ON NON-NEGATIVE UNBIASED ESTIMATORS

BY P.E. JACOB AND A.H. THIERY

University of Oxford and National University of Singapore

We study the existence of algorithms generating almost surely non-negative unbiased estimators. We show that given a non-constant real-valued function \( f \) and a sequence of unbiased estimators of \( \lambda \in \mathbb{R} \), there is no algorithm yielding almost surely non-negative unbiased estimators of \( f(\lambda) \in \mathbb{R}^+ \). The study is motivated by pseudo-marginal Monte Carlo algorithms that rely on such non-negative unbiased estimators. These methods allow “exact inference” in intractable models, in the sense that integrals with respect to a target distribution can be estimated without any systematic error, even though the associated probability density function cannot be evaluated pointwise. We discuss the consequences of our results on the applicability of those sampling algorithms in various statistical settings, such as inference for diffusions, inference in the regime of large datasets, doubly intractable distributions and posterior distributions arising from reference priors.

1. Introduction.

1.1. Motivation. Consider the problem of estimating the integral of a function \( \varphi \) with respect to a probability distribution with density \( \pi \). A successful Markov chain Monte Carlo or sequential Monte Carlo method allows to estimate integrals with respect to \( \pi \) in such a way that the error can be reduced down to zero by producing more samples. We call these methods “exact” since there is no systematic error in the estimation, even though the sampling error can be large for a given computational budget. Using the Metropolis-Hastings algorithm, exact inference is possible when the target probability density function can be evaluated pointwise up to a multiplicative constant.

The possibility of performing exact inference without relying on evaluations of the target probability density function is an important open question. A class of exact methods, called pseudo-marginal Metropolis-Hastings, has been proposed in Andrieu and Roberts (2009), generalizing and validating methods developed in population genetics (Beaumont, 2003) and lattice quantum chromodynamics (Kennedy and Kuti, 1985). Pseudo-marginal methods rely on unbiased estimators of density evaluations \( \pi(x) \) instead of the evaluations themselves. In a related manner, Liu and Chen (1998); Del Moral, Doucet and Jasra (2007); Fearnhead, Papaspiliopoulos and Roberts (2008); Fearnhead et al. (2010); Tran et al. (2013) show that sequential Monte Carlo

Keywords and phrases: Unbiased estimator; Poisson estimator; Monte Carlo methods; Sign problem
methods remain exact when the importance weights are replaced by unbiased esti-

mators thereof.

The applicability of exact methods has thus been considerably extended since
estimating $\pi(x)$ is generally easier than evaluating it. For instance, in the common
case where the cost of evaluating the likelihood function grows at least linearly with
the size of the dataset, pointwise posterior density evaluations become prohibitive
for large datasets (Welling and Teh, 2011; Kleiner et al., 2014). In other settings,
the likelihood cannot be evaluated because it involves an intractable quantity, as
“doubly intractable” posterior distributions commonly found in spatial statistics
and graphical models (Møller et al., 2006; Everitt, 2012). Even for simple models
and small datasets, the use of reference priors for an objective Bayesian analysis
leads to posterior probability density functions that cannot be evaluated pointwise
(Berger, Bernardo and Sun, 2009) because they involve intractable limits or in-
finites. In each case, exact inference could still be achieved through a pseudo-
marginal approach, provided that an appropriate unbiased estimator $\hat{\pi}(x)$ is avail-
able, as will be discussed in Section 4.

Generic techniques to obtain unbiased estimators have been developed inde-
pendently in various fields and have been recently reviewed and generalized in
Rhee and Glynn (2012, 2013). We refer to these as “debiasing techniques” and
recall the main ideas in Section 1.2 for the sake of completeness. The combina-
tion of debiasing techniques and pseudo-marginal methods provides a promising
roadmap to perform exact inference in a very general setting. Unfortunately unbi-
asied estimators $\hat{\pi}(x)$, as produced by current debiasing techniques, can take neg-
ative values with positive probability even if their expectations $\pi(x)$ are known to
be non-negative. These negative values make the direct use of these kind of unbi-
asied estimator useless within a pseudo-marginal Markov Chain algorithm. Like-
wise, standard sequential Monte Carlo methods cannot be directly implemented
when negative values can be encountered. Work-arounds have been proposed to
deal with this so-called sign problem (Girolami et al., 2013), as will be discussed
in Section 5.2.

One might want to avoid the sign problem completely by using unbiased esti-
mators that only take non-negative values. In other words, one might hope to find a
debiasing technique which satisfies a sign constraint. We propose to study the de-
sign of such algorithms. In Section 2 we present a result stating the non-existence
of generic schemes to obtain non-negative unbiased estimators. In Section 3 we
discuss their existence under additional conditions, which in practice require addi-
tional model-specific information. In Section 4 we discuss some consequences in
statistics. Finally we discuss the results and further research venues in Section 5.
1.2. Designing unbiased estimators. In the context of lattice fermions, Kuti (1982) uses a method to unbiasedly estimate some elements of the inverse of a matrix without fully inverting it, based on an infinite series representation and a random truncation; the method is attributed to an unpublished work by J. von Neumann and S.M. Ulam. A similar idea has been proposed by Rychlik (1990) for estimating the derivative of a regression function and by Rychlik (1995) for kernel density estimation. More recently McLeish (2012); Rhee and Glynn (2012, 2013) propose a more general scheme to remove the bias of a sequence of consistent estimators \((S_n)_{n \geq 0}\) of a quantity \(\lambda \in \mathbb{R}\), satisfying

\[
\lim_{n \to \infty} \mathbb{E}(S_n) = \mathbb{E}(S) = \lambda.
\]

In Equation (1.1), the quantity \(S\) can either be thought of as a random variable that is impossible to generate in finite time and \(S_n\) as an approximation of \(S\), or simply as the desired, and generally unknown, value \(S = \lambda\). Suppose that one can sample from \(S_n\) for each \(n \geq 0\). Let \(N\) be an integer-valued random variable that is independent of the sequence \((S_n)_{n \geq 0}\) and that can take arbitrary large values. Under mild assumptions, with the convention \(S_{-1} = 0\), the weights \(w_n = 1/\mathbb{P}(N \geq n)\) are such that the random sum

\[
Y = \sum_{n=0}^{N} w_n \times (S_n - S_{n-1})
\]

is an unbiased estimator of \(\lambda\). The article Rhee and Glynn (2013) establishes the following result.

**THEOREM 1.1 (Theorem 1 of Rhee and Glynn (2013)).** Introduce a random variable \(S\) with \(\mathbb{E}(S) = \lambda \in \mathbb{R}\). Let \((S_n)_{n \geq 0}\) be a sequence of random variables, let \(N\) be an integer valued random variable that can take arbitrarily large values and set \(w_n = 1/\mathbb{P}(N \geq n)\). Under the condition

\[
\sum_{n=1}^{\infty} w_n \times \mathbb{E}(|S - S_{n-1}|^2) < \infty
\]

the random variable \(Y = \sum_{n=0}^{N} w_n \times (S_n - S_{n-1})\), with the convention \(S_{-1} = 0\), is well defined, has expectation \(\mathbb{E}(Y) = \mathbb{E}(S) = \lambda\) and a finite second moment

\[
\mathbb{E}(Y^2) = \sum_{n=0}^{\infty} w_n \times \left(\mathbb{E}(|S - S_{n-1}|^2) - \mathbb{E}(|S - S_n|^2)\right) < \infty.
\]

The “debiased” estimator \(Y\) also generalizes the random truncation approach discussed in Papaspiliopoulos (2011); Girolami et al. (2013) and references therein.
The random variable \( N \) could be replaced by a stopping time. Since the random sum in Equation (1.2) only involves an almost surely finite number of terms, the estimator \( Y \) is straightforward to simulate.

**Example 1.1.** Consider the problem of estimating \( \lambda = \exp(\mathbb{E}[X]) \) given independent copies \((X_n)_{n \geq 1}\) of a random variable \( X \) with expectation \( \mathbb{E}[X] \in \mathbb{R} \). Define \( S_n = 1 + \sum_{j=1}^{n}(\prod_{k=1}^{n}X_k)/n! \). For an integer valued random variable \( N \), the so-called Poisson estimator \( Y \) as defined in Equation (1.2) takes the form

\[
1 + \sum_{n=1}^{N}(\prod_{k=1}^{n}X_k)/(\mathbb{P}(N \geq n)n!)
\]

and is potentially an unbiased estimator of \( \exp(\mathbb{E}[X]) \) with finite variance provided that \( X \) and \( N \) can be chosen so that Condition (1.3) holds. For instance, one can readily check that if \( X \) has a finite variance and if the random variable \( N \) does not decay too quickly to zero in the sense that \( \mathbb{P}(N \geq n) \geq C/(1+\varepsilon)^n \) for some constants \( C, \varepsilon > 0 \) then Equation (1.3) holds with \( S = 1 + \sum_{n=1}^{\infty}(\prod_{k=1}^{n}X_k)/(\mathbb{P}(N \geq n)n!) \).

In the case where the quantity of interest \( \lambda \) is non-negative, the random sum in Equation (1.2) can still take negative values, even if the original estimators \((S_n)_{n \geq 0}\) were all almost surely non-negative; this is because each increment \((S_n - S_{n-1})\) can potentially be negative. An important exception occurs when the estimators \((S_n)_{n \geq 0}\) are ordered, id est \( S_n \geq S_{n-1} \) almost surely; this is for example the case in Example 1.1 if \( X \) is almost surely non-negative. With exact inference in mind, one can wonder about the existence of other debiasing techniques which, unlike \( Y \) of Equation (1.2), would only yield non-negative values. The following section introduces a framework to study that question.

2. Existence of non-negative unbiased estimation schemes.

2.1. Algorithms and factories. For any non-empty measurable space \( \mathcal{X} \subset \mathbb{R} \), let \( M_1(\mathcal{X}) \) be the set of probability distributions on \( \mathcal{X} \) with finite first moment and \( \text{conv}(\mathcal{X}) \) the smallest convex set containing \( \mathcal{X} \). For \( \mu \in M_1(\mathcal{X}) \) we use the notation \( m_1(\mu) = \int_\mathcal{X} x \mu(dx) \) for the mean of \( \mu \); indeed, \( m_1(\mu) \in \text{conv}(\mathcal{X}) \) for any \( \mu \in M_1(\mathcal{X}) \). The distribution of the random variable \( X \) is denoted by \( D(X) \). Let \( L^2(\mathcal{X}) \) be the space of square integrable random variables on \( \mathcal{X} \). The indicator function of a set \( A \) is denoted by \( 1_A \), and \( 1_x \) for some \( x \in \mathbb{R} \) denotes the Dirac delta function centered at \( x \). An unbiased estimator of a quantity \( \lambda \in \mathbb{R} \) is called a \( U \)-estimator of \( \lambda \), or a \( U^+ \)-estimator in the case where it is almost surely non-negative.

For a function \( f : \text{conv}(\mathcal{X}) \rightarrow \mathbb{R}^+ \), we propose to study the existence of \( f \)-factories, defined as devices taking as input \( U \)-estimators of \( \lambda \in \text{conv}(\mathcal{X}) \) with support on \( \mathcal{X} \), and producing \( U^+ \)-estimators of \( f(\lambda) \). Borrowing ideas from
Keane and O’Brien (1994), we first define rigorously a class of algorithms that we will consider practical.

**Definition 2.1.** Let $\mathcal{X}$ be a subset of $\mathbb{R}$. An $\mathcal{X}$-algorithm $A$ is a pair $(T, \varphi)$ where $T = (T_n)_{n \geq 1}$ is a sequence of functions $T_n : (0, 1) \times \mathcal{X}^n \to \{0, 1\}$ and $\varphi = (\varphi_n)_{n \geq 1}$ is a sequence of functions $\varphi_n : (0, 1) \times \mathcal{X}^n \to \mathbb{R}^+$.  

An $\mathcal{X}$-algorithm $A \equiv (T, \varphi)$ takes an infinite sequence $x = (x_n)_{n \geq 1} \in \mathcal{X}^\infty$ and an auxiliary variable $u \in (0, 1)$ as inputs and produces as output 

$$A(u, x) = \varphi_T(u, x_1, \ldots, x_T)$$

with $\tau = \tau(u, x) = \inf\{n \geq 1 : T_n(u, x_1, \ldots, x_n) = 1\}$. We adopt the convention $A(u, x) = \infty$ when $\{n \geq 1 : T_n(u, x_1, \ldots, x_n) = 1\} = \emptyset$ and say in this case that the algorithm does not terminate. In most of the applications that we have in mind, the infinite sequence $x = (x_n)_{n \geq 1} \in \mathcal{X}^\infty$ is the realisation of an independent sequence of random variables $X = (X_n)_{n \geq 1}$ and the variable $u \in (0, 1)$ is the realisation of a random variable $U \sim \text{Uniform}(0, 1)$ independent of $X$. In that case, we say that the algorithm almost surely terminates if $P(\tau < \infty) = 1$. Definition 2.1 translates the fact that a valid algorithm uses a possibly random amount of inputs and that the decision to stop acquiring more inputs only relies on the information contained in the inputs already acquired.

The variable $U$ allows the algorithm to be randomized: on top of the sequence $(X_n)_{n \geq 1}$ it can sample additional random variables. Specifying a single auxiliary variable $U \sim \text{Uniform}(0, 1)$ or an infinite independent sequence $(U_n)_{n \geq 1}$ is equivalent. Indeed, one can construct an infinite sequence of independent Bernoulli random variables by considering the binary expansion of $U \sim \text{Uniform}(0, 1)$, and then partition the expansion into disjoint infinite subsequences to obtain an infinite number of binary representations of independent uniform random variables.

**Definition 2.2.** Let $\mathcal{X}$ be a subset of $\mathbb{R}$ and $f : \text{conv}(\mathcal{X}) \to \mathbb{R}^+$ a function. An $f$-factory $A \equiv (\varphi, T)$ is an $\mathcal{X}$-algorithm such that for any distribution $\pi \in M_1(\mathcal{X})$ and independent sequence $X = (X_n)_{n \geq 1}$ marginally distributed as $\pi$ and auxiliary random variable $U \sim \text{Uniform}(0, 1)$ independent of $(X_n)_{n \geq 1}$, the random variable $Y = A(U, X)$ is a non-negative unbiased estimator of $f(m_1(\pi))$.

The condition $E(A(U, X)) = f(m_1(\pi))$ implies that the algorithm terminates with probability one when fed with the independent sequence $X = (X_n)_{n \geq 1}$ and $U \sim \text{Uniform}(0, 1)$. Importantly the definition implies that an $f$-factory should work for any distribution $\pi \in M_1(\mathcal{X})$.

The article of Keane and O’Brien (1994) introduces a similar definition of algorithm referred to as a Bernoulli factory. Given a subset $\mathcal{P} \subset [0, 1]$ and a function
3.2. Non-existence of general f-factories. We first consider the general case \( X = \mathbb{R} \).

**Theorem 2.1.** For any non-constant function \( f : \mathbb{R} \to \mathbb{R}^+ \), no f-factory exists.

**Proof.** For the sake of contradiction, suppose that there exists a non-constant function \( f : \mathbb{R} \to \mathbb{R}^+ \) and an \( \mathbb{R} \)-algorithm \((\varphi, T)\) as in Definition 2.2; because \( f \) is not constant, there exist two real numbers \( \lambda_X, \lambda_Y \in \mathbb{R} \) with \( f(\lambda_X) > f(\lambda_Y) \).

Choose any distribution \( \mu_X \in M_1(\mathbb{R}) \) with \( m_1(\mu_X) = \lambda_X \) and consider a sequence \( X = (X_n)_{n \geq 1} \) marginally distributed according to \( \mu_X \). For \( \varepsilon > 0 \) and an independent sequence of Bernoulli random variables \((B_n)_{n \geq 1}\) with success probability \( P(B_n = 1) = 1 - P(B_n = 0) = 1 - \varepsilon \), independent from any other source of randomness, the sequence \( Y = (Y_n)_{n \geq 1} \) defined by

\[
Y_n = B_n X_n + \frac{\lambda_Y - \lambda_X(1 - \varepsilon)}{\varepsilon} (1 - B_n)
\]

is such that \( E(Y_n) = \lambda_Y \). For any integer \( n \) we have \( Y_n = X_n \) with arbitrarily large probability \( 1 - \varepsilon \), where \( \varepsilon \) can be chosen arbitrary small, while \( \lambda_Y \) and \( \lambda_X \) are distinct and fixed; this construction is pivotal in all the proofs of this article.

Let us first give an informal description of the proof. We will compare the outputs of the algorithm for the two sequences \((X_n)_{n \geq 1}\) and \((Y_n)_{n \geq 1}\) and a common auxiliary variable \( U \). Suppose first that the algorithm, when applied to the sequence \((X_n)_{n \geq 1}\) terminates after \( n \) steps, say. Tuning the value of \( \varepsilon \) can make the events \( \{(Y_1, \ldots, Y_n) \neq (X_1, \ldots, X_n)\} \) arbitrarily rare. On the other hand the expected outputs are set to \( f(\lambda_X) \) for \((X_n)_{n \geq 1}\) and \( f(\lambda_Y) < f(\lambda_X) \) for \((Y_n)_{n \geq 1}\). Hence,
The dominated convergence theorem yields
\begin{equation}
(2.3)
\end{equation}
so that for all
\begin{align*}
\text{The random variables } \phi_{\tau X}(U, X_1, \ldots, X_{\tau X}) &= f(\lambda_X), \\
\mathbb{E}_{U,Y}(\phi_{\tau Y}(U, Y_1, \ldots, Y_{\tau Y})) &= f(\lambda_Y).
\end{align*}

Notice further that
\begin{equation}
(2.2)
\end{equation}
where we have defined the sets \( L_n = \{ \omega : \tau_X \leq n \} \) and \( M_n = \{ \omega : B_1 = \ldots = B_n = 1 \} = \{ \omega : X_1 = Y_1, \ldots, X_n = Y_n \} \). Since \( \phi_{\tau Y} \) is almost surely non-negative we have for all \( n \geq 1 \)
\begin{align*}
\mathbb{E}_{U,Y}(\phi_{\tau Y}(U, Y_1, \ldots, Y_{\tau Y})) &= \mathbb{E}(\phi_{\tau Y}(U, Y_1, \ldots, Y_{\tau Y})) \\
&\geq \mathbb{E}(\phi_{\tau Y}(U, Y_1, \ldots, Y_{\tau Y}) 1_{L_n \cap M_n}) \\
&= \mathbb{E}(\phi_{\tau X}(U, X_1, \ldots, X_{\tau X}) 1_{L_n \cap M_n}).
\end{align*}

The random variables \( (B_n)_{n \geq 1} \) are independent of all other source of randomness so that for all \( n \geq 1 \) we have
\begin{equation}
(2.3)
\end{equation}
The dominated convergence theorem yields
\begin{align*}
\lim_{n \to \infty} \mathbb{E}_{U,X}(\phi_{\tau X}(U, X_1, \ldots, X_{\tau X}) 1_{L_n}) &= \mathbb{E}_{U,X}(\phi_{\tau X}(U, X_1, \ldots, X_{\tau X})) \\
&= f(\lambda_X).
\end{align*}
so that for any $\delta > 0$ there exists $n_0 = n_0(\delta) \in \mathbb{N}$ such that for all $n \geq n_0$

\[(2.4) \quad f(\lambda_X) - \delta \leq \mathbb{E}_{U,X}(\varphi_{TX}(U,X_1,\ldots,X_{\tau_X})1_{L_n}) \leq f(\lambda_X).\]

One can choose $\delta > 0$ and $\eta > 0$ such that $f(\lambda_Y) + \eta < f(\lambda_X) - \delta$. Equations (2.2), (2.3) and (2.4) yield that for some integer $n_0 = n_0(\delta)$ and any $\varepsilon > 0$ we have

\[
f(\lambda_Y) = \mathbb{E}_{U,Y}(\varphi_{\tau_Y}(U,Y_1,\ldots,Y_{\tau_Y})) \geq \tilde{\mathbb{E}}(\varphi_{TX}(U,X_1,\ldots,X_{\tau_X})1_{L_{n_0}\cap M_{n_0}}) = (1 - \varepsilon)^{n_0}\mathbb{E}_{U,X}(\varphi_{\tau_X}(U,X_1,\ldots,X_{\tau_X})1_{L_{n_0}}) \geq (1 - \varepsilon)^{n_0}(f(\lambda_X) - \delta) > (1 - \varepsilon)^{n_0}(f(\lambda_Y) + \eta).
\]

We obtain a contradiction for $\varepsilon > 0$ small enough. \hfill \Box

Theorem 2.1 indicates in particular that given $U$-estimators $(X_n)_{n \geq 1}$ of a quantity $\lambda$ and without additional knowledge on these estimators we cannot obtain $U^+$-estimators of $\exp(\lambda)$ or $\exp(-\lambda)$.

Another question of interest arises in the case where $\mathcal{X} = \mathbb{R}$ and we are given $U$-estimators of a quantity $\lambda > 0$ and want to construct a $U^+$-estimator $Y$ of the same quantity $\lambda$. This is not exactly equivalent to asking whether there exists an $f$-factory for $f : x \mapsto x$, firstly because we have only defined $f$-factories for $f$ taking values in $\mathbb{R}^+$, and secondly because in Definition 2.2 the algorithm should work for any variable distributed as $\pi \in \mathcal{M}_1(\mathbb{R})$, whereas here we only consider distributions with expectation in $\mathbb{R}^+$.

**Lemma 2.1.** Let $\eta \geq 0$ be a known constant. There does not exist an $\mathbb{R}$-algorithm $\mathcal{A} \equiv (\varphi,T)$ such that for any independent sequence $X = (X_n)_{n \geq 1}$ marginally distributed as $\pi \in \mathcal{M}_1(\mathbb{R})$ with $m_1(\pi) > \eta$ and an auxiliary random variable $U \sim \text{Uniform}(0,1)$ independent from $(X_n)_{n \geq 1}$, the random variable $Y = \mathcal{A}(U,X)$ is a non-negative unbiased estimator of $m_1(\pi)$.

**Proof.** We follow the same arguments as in the proof of Theorem 2.1. Consider $\lambda_X, \lambda_Y \in \mathbb{R}^+$ with $\lambda_X > \lambda_Y > \eta$, and an algorithm $\mathcal{A} \equiv (\varphi,T)$ as in the statement of Lemma 2.1. Let $\mu_X \in \mathcal{M}_1(\mathbb{R})$ with $m_1(\mu_X) = \lambda_X$ and consider an sequence $X = (X_n)_{n \geq 1}$ marginally distributed according to $\mu_X$. One can define $Y$ as in Equation (2.1). Since $\mathbb{E}(Y) = \lambda_Y \geq 0$, one can construct the same contradiction as in the proof of Theorem 2.1. \hfill \Box

The presence of $\eta \geq 0$ in the statement might seem cumbersome but emphasizes that the contradiction does not stem from distributions with expectation arbitrarily close to zero. According to the Lemma 2.1, even if one knows that the sequence of estimators has expectation larger than one, say, one still cannot design an algorithm transforming that sequence into a non-negative random variable with the same expectation.
3. Existence and non-existence under stronger assumptions.

3.1. Case where $X = [a, +\infty)$ or $X = (-\infty, b]$.

**Lemma 3.1.** Let $a, b \in \mathbb{R}$ be two real numbers.

- For an $f$-factory to exist with $X = [a, \infty)$ and $f : X \to \mathbb{R}^+$ continuous, $f$ must be increasing.
- For a $g$-factory to exist with $X = (-\infty, b]$ and $g : X \to \mathbb{R}^+$ continuous, $g$ must be decreasing.

**Proof.** By symmetry we prove only the first assertion. For the sake of contradiction assume that there exist $a \leq \lambda_X < \lambda_Y$ with $f(\lambda_X) > f(\lambda_Y)$ and an algorithm $A \equiv (\varphi, T)$ as in Definition 2.2. Choose any distribution $\mu_X \in \mathcal{M}_1([a, \infty))$ with $m_1(\mu_X) = \lambda_X$ and an independent sequence $X = (X_n)_{n \geq 1}$ marginally distributed according to $\mu_X$. For $\varepsilon \in (0, 1)$, consider the sequence $Y = (Y_n)_{n \geq 1}$ as defined in Equation (2.1). For $\varepsilon > 0$ small enough we have $D(Y) \in \mathcal{M}_1([a, \infty))$ since $\lambda_Y > \lambda_X$. One can then construct exactly the same contradiction as in the proof of Theorem 2.1.

Lemma 3.1 indicates in particular that it is impossible to obtain $U^+$-estimators of $\exp(-\lambda)$ nor of $1/\lambda$ given $U^+$-estimators of a quantity $\lambda > 0$ without exploiting any other additional information on the distribution of these $U^+$-estimators.

For $X = [a, \infty)$ and some increasing functions $f$, there can be explicit constructions of $f$-factories. The case $f : x \mapsto x - a$ is trivial. It is also well known that there exists an $f$-factory for any function $f : [a, \infty) \to \mathbb{R}^+$ that can be expressed as a power series of the type

$$f(x) = \sum_{n=0}^{\infty} c_n (x - a)^n$$

(3.1)

with non-negative coefficients $c_n \geq 0$ for all $n \geq 0$. Indeed, introduce an independent sequence of random variables $(X_k)_{k \geq 1}$ marginally distributed as $\mu_X \in \mathcal{M}_1([a, \infty))$ and an integer-valued random variable $N$; setting the weights $w_n = 1/\mathbb{P}(N \geq n)$ as in Section 1.2, Tonelli’s Theorem yields that the estimator

$$Y = \sum_{n=0}^{N} w_n c_n \prod_{k=1}^{n} (X_k - a),$$

where the product is equal to 1 when $n = 0$, is well-defined, is almost surely non-negative and has expectation $m_1(\mu_X)$. This construction generalizes the construction given in Example 1.1 for the exponential function, and will be discussed again.
in the case of large datasets in Section 4.2. For increasing functions in general, the existence of $f$-factories remains an open question.

We conjecture that the only functions $f : [a, +\infty) \to \mathbb{R}^+$ for which an $f$-factory exists are of the form described in Equation (3.1), with non-negative coefficients $c_n \geq 0$.

3.2. Case where $\mathcal{X} = [a, b]$. The case of a bounded interval $\mathcal{X} = [a, b]$ is the most related to the Bernoulli factory, recalled in Section 2.1. We highlight in this section the similarities and differences between the construction of non-negative estimators and Bernoulli factories. We then give a complete characterization of functions $f : \mathcal{X} = [a, b] \to \mathbb{R}$ for which $f$-factories exist.

Arguments similar to the proof of Theorem 2.1 show that for an $f$-factory to exist, the function $f : \mathcal{X} \to \mathbb{R}^+$ has to be continuous. Such a function $f : \mathcal{X} \to \mathbb{R}^+$ is thus necessarily bounded and we consider a non-trivial interval $[0, \gamma]$ containing its range. If a Bernoulli factory exists for the function $g : [0, 1] \to [0, 1]$ with $g(x) = f(a(1 - x) + bx)/\gamma$, then there exists an $f$-factory. Indeed, consider an i.i.d sequence $X = (X_n)_{n \geq 1}$ marginally distributed according to $\mu_X \in \mathcal{M}_1(\mathcal{X})$. Introduce random variables $(B_n)_{n \geq 1}$, with $B_n := 1_{U_n \leq (X_n - a)/(b - a)}$ where $(U_n)_{n \geq 1}$ is an i.i.d sequence of random variables uniformly distributed on $(0, 1)$. Then $(B_n)_{n \geq 1}$ forms an i.i.d sequence of Bernoulli random variables with mean $(m_1(\mu_X) - a)/(b - a)$. Therefore the Bernoulli factory for $g$ takes the sequence $(B_n)_{n \geq 1}$ as inputs and produces a Bernoulli random variable $\tilde{B}$ with mean $g((m_1(\mu_X) - a)/(b - a)) = f(m_1(\mu_X))/\gamma$. The random variable $\gamma \tilde{B}$ is thus a non-negative unbiased estimator of $m_1(\mu_X)$. As proved in Keane and O’Brien (1994), a necessary and sufficient condition on $g : [0, 1] \to [0, 1]$ for the existence of a Bernoulli factory is

$$\exists \varepsilon > 0 \quad \exists n \in \mathbb{N} \quad \forall x \in [0, 1] \quad \min (g(x), 1 - g(x)) \geq \varepsilon \min \left( x^n, (1 - x)^n \right).$$

It follows that an $f$-factory exists as soon as the condition $\min (f(x), \gamma - f(x)) \geq \varepsilon \min \left( (x - a)^n, (b - x)^n \right)$ is satisfied for some $\varepsilon > 0$, $n \in \mathbb{N}$ and all $x \in [a, b]$. Theorem 3.1 shows in fact that

$$\exists \varepsilon > 0 \quad \exists n \in \mathbb{N} \quad \forall x \in [a, b] \quad f(x) \geq \varepsilon \min \left( (x - a)^n, (b - x)^n \right)$$

is a necessary and sufficient condition for an $f$-factory to exist. The necessary condition $1 - g(x) \geq \varepsilon \min \left( x^n, (1 - x)^n \right)$ for the Bernoulli factory problem to have a solution comes from the fact that the Bernoulli factory has to produce a $\{0, 1\}$-valued estimator; we only need to construct a $[0, \infty)$-valued estimator and can thus get away with the weaker Condition (3.2).
THEOREM 3.1. Let $\mathcal{X} = [a, b]$ be a real interval and $f : \mathcal{X} \to \mathbb{R}^+$ a continuous function that is not identically zero. There exists an $f$-factory if and only if Condition (3.2) holds.

PROOF. The sufficiency is proved as a consequence of the results proved in Keane and O’Brien (1994). The proof of the necessity requires different arguments.

Sufficiency. Let $f : \mathcal{X} \to \mathbb{R}^+$ be a continuous function that satisfies Condition (3.2). Since $f$ is bounded on $\mathcal{X}$, one can find $\gamma \geq \max_{x \in \mathcal{X}} f(x)$ large enough such that $\gamma - f(x) > \varepsilon \min((x - a)^n, (b - x)^n)$ for all $x \in \mathcal{X}$. The discussion before the statement of Theorem 3.1 thus shows that an $f$-factory can be constructed.

Necessity. For notational convenience, we present the proof in the case $\mathcal{X} = [0, 1]$. The general case $\mathcal{X} = [a, b]$ is identical. Let $\mathcal{A} \equiv (T, \varphi)$ be an $f$-factory for some function $f : [0, 1] \to \mathbb{R}^+$. For $x_{1:n} = (x_1, \ldots, x_n) \in \{0, 1\}^n$ and a random variable $U$ uniformly distributed on $(0, 1)$, we denote by $F_n(x_{1:n})$ the set of events such that the algorithm terminates after having processed $x_{1:n}$:

$$F_n(x_{1:n}) = \{ \omega : \inf\{1 \leq k \leq n : T_k(U, x_1, \ldots, x_k) = 1\} = n \}$$

with the convention $\inf\{\emptyset\} = \infty$. We define the expected output at time $n$ by

$$\Psi_n(x_{1:n}) = E\left(1_{F_n(x_{1:n})} \varphi_n(U, x_1, \ldots, x_n)\right).$$

For any index $n \geq 1$ and $x_{1:n} \in \{0, 1\}^n$, $\Psi_n(x_{1:n})$ is a non-negative real number. By Definition 2.2 for any $z \in [0, 1]$, $\Psi_n(x_{1:n})$ is a non-negative real number. By Definition 2.2 for any $z \in [0, 1]$, and an i.i.d sequence $(X_n)_{n \geq 1}$ of Bernoulli random variables with mean $z \in [0, 1]$, we have

$$f(z) = E\left(\sum_{n=1}^{\infty} \Psi_n(X_{1:n})\right)$$

$$= \sum_{n=1}^{\infty} \sum_{x_{1:n} \in \{0, 1\}^n} P(X_{1:n} = x_{1:n}) \Psi_n(x_{1:n}).$$

For any index $n \geq 1$ and $x_{1:n} \in \{0, 1\}^n$, defining $r = \text{Card}\{1 \leq i \leq n : x_i = 1\}$ we have $P(X_{1:n} = x_{1:n}) = z^r(1 - z)^{n-r}$, and the above double sum can be written

$$f(z) = \sum_{n=1}^{\infty} \sum_{r=0}^{n} \binom{n}{r} z^r(1 - z)^{n-r} \Psi_n(x_{1:n})$$

$$= \sum_{p,q \in \mathbb{N}^2} c_{p,q} z^p(1 - z)^q$$

for some non-negative coefficient $c_{p,q} \geq 0$. Condition (3.2) follows. □

imsart-aos ver. 2014/01/08 file: unbiased-aos.tex date: May 13, 2014
4. Applications.

4.1. Exact inference for diffusions. Some specific choices of function $f$ are of special interest in applied probability and statistics. Various articles (Beskos and Roberts, 2005; Beskos et al., 2006; Beskos, Papaspiliopoulos and Roberts, 2006; Jourdain and Sbai, 2007; Fearnhead, Papaspiliopoulos and Roberts, 2008; Olsson and Strojby, 2011; Sermaidis et al., 2012) describe the case $f : x \mapsto \exp(x)$ and its applications to perfect simulation and inference for diffusion processes, through the so-called Poisson estimator.

Let us mention a negative consequence of Theorem 2.1 in this context. On a finite interval $[0, T]$, the probability distribution $Q$ on the space of continuous functions $C([0, T], \mathbb{R})$ generated by a scalar diffusion processes with unit volatility coefficient $dX_t = \mu(X_t) \, dt + dW_t$ has, under mild regularity assumption on the drift function $\mu : \mathbb{R} \to \mathbb{R}$, a Radon-Nikodym derivative with respect to the standard Wiener measure $W$ that can be expressed as

$$
\frac{dQ}{dW}(x_t|_{t=0}) = \exp \left( \int_{t=0}^{T} \Phi(x_t) \, dt \right)
$$

for an explicit function $\Phi : \mathbb{R} \to \mathbb{R}$ given by Girsanov’s theorem. As described in Beskos et al. (2006), unbiased estimates of the integral $\int_{t=0}^{T} \Phi(x_t) \, dt$ can be obtained by standard importance sampling. If one could obtain non-negative and unbiased estimates of the Radon-Nikodym derivative (4.1) from unbiased estimates of $\int_{t=0}^{T} \Phi(x_t) \, dt$, a pseudo-marginal algorithm (Andrieu and Roberts, 2009) could be used to sample exactly from $Q$. Theorem 2.1 shows that such an approach is hopeless unless additional knowledge on the law of $Q$ is taken into account.

4.2. Large datasets. The exponential function also appears naturally in the context of inference for large datasets, where the posterior probability density function $\pi$ is expensive to evaluate point-wise. Numerous articles Bardenet, Doucet and Holmes (2014); Korattikara, Chen and Welling (2014) describe ways of approximating the Metropolis-Hastings ratio in the generic context of large datasets; Maclaurin and Adams (2014) uses a lower bound on the local likelihood to simulate from the exact posterior distribution while evaluating only a subset of the data at each iteration. It is hinted in Korattikara, Chen and Welling (2014) that one can use the Poisson estimator for Bayesian inference in the regime of large data. For example, the log-likelihood $\ell(\theta) = \sum_{i=1}^{n} \log f(y_i | \theta)$ of $n \gg 1$ independent observations $(y_i)_{i=1}^{n}$ is potentially computationally expensive to evaluate and can be unbiasedly estimated at reduced cost by using a random subsample of only $m \ll n$ observations. Indeed, for any $m \geq 1$ the quantity $\hat{\ell}(\theta) = (n/m) \sum_{i=1}^{m} \log f(y_{\sigma_i} \mid \theta)$, where $(\sigma_i)_{i \geq 1}$ is a random subset of $\{1, \ldots, n\}$, is an unbiased estimator of $\ell(\theta)$. If one
could efficiently obtain an $U^+$-estimator of the likelihood $L(\theta) = \exp(\ell(\theta))$ from cheap $U$-estimators $\hat{\ell}(\theta)$, an exact approximation algorithm for posterior inference could be devised in this context. Theorem 2.1 indicates that without additional knowledge on the law of $\hat{\ell}(\theta)$ it is in fact impossible to construct $U^+$-estimators of $L(\theta)$. However the construction is possible if we know for instance that $\hat{\ell}(\theta) > a$ almost surely for some $a \in \mathbb{R}$, as described in Section 3.1.

4.3. Doubly intractable distributions. The choice $f : x \mapsto 1/x$ appears in the context of doubly intractable models (Walker, 2011; Girolami et al., 2013) where the observations are assumed to follow a distribution with density defined as

$$f(y \mid \theta) = \frac{g(y, \theta)}{\int g(s, \theta) ds},$$

for a function $(y, \theta) \mapsto g(y, \theta)$ that can be evaluated pointwise. The denominator $Z(\theta) = \int g(s, \theta) ds$ is generally intractable. Consequently, the posterior density itself generally cannot be evaluated pointwise, which prevents the use of the standard Metropolis-Hastings algorithm. Whilst $Z(\theta)$ itself might be unbiasedly estimated by standard Monte Carlo techniques, our analysis reveals that in general one cannot hope to produce $U^+$-estimators of $1/Z(\theta)$ given a sequence of independent $U$-estimators of $Z(\theta)$ (by Theorem 2.1) or given a sequence of $U^+$-estimators of $Z(\theta)$ (by Lemma 3.1). However, by Theorem 3.1 an $U^+$-estimator of $1/Z(\theta)$ can be generated if one has access to an $U$-estimator of $Z(\theta)$ with support in some known interval $[a, b]$.

4.4. Reference priors. The posterior density induced by a reference prior, as described in Berger, Bernardo and Sun (2009), is typically intractable. Given an arbitrary starting prior $\pi^*(\theta)$ and an observation density $p(y \mid \theta)$ the reference prior density $\pi_r$ can be defined as follows

1. For any $k$ define $f_k$ for all $\theta$ as

$$f_k(\theta) = \exp \left( \int p(y_1, \ldots, y_k \mid \theta) \log \pi^*(\theta \mid y_1, \ldots, y_k) dy_1 \ldots dy_k \right).$$

2. Then for any $\theta_0$ define $\pi_r$ for all $\theta$ as

$$\pi_r(\theta) = \lim_{k \to \infty} \frac{f_k(\theta)}{f_k(\theta_0)}.$$  

Although $\pi_r$ might not be proper, it leads to a well-defined posterior distribution, from which we might want to sample using the Metropolis-Hastings algorithm. In this case, even if the observation density $p(y \mid \theta)$ is tractable, the posterior...
density is not. Indeed the normalisation constant of \( \pi^*(\theta \mid y_1, \ldots, y_k) \) would be intractable in most cases, and the integral over \( y_1, \ldots, y_k \) in the definition of each function \( f_k \) would be intractable as well. An unbiased estimator of this integral might still exist. Then the exponential function would have to be applied to yield an unbiased estimator of \( f_k(\theta) \). The limit \( k \to \infty \) then has to be taken, for instance using an infinite telescoping sum representation and a random truncation technique as in Section 1.2, in order to finally obtain an unbiased estimator of \( \pi_r(\theta) \). Thus if getting an unbiased estimator of \( \pi_r(\theta) \) appears feasible in principle, we see that guaranteeing almost sure non-negativity is impossible in general.

5. Discussion.

5.1. Limits of the analysis. The results of Section 2.2 suggest that, for a non-trivial function \( f : \mathbb{R} \to \mathbb{R}^+ \), the ability to sample an unbiased estimator \( X \) of a quantity \( \lambda \) is not enough to obtain a non-negative unbiased estimator of \( f(\lambda) \). However, as described in Section 3, when more information such as almost sure lower and/or upper bounds on \( X \) is available, an \( f \)-factory might exist; this allows exact inference based on the pseudo-marginal approach. We have identified functions \( f \) and sets \( \mathcal{X} \) supporting \( X \) such that \( f \)-factories exist; the case where \( f \) is increasing and \( \mathcal{X} = [a, \infty) \) remains however partly unsettled.

We have prescribed as inputs of \( f \)-factories unbiased estimators of arbitrary quantities \( \lambda \in \mathbb{R} \); other types of inputs could be envisioned, such as estimators consistent in \( L_2 \). However in this case we could first apply their debiasing technique recalled in Section 1.2 and then feed the output to our \( f \)-factories, and hence the conclusions would be similar.

Finally we have not considered the multi-dimensional case \( f : \mathbb{R}^d \to \mathbb{R}^+ \) for \( d > 1 \) since, in the context of exact inference, quantities of interest are posterior density evaluations. Note nevertheless that the statement and proof of Theorem 2.1 readily generalize to much more general settings.

5.2. Dealing with negative values. Exact inference is still possible using signed unbiased estimators. One approach consists in using the absolute value of the estimates and then to correct for the discrepancy using importance sampling, as is commonly done in the statistical physics literature (Lin, Liu and Sloan, 2000). Assume that we want to approximate an integral with respect to a target distribution \( \pi \) and that we have access to a unbiased estimator of \( C\pi(x) \) for each \( x \) and some normalising constant \( C \), that is we can sample \( U \sim q(u \mid x) \) and evaluate \( z(x, u) \) such that \( \int z(x, u)q(u \mid x)du = C\pi(x) \). Even if \( z(x, u) \) is allowed to take negative values, we can introduce the function \( \pi \) defined by \( \pi(x, u) = z(x, u)q(u \mid x) \). Then an integral \( \int \varphi(x)\pi(x)dx \) can be approximated using an extended space rep-
presentation as follows
\[
\int \varphi(x)\pi(x)dx = C^{-1} \int \int \varphi(x)z(x, u)q(u \mid x)dudx = \frac{\int \int \varphi(x)\pi(x, u)dxdudx}{\int \pi(x, u)dxdudx}.
\]

Even if \(z(x, u)\) can take negative values, we can write \(\sigma(x, u)\) for the sign of \(z(x, u)\) and express the integral as
\[
\frac{\int \int \varphi(x)\sigma(x, u)|z(x, u)|q(u \mid x)dxdudx}{\int \int \sigma(x, u)|z(x, u)|q(u \mid x)dxdudx}.
\]

This amounts to transferring the sign from the integrating measure to the test function. Using this representation a Markov chain \((X_t, U_t)_{t \geq 1}\) can be designed to target a probability distribution \(\tilde{\pi}\) with density \(\tilde{\pi}(x, u) \propto |z(x, u)|q(u \mid x)\). Then the ratio of estimates
\[
(5.1) \quad \frac{T^{-1} \sum_{t=1}^{T} \varphi(X_t)\sigma(X_t, U_t)}{T^{-1} \sum_{t=1}^{T} \sigma(X_t, U_t)}
\]
converges to \(\int \varphi(x)\pi(x)dx\) as \(T\) goes to infinity, as described in Girolami et al. (2013). This method typically scales exponentially poorly with the dimension of the sampling space (Troyer and Wiese, 2005), an issue referred to as the “sign problem” in the lattice quantum chromodynamics literature. The performance of the estimator given in Equation (5.1) could be compared to the estimator provided by an ideal Metropolis-Hastings sampler directly targeting \(\pi\), in a similar spirit as the study of Andrieu and Vihola (2014); this is beyond the scope of the present article. Interestingly, the estimator of Equation (5.1) allows exact inference using a Metropolis-Hastings scheme and without non-negative unbiased estimators. Applications to doubly intractable distributions are proposed in Girolami et al. (2013).

5.3. Computationally practical estimators. When \(f\)-factories exist as in Section 3.2, we have discussed an implementable scheme based on the Bernoulli factory. The algorithms considered in Definition 2.2 terminate with probability one but the expectation of the computational time is not necessarily finite.

For the debiasing scheme described in Section 1.2, Rhee and Glynn (2013) consider conditions for the unbiased estimator of Equation (1.2) to have a finite expected computational cost. Taking both the variance and the computational cost into account, an optimal choice of the random truncation variable \(N\) can be made given the convergence rate of the consistent estimators \((S_n)_{n \geq 0}\) used as inputs. When the original estimators \((S_n)_{n \geq 0}\) converge too slowly to their limit, it is impossible to obtain an unbiased estimator with both a finite variance and a finite expected computational cost.
Similarly the existence of \( f \)-factories for a particular function \( f \) does not ensure the practicality of the resulting estimator. In relation to the case \( \mathcal{X} = [a, b] \) of Section 3.2, the recent literature on Bernoulli factories has indeed focused on characterizing algorithms that generate the desired output using as little inputs as possible (Latuszyński et al., 2011; Thomas and Blanchet, 2011; Flegal and Herbei, 2012). The computational cost of \( f \)-factories in this setting is expected to be similar to the cost of Bernoulli factories, the only difference being that the output variable does not have to be Bernoulli distributed. It is not clear that much computational efficiency could be obtained from this additional freedom.

5.4. Exact or inexact inference. An advantage of exact methods, where no systematic bias remains, is that the error is entirely due to the variation in the Monte Carlo algorithm and thus is straightforward to quantify and to interpret (Wagner, 1987). The trade-off between computational feasibility and exactness is ubiquitous in statistics, for instance between Ensemble Kalman Filters and Particle Filters (Frei and Künsch, 2013) or between Approximate Bayesian Computation and Markov Chain Monte Carlo (Marin et al., 2012). In some contexts a non-negative unbiased estimator of the likelihood can be directly obtained, for instance in the context of discrete time state space models or hidden Markov models, where particle filters provide an efficient likelihood estimator that is non-negative by design. In that setting the pseudo-marginal approach leads to the popular class of particle Markov chain Monte Carlo methods (Andrieu, Doucet and Holenstein, 2010) which are both exact and computationally realistic.

Our study indicates that in some contexts non-negative unbiased estimators cannot be obtained and thus the pseudo-marginal approach cannot be applied. We stress that even when the method could be applied in principle, its computational cost might prevent any practical implementation, as noted in Section 5.3. Among inexact methods approximate versions of Metropolis–Hastings (Ceperley and Dewing, 1999; Nicholls, Fox and Watt, 2012; Raftery et al., 2012; Korattikara, Chen and Welling, 2014; Bardenet, Doucet and Holmes, 2014) that do not directly rely on \( U^+ \)-estimators have appeared. In the case of Bardenet, Doucet and Holmes (2014) the bias is precisely controlled in terms of total variation distance between the original distribution of interest and the distribution targeted by the algorithm. Other approaches (Welling and Teh, 2011; Ahn, Korattikara and Welling, 2012; Chen, Fox and Guestrin, 2014) bypass the Metropolis-Hastings acceptance step by discretizing a Langevin diffusion. Given a sample point \( \theta_t \) at step \( t \), the next point is obtained as

\[
\theta_{t+1} = \theta_t + \frac{\varepsilon C}{2} \hat{G}(\theta_t) + \nu_t
\]

where \( \varepsilon \) corresponds to the discretisation step size, \( C \) a preconditioning matrix, \( \hat{G}(\theta_t) \) an unbiased estimator of the gradient of the log target density at \( \theta_t \), and \( \nu_t \)
is a centred Gaussian variable with variance $\varepsilon C$. As $\varepsilon$ goes to zero the sampling method becomes exact, but for a fixed step size it is biased and controlling its error is an active area of research. The approach appears particularly suited to the context of large datasets mentioned in Section 4.2, where the gradient of the log-likelihood can be unbiasedly estimated by subsampling the dataset.

**Acknowledgements.** All authors were supported by a Singapore MoE grant. We are grateful to Mourad Sabilellah for stimulating discussions.

**References.**

AHN, S., KORATTIKARA, A. and WELLING, M. (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo (with discussion). *J. Royal Statist. Society Series B* 72 357–385.

ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* 37 697–725.

ANDRIEU, C. and VIHOLA, M. (2014). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Applied Probability* (to appear).

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of The 31st International Conference on Machine Learning* 405–413.

BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* 164 1139-1160.

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* 905–938.

BESKOS, A., PAPASPILIPOULOS, O. and ROBERTS, G. O. (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli* 12 1077–1098.

BESKOS, A. and ROBERTS, G. O. (2005). Exact simulation of diffusions. *Ann. Applied Probability* 15 2422–2444.

BESKOS, A., PAPASPILIPOULOS, O., ROBERTS, G. O. and FEARNSHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. Royal Statist. Society Series B* 68 333–382.

CEPERLEY, D. and DEWING, M. (1999). The penalty method for random walks with uncertain energies. *The Journal of chemical physics* 110 9812.

CHEN, T., FOX, E. B. and GUESTRIN, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo. In *Proceedings of The 31st International Conference on Machine Learning*.

DEL MORAL, P., DOUCET, A. and JASRA, A. (2007). Sequential monte carlo for bayesian computation. In *Bayesian statistics 8: proceedings of the eighth Valencia International Meeting, June 2-6, 2006 (J. M. Bernardo, M. J. Bayarri, J. O. Degroot, A. P. Dawid, D. Heckerman, A. M. Smith and M. West, eds.)* 8 115. Oxford University Press, USA.

EVERITT, R. G. (2012). Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks. *Journal of Computational and Graphical Statistics* 21 940-960.

FEARNSHEAD, P., PAPASPILIPOULOS, O. and ROBERTS, G. O. (2008). Particle filters for partially observed diffusions. *J. Royal Statist. Society Series B* 70 755–777.

FEARNSHEAD, P., PAPASPILIPOULOS, O., ROBERTS, G. O. and STUART, A. (2010). Random weight particle filtering of continuous time processes. *J. Royal Statist. Society Series B* 72 497–513.
Flegal, J. M. and Herbei, R. (2012). Exact sampling for intractable probability distributions via a Bernoulli factory. *Electronic Journal of Statistics* **6** 10–37.

Frei, M. and Künsch, H. R. (2013). Bridging the ensemble Kalman and particle filters. *Biometrika* **100** 781–800.

Girolami, M., Lyne, A. M., Strathmann, H., Simpson, D. and Atchade, Y. (2013). Playing Russian Roulette with Intractable Likelihoods. *ArXiv e-prints*.

Jourdain, B. and Sbai, M. (2007). Exact retrospective Monte Carlo computation of arithmetic average Asian options. *Monte Carlo Methods and Applications* **13** 135–171.

Keane, M. S. and O’Brien, G. L. (1994). A Bernoulli factory. *ACM Trans. Model. Comput. Simul.* **4** 213–219.

Kennedy, A. D. and Kuti, J. (1985). Noise without Noise: A New Monte Carlo Method. *Phys. Rev. Lett.* **54** 2473–2476.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014). A scalable bootstrap for massive data. *J. Royal Statist. Society Series B*.

Korattikara, A., Chen, Y. and Welling, M. (2014). Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *Proceedings of The 31st International Conference on Machine Learning*.

Kuti, J. (1982). Stochastic method for the numerical study of lattice fermions. *Physical Review Letters* **49** 183.

Łatuszyński, K., Kosmidis, I., Papaspiliopoulos, O. and Roberts, G. O. (2011). Simulating events of unknown probabilities via reverse time martingales. *Random Structures & Algorithms* **38** 441–452.

Lin, L., Liu, K. and Sloan, J. (2000). A noisy Monte Carlo algorithm. *Physical Review D* **61** 074505.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association* **93** 1032–1044.

Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with Subsets of Data. *arXiv preprint arXiv:1403.5693*.

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing* **22** 1167–1180.

McLeish, D. (2012). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications* **17** 301–315.

Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93** 451–458.

Nacu, Ş. and Peres, Y. (2005). Fast simulation of new coins from old. *Ann. Applied Probability* **15** 93–115.

Nicholls, G. K., Fox, C. and Watt, A. M. (2012). Coupled MCMC with a randomized acceptance probability. *ArXiv e-prints*.

Olsson, J. and Ströjby, J. (2011). Particle-based likelihood inference in partially observed diffusion processes using generalised Poisson estimators. *Electronic Journal of Statistics* **5** 1090–1122.

Papaspiliopoulos, O. (2011). A methodological framework for Monte Carlo probabilistic inference for diffusion processes. In *Bayesian time series models* (D. Barber, A. T. Cemgil and S. Chiappa, eds.). Cambridge University Press.

Raftery, A. E., Niú, X., Hoff, P. D. and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics* **21** 901–919.

Rhee, C.-H. and Glynn, P. W. (2012). A new approach to unbiased estimation for SDE’s. In *Proceedings of the Winter Simulation Conference. WSC ’12* 17:1–17:7. Winter Simulation Con-
ference.

Rhee, C.-H. and Glynn, P. W. (2013). Unbiased Estimation with Square Root Convergence for SDE Models. Technical Report, Stanford University.

Rychlik, T. (1990). Unbiased nonparametric estimation of the derivative of the mean. Statistics & probability letters 10 329–333.

Rychlik, T. (1995). A class of unbiased kernel estimates of a probability density function. Applications Math 22 485–497.

Sermaidis, G., Papaspiliopoulos, O., Roberts, G. O., Beskos, A. and Fearnhead, P. (2012). Markov chain Monte Carlo for exact inference for diffusions. Scandinavian Journal of Statistics.

Thomas, A. and Blanchet, J. H. (2011). A Practical Implementation of the Bernoulli Factory. ArXiv e-prints.

Tran, M. N., Schartz, M., Pitt, M. K. and Kohn, R. (2013). IS’2 for Bayesian inference in latent variable models. ArXiv e-prints.

Troyer, M. and Wiese, U.-J. (2005). Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. Physical review letters 94 170201.

Wagner, W. (1987). Unbiased Monte Carlo evaluation of certain functional integrals. Journal of Computational Physics 71 21–33.

Walker, S. G. (2011). Posterior sampling when the normalizing constant is unknown. Communications in Statistics-Simulation and Computation 40 784–792.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) 681–688.

Department of Statistics
University of Oxford
E-MAIL: pierre.jacob@stats.ox.ac.uk

Department of Statistics and Applied Probability
National University of Singapore
E-MAIL: a.h.thiery@nus.edu.sg