}\{0, 1, \ldots, k^{(i-1)}\}$.

  · With probability

$$\min\left\{1, \frac{\pi_1\left(k^{(i-1)} - 1, \left(X_{0:J-1}^{(i-1)}, X_{J+1:k^{(i-1)}}^{(i-1)}\right)\right) q_b(X_J^{(i-1)}) b_{k^{(i-1)}-1}}{\pi_1\left(k^{(i-1)}, X_{0:k^{(i-1)}}^{(i-1)}\right) d_{k^{(i-1)}}}\right\}, \tag{39}$$

  set $k^{(i)} = k^{(i-1)} - 1, X_{0:k^{(i)}}^{(i)} = (X_{0:J-1}^{(i-1)}, X_{J+1:k^{(i-1)}}^{(i-1)})$, otherwise set $k^{(i)} = k^{(i-1)}, X_{0:k^{(i)}}^{(i)} = X_{0:k^{(i-1)}}^{(i-1)}$.

To compute (37)–(39), one needs to be able to compute ratios of the form

$$\frac{\pi_1(l, x_{0:l})}{\pi_1(k, x_{0:k})} = \frac{c_l \pi_{1,l}(x_{0:l})}{c_k \pi_{1,k}(x_{0:k})} = \left|\frac{f_l(x_{0:l})}{f_k(x_{0:k})}\right|.$$

This can be performed easily as $f_l(x_{0:l})$ and $f_k(x_{0:k})$ are given by (5). It is easy to check that the invariant distribution of this Markov chain is π_1. Ergodicity must be established on a case-by-case basis.

It is not our intention to suggest that this precise algorithm will work well in all circumstances. Indeed, this is certainly not the case: it is always necessary to design MCMC algorithms which are appropriate for the target distribution and this is no exception. However, this simple approach works adequately in the examples presented below and there is a great deal of literature on the design of efficient MCMC algorithms which can be employed when dealing with more challenging problems.

3. Examples

We begin by motivating the MCMC approach with a simple example in which the optimal importance distribution can be obtained analytically but for which the straightforward SIS estimator could easily have infinite variance. This is followed with a toy example for which the solution can be obtained analytically and a realistic example taken from the econometrics literature.

3.1. Motivation: an application to value function estimation

Our motivating application is related to control. We consider a Markov process $\{X_k\}_{k \geqslant 0}$ on E with transition kernel P. Let us introduce a reward function $r : E \to \mathbb{R}^+$ and a discount factor $\gamma \in (0,1)$. When the process is in state x at time k it accumulates a reward $\gamma^k r(x)$. Thus the expected reward starting from $X_0 = x$ is given by

$$V(x) = \mathbb{E}_{X_0=x}\left[\sum_{k=0}^{\infty} \gamma^k r(X_k)\right].$$

The expected reward is called the value function in the optimal control literature [1]. Under standard regularity assumptions, it can be established that the value function satisfies

$$V(x) = \gamma \int_E P(x,y)V(y)dy + r(x),$$

that is a Fredholm equation of the second kind (1) with $f(x) = V(x)$, $K(x, y) = \gamma P(x, y)$ and $g(x) = r(x)$.

We present here a simple example for which all calculations can be performed analytically that emphasizes the limitations of SIS in this context. We denote by $\mathcal{N}(m, \sigma^2)$ the Gaussian distribution of mean m and variance σ^2 and

$$\mathcal{N}(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

We set $P(x,y) = \mathcal{N}(y; \alpha x, \sigma_1^2)$ (with $|\alpha| < 1$) and $r(x) = \mathcal{N}(x; 0, \sigma_r^2)$. In this case, one has

$$X_k|(X_0 = x) \sim \mathcal{N}(m_k(x), \sigma_k^2),$$

with $m_0(x) = x, \sigma_0^2 = 0$ and for $k \geqslant 1$

$$m_k(x) = \alpha^k x, \quad \sigma_k^2 = \left(\sum_{i=1}^{k} \alpha^{2(i-1)}\right)\sigma_1^2.$$

It follows that

$$f(x) = \sum_{k=0}^{\infty} \gamma^k \mathcal{N}(m_k(x); 0, \sigma_k^2 + \sigma_r^2).$$

Consider using an SIS method to solve this problem. A sensible choice for M is

$$M(x,y) = (1 - P_d)P(x,y) + P_d \delta(y - \dagger).$$

If one is interested in estimating the function at a given point $x_0 = x$, then the importance weights are given by (9); that is

$$W_2\left(X^{(i)}_{0:k^{(i)}}\right) = \begin{cases} \left(\frac{\gamma}{(1-P_d)}\right)^{k^{(i)}} \frac{g\left(X^{(i)}_{k^{(i)}}\right)}{P_d} & \text{if } k^{(i)} \geqslant 1, \\ \frac{g(x)}{P_d} & \text{if } k^{(i)} = 0. \end{cases}$$

The variance of the importance weights is given by

$$var\left[W_2\left(x, X^{(i)}_{1:k^{(i)}}\right)\right] = \frac{1}{2P_d\sqrt{\pi}\sigma_r} \sum_{k=0}^{\infty} \left(\frac{\gamma^2}{1-P_d}\right)^k \mathcal{N}(m_k(x); 0, \sigma_k^2 + \sigma_r^2/2) - f^2(x). \tag{40}$$

This variance (40) will be finite only if $\frac{\gamma^2}{1-P_d} < 1$. In this case, the optimal importance function $\pi_{1,n}$ can easily be computed in closed-form as $p_{1,n}$ is known and $\pi_{1,n}(x_{0:n})$ is a Gaussian; the variance of the associated estimate is zero.

When estimating the function $f(x_0)$, we consider the importance weights (7) given by

$$W_1\left(X_{0:k^{(i)}}^{(i)}\right) = \begin{cases} \frac{1}{\mu(X_0^{(i)})}\left(\frac{\gamma}{(1-P_d)}\right)^{k^{(i)}} \frac{g\left(X_{k^{(i)}}^{(i)}\right)}{P_d} & \text{if } k^{(i)} \geqslant 1, \\ \frac{g(X_0^{(i)})}{\mu(X_0^{(i)})P_d} & \text{if } k^{(i)} = 0. \end{cases}$$

The variance of the importance weights is equal to

$$var\left[W_1\left(X_{0:k^{(i)}}^{(i)}\right)\right] = \frac{1}{2P_d\sqrt{\pi}\sigma_r}\left(\sum_{k=1}^{\infty}\left(\frac{\gamma^2}{1-P_d}\right)^k \int \frac{1}{\mu(x_0)}\mathcal{N}(m_k(x_0); 0, \sigma_k^2 + \sigma_r^2/2)dx_0\right) - \left(\int f(x_0)dx_0\right)^2. \tag{41}$$

Assume we consider $\mu(x_0) = \mathcal{N}(x_0; 0, \sigma^2)$, then to ensure that the variance (41) is finite, it requires $\frac{\gamma^2}{1-P_d} < 1$ and

$$\sigma^2 > \frac{\sigma_1^2}{1-\alpha^2} + \frac{\sigma_r^2}{2}.$$

In this case, the optimal importance function $\pi_{2,n}$ admits a closed-form and the variance of the associated estimate is zero.

For more complex problems, it could be impossible to obtain necessary conditions on μ to ensure the variance is finite by analytic means.

3.2. Analytically tractable example

To verify the proposed technique, it is useful to consider a simple, analytically-tractable model. The MCMC algorithm described above was applied to the solution of:

$$f(x) = \int_0^1 \frac{1}{3}\exp(x-y)f(y)dy + \frac{2}{3}\exp(x), \tag{42}$$

which has the solution $f(x) = \exp(x)$.

For simplicity the birth, death and update probabilities were set to 1/3 and a uniform distribution over the unit interval was used for all proposals. Note that previously it has been mentioned that an empirical measure approximates the solution to the Fredholm equation in a weak sense. This approach amounts to using the empirical measure associated with a sample from a related distribution as an approximation of that distribution. In order to recover a continuous representation of the solution it is possible to use standard density estimation techniques. There is much literature in this field: the details of such estimation pose no great technical difficulties and are outside the scope of this paper.

Fig. 1 illustrates that even a simple histogram provides a reasonable representation of the solution. The large number of samples (250,000) used to generate this histogram were required only to produce a reasonably high-resolution depiction of the function of interest using a crude density-estimation technique: many fewer samples would suffice if a more sophisticated density estimation strategy was employed, or integrals with respect to the associated measure were the objects of interest.

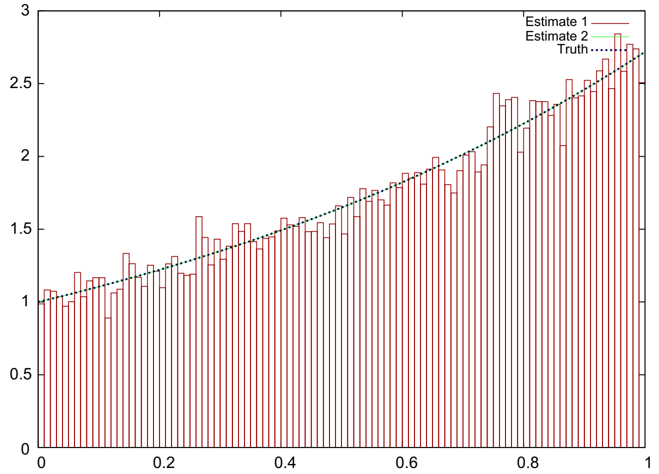

Fig. 1. Histogram of 250,000 samples scaled by the estimated normalising constant (estimate 1), smooth estimate of the same (estimate 2) and the analytic solution for the toy example.

The figure also illustrates one particularly appealing approach, and the one which we would recommend. The Fredholm equation itself provides a natural device for obtaining smooth approximations to the solution of Fredholm equations (with smooth kernels and potentials) from a sample approximation: such an estimate can be obtained by approximating the right hand side of (1) using the sample approximation to the integral on the right hand side. That is, rather than using (21) directly, we use it to approximate the right hand side of the Fredholm equation, obtaining the estimator:

$$\hat{\hat{f}}(x_0) = \int K(x_0,y)\left[\frac{1}{N}\sum_{i=1}^{N}\frac{f_{k^{(i)}}\left(X^{(i)}_{0:k^{(i)}}\right)}{\pi_1\left(k^{(i)},X^{(i)}_{1:k^{(i)}}\right)}\delta_{X^{(i)}_0}(y)\right]dy + g(x_0) \tag{43}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{f_{k^{(i)}}\left(X^{(i)}_{0:k^{(i)}}\right)}{\pi_1\left(k^{(i)},X^{(i)}_{1:k^{(i)}}\right)}K\left(x_0,X^{(i)}_0\right) + g(x_0), \tag{44}$$

which, of course, takes a particularly simple form when the optimal π_1 is chosen. It is clear that this produces a smooth curve in good agreement with the analytic solution (indeed, we cannot distinguish this estimate from the truth).

Table 1 shows the performance of the second MCMC estimator when used for estimating f point-wise. Figures obtained are consistent with an estimator variance proportional to the square of the value being estimated and the reciprocal of the number of samples used.

3.3. An asset-pricing problem

The rational expectation pricing model (see, for example, [12]) requires that the price of an asset in some state $s \in E, V(s)$ must satisfy

$$V(s) = \pi(s) + \beta\int_E V(t)p(t|s)dt. \tag{45}$$

In this equation $\pi(s)$ denotes the return on investment (or the perceived utility of that return), β is a suitable discount factor and $p(t|s)$ is a Markov kernel which models the evolution of the asset's state. E is generally taken to be some compact subset of $\mathbb{R}^n$.

For simplicity, we consider $E = [0,1]$, although there is no difficulty in using the proposed method in the multivariate case, and employ the risk-seeking utility function $\pi(s) = \exp(s^2) - 1$. As suggested by [12], we take $p(t|s)$ a truncated normal distribution, which leads to the following Fredholm equation:

$$V(s) = \frac{\beta}{\sqrt{2\pi\lambda}}\int_0^1 \frac{\exp\left(-\frac{1}{2\lambda}(t-[as+b])^2\right)}{\Phi\left(\frac{1-[as+b]}{\sqrt{\lambda}}\right) - \Phi\left(-\frac{as+b}{\sqrt{\lambda}}\right)}V(t)dt + (\exp(s^2)-1), \tag{46}$$

with Φ denoting the standard normal distribution function (which has associated density ϕ).Thus the potential is $g(s) = \exp(s^2) - 1$ and the kernel may be written in the form

$$K(s,t) = \frac{\beta\phi\left(\frac{t-[as+b]}{\sqrt{\lambda}}\right)}{\Phi\left(\frac{1-[as+b]}{\sqrt{\lambda}}\right) - \Phi\left(-\frac{as+b}{\sqrt{\lambda}}\right)}.$$

For the purposes of this paper we will use the following parameter values: $a = 0.05$, $b = 0.9$, $\beta = 0.85$ and $\lambda = 100$. Note that using such a large value for λ has the effect of making the distribution of X_t almost independent of X_{t-1}. This has been done to

Table 1
Performance of MCMC point-wise-estimation. Figures are obtained from 100 independent instances of the algorithm.

		$N = 100$		$N = 1000$		$N = 10{,}000$	
x	$f(x)$	Mean	Variance	Mean	Variance	Mean	Var./10^{-4}
0	1	1.0516	0.1207	1.0026	0.0006	1.0002	0.4
0.1	1.1052	1.1081	0.0060	1.1047	0.0005	1.1038	0.4
0.2	1.2214	1.2259	0.0103	1.2199	0.0006	1.2214	0.7
0.3	1.3499	1.3864	0.0281	1.3483	0.0008	1.3508	0.9
0.4	1.4918	1.5232	0.0193	1.4893	0.0009	1.4909	1.0
0.5	1.6487	1.6706	0.0430	1.6418	0.0009	1.6488	1.0
0.6	1.8221	1.8277	0.0376	1.8164	0.0019	1.8206	1.2
0.7	2.0138	2.0340	0.0259	2.0178	0.0018	2.0148	2.2
0.8	2.2255	2.2482	0.0471	2.2354	0.0021	2.2245	2.0
0.9	2.4596	2.5316	0.1117	2.4634	0.0034	2.4623	2.7
1	2.7183	2.7693	0.0964	2.7232	0.0037	2.7192	3.4

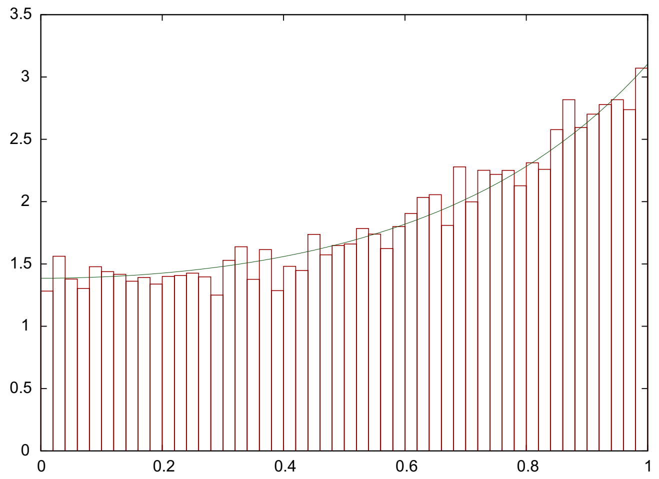

Fig. 2. Histogram and our smooth estimate of V obtained with 100,000.

demonstrate that even in such a simple scenario, it can be impossible for the SIS algorithm to use a good approximation of the optimal importance distribution. Details are provided below.

Within the literature, it is common to compare residuals to assess the performance of an algorithm which provides numerical solutions to Fredholm equations for which no analytic solution is available. Fig. 2 shows a histogram estimate of V obtained using (35) as well as an estimate obtained by approximating the right hand side of (45) using the same sample to approximate the integral (i.e. the approach proposed in the previous section). This shows two things: the agreement is good, suggesting that a valid solution to the equation has indeed been found and a smooth estimate is obtained by the second technique.

Fig. 3 illustrates the distribution of path lengths for samples with values of X_0 close to 0 and *1*. Notice that even in this situation, the distribution is very different for the two regimes. As the length of chains has a distribution independent of their starting points in the SIS case, it would not be possible for such an algorithm to approximate both of these regimes well. It is the non-uniform potential, g, which is responsible for this phenomenon: if the initial sample lies somewhere with a large value of g, then there is little advantage in extending the chain; if the initial value has a very small value of g associated with it then there is a large expected gain. The near-independence of consecutive samples ensures that the distribution of chain length conditional upon the length exceeding 1 is approximately the same for the two regimes, but this would not be true for a more general model.

It would be straightforward to employ the algorithms developed here for more challenging models, although some effort may be required to design good MCMC moves in the case of complex models.

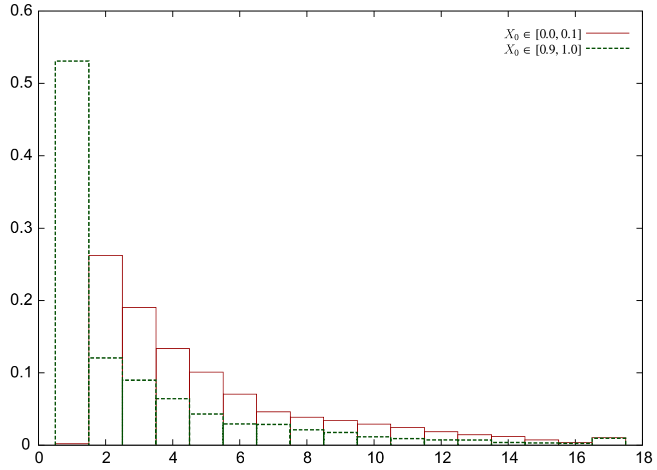

Fig. 3. Distribution of path lengths for two ranges of X_0.

4. Discussion

We have demonstrated that it is possible to solve Fredholm (and by extension, Volterra and other related) equations of the second kind by using trans-dimensional MCMC and an appropriately defined distribution. It is clear that other methods for sampling from such distributions could also be used. The principal rôle of this paper has been to introduce a novel approach and to provide a ‘‘proof of concept”.

The proposed method is qualitatively different to the Monte Carlo methods which have previously been developed for the solution of integral equations. Existing techniques almost all depend upon SIS techniques or closely-related importance sampling strategies. The approach proposed here operates by explicitly defining a distribution over a trans-dimensional space and obtaining samples from that distribution using MCMC (other sampling strategies could also be adopted within the same framework). The one existing approach which appears to be related to the method developed here is described by [13]. This is a specialised technique used to solve a particular problem which arises in ray tracing. It is not clear how the method developed in this context relates to the solution of more general integral equations.

As discussed previously, SIS-based approaches to the solution of integral equations have certain limitations, which the proposed approach avoids. The examples presented above are simple ones, with regular transition kernels in low-dimensional spaces. This choice was made to allow the paper to focus upon methodological developments, but should not be taken as an indication that these are the most sophisticated problems which could be addressed by the above method. Indeed, it is well known that MCMC methods are able to provide samples from extremely complex distributions on spaces of high dimension, albeit at the cost of some design effort and computational time. It is our belief that the proposed technique extends the range of integral equations which can be addressed using Monte Carlo techniques.

References

[1] D.P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, 1990.
[2] R.C. Griffiths, S. Tavaré, Simulating probability distributions in the coalescent, Theoretical Population Biology 46 (1994) 131–159.
[3] J. Spanier, E.M. Gelbard, Monte Carlo Principles and Neutron Transportation Problems, Addison-Wesley, Reading, Massachusetts, 1969.
[4] R. Rubinstein, Simulation and the Monte Carlo Method, Wiley, New York, 1981.
[5] J. Saberi-Nadjafi, M. Heidari, Solving linear integral equations of the second kind with repeated modified trapezoid quadrature method, Applied Mathematics and Computation 189 (1) (2007) 980–985.
[6] E. Babolian, H.R. Marsban, M. Salmani, Using triangular orthogonal functions for solving Fredholm integral equations of the second kind, Applied Mathematics and Computation 201 (1–2) (2008) 452–464.
[7] D.P. Bertsekas, J. Tsitsiklis, Neuro-dynamic Programming, Athena Scientific, 1996.
[8] P. Del Moral, L. Miclo, Branching and interacting particle systems and approximations of Feynman–Kac formulæ with applications to non-linear filtering, in: J. Azéma, M. Emery, M. Ledoux, M. Yor (Eds.), Séminaire de Probabilités XXXIV, Lecture Notes in Mathematics, vol. 1729, Springer-Verlag, Berlin, 2000, pp. 1–145.
[9] A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice, Statistics for Engineering and Information Science, Springer-Verlag, New York, 2001.
[10] P.J. Green, Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (1995) 711–732.
[11] P.J. Green, Trans-dimensional Markov chain Monte Carlo, in: P.J. Green, N.L. Hjort, S. Richardson (Eds.), Highly Structured Stochastic Systems, Oxford Statistical Science Series, Oxford University Press, 2003, pp. 179–206. Ch. 6.
[12] J. Rust, J.F. Traub, H. Wózniakowski, Is there a curse of dimensionality for contraction fixed points in the worst case?, Econometrica 70 (1) (2002) 285–329
[13] E. Veach, L.J. Guibas, Metropolis light transport, in: Proceedings of SIGGRAPH, Addison-Wesley, 1997, pp. 65–76.