On non-negative unbiased estimators

Pierre E. Jacob
© University of Oxford
& Alexandre H. Thiéry
© National University of Singapore

Banff – March 2014
Outline

1. Exact inference
2. Unbiased estimators and sign problem
3. Existence and non-existence results
4. Discussion
Outline

1. Exact inference
2. Unbiased estimators and sign problem
3. Existence and non-existence results
4. Discussion
For a target probability distribution with unnormalised density $\pi$, a numerical method is “exact” if for any test function $\varphi$,

$$\frac{\int \varphi(\theta)\pi(\theta) d\theta}{\int \pi(\theta) d\theta}$$

can be approximated with arbitrary precision, at the cost of more computational effort.

$\Rightarrow$ In this sense MCMC is exact.
Exact inference

- With exact methods, no systematic error.

- No guarantees that for a fixed computational budget, exact methods should be preferred over approximate methods.

- Still important to know in which settings exact methods are available.
Metropolis-Hastings algorithm.

1: Set some $\theta^{(1)}$.
2: for $i = 2$ to $N_\theta$ do
3:   Propose $\theta^* \sim q(\cdot | \theta^{(i-1)})$.
4:   Compute the ratio:
\[
\alpha = \min \left( 1, \frac{\pi(\theta^*)}{\pi(\theta^{(i-1)})} \frac{q(\theta^{(i-1)} | \theta^*)}{q(\theta^* | \theta^{(i-1)})} \right).
\]
5:   Set $\theta^{(i)} = \theta^*$ with probability $\alpha$, otherwise set $\theta^{(i)} = \theta^{(i-1)}$.
6: end for
Exact inference with unbiased estimators

Pseudo-marginal Metropolis-Hastings algorithm.
For each $\theta$, we can sample $Z(\theta)$ with $\mathbb{E}(Z(\theta)) = \pi(\theta)$.

1: Set some $\theta^{(1)}$ and sample $Z(\theta^{(1)})$.
2: for $i = 2$ to $N_\theta$ do
3: Propose $\theta^* \sim q(\cdot|\theta^{(i-1)})$ and sample $Z(\theta^*)$.
4: Compute the ratio:

$$\alpha = \min \left( 1, \frac{Z(\theta^*)}{Z(\theta^{(i-1)})} \frac{q(\theta^{(i-1)}|\theta^*)}{q(\theta^*|\theta^{(i-1)})} \right).$$

5: Set $\theta^{(i)} = \theta^*$, $Z(\theta^{(i)}) = Z(\theta^*)$ with probability $\alpha$, otherwise set $\theta^{(i)} = \theta^{(i-1)}$, $Z(\theta^{(i)}) = Z(\theta^{(i-1)})$.
6: end for
Exact inference with unbiased estimators

- Game-changer when one has access to efficient unbiased estimators of the target density.

- Especially in parameter inference for hidden Markov models, with particle MCMC methods based on particle filters to estimate the likelihood.

- What if one doesn’t have access to straightforward unbiased estimators? Are there general schemes to obtain those unbiased estimators?
Example: big data

Observations $y_i \overset{iid}{\sim} f_\theta$ for $i = 1, \ldots, n$, and $n$ is very large. Can we do exact inference without computing the full likelihood every time we try a new parameter value?

- Unbiased estimator of the log-likelihood

$$\hat{\ell}(\theta) = \frac{n}{m} \sum_{i=1}^{m} \log f(y_{\sigma_i} \mid \theta)$$

for $m < n$ and $\sigma_i$ corresponding to some subsampling scheme.

- It doesn’t directly provide an unbiased estimator of the likelihood.
One can typically get an unbiased estimator of $C(\theta)$ using importance sampling.

- It doesn't directly provide an unbiased estimator of $1/C(\theta)$. 
Reference priors

Starting from an arbitrary prior $\pi^*$, define

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

and the reference prior is, for any $\theta_0$ in the interior of $\Theta$,

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

(Berger, Bernardo, Sun 2009.)

Unbiased estimators of $f(\theta)$?
Unbiased estimators

Von Neuman & Ulam (∼ 1950), Kuti (∼ 1980), Rychlik (∼ 1990), McLeish, Rhee & Glynn (∼ 2010)...

Removing the bias off consistent estimators

Introduce

- a random variable $S$ with $E(S) = \lambda \in \mathbb{R}$,
- a sequence $(S_n)_{n \geq 0}$ converging to $S$ in $L^2$,
- $N$ be an integer valued random variable and $w_n = 1/P(N \geq n) < \infty$ for all $n \geq 0$,

then

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

has expectation $E(Y) = E(S) = \lambda$. 
Unbiased estimators

■ If
\[
\sum_{n=1}^{\infty} w_n \times \mathbb{E}(|S - S_{n-1}|^2) < \infty,
\]
then the variance of $Y$ is finite.

■ Denote by $\bar{t}_n$ the expected computing time to obtain $S_n - S_{n-1}$. Then the computing time of $Y$, denoted by $\bar{\tau}$, should preferably satisfy
\[
\mathbb{E}(\bar{\tau}) = \sum_{n=0}^{\infty} w_n^{-1} \times \bar{t}_n < \infty.
\]

Success story in multi-level Monte Carlo.
Unbiased estimators

Even if the consistent estimators $S_n$ are each almost-surely non-negative, $Y$ is not in general almost-surely non-negative:

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

unless we manage to construct ordered consistent estimators, ie:

$$\mathbb{P}(S_{n-1} \leq S_n) = 1.$$ 

Direct implementation of the pseudo-marginal approach is difficult in the presence of possibly negative acceptance probabilities.
Dealing with negative values

- One can still perform exact inference by noting

\[ \int \frac{\varphi(\theta)\pi(\theta) d\theta}{\int \pi(\theta) d\theta} = \frac{\int \varphi(\theta)\sigma(\pi(\theta))|\pi(\theta)| d\theta}{\int \sigma(\pi(\theta))|\pi(\theta)| d\theta} \]

which suggests using the absolute values of \( Z(\theta) \) in the MH acceptance ratio.

- The integral is recovered using the importance sampling estimator:

\[ \sum_{i=1}^{N} \frac{\sigma(Z(\theta^{(i)}))\varphi(\theta^{(i)})}{\sum_{i=1}^{N} \sigma(Z(\theta^{(i)}))} \]

- As an importance sampler, deteriorates with the dimension. Called the sign problem in lattice quantum chromodynamics.
Avoiding the sign problem

- Can we avoid the sign problem by directly designing non-negative unbiased estimators?

- Given an unbiased estimator of $\lambda > 0$, can I generate a non-negative unbiased estimator of $\lambda$?

- Let $f$ be any function $f : \mathbb{R} \rightarrow \mathbb{R}^+$. Given an unbiased estimator of $\lambda \in \mathbb{R}$, can I generate a non-negative unbiased estimator of $f(\lambda)$?
Outline

1. Exact inference
2. Unbiased estimators and sign problem
3. Existence and non-existence results
4. Discussion
Let \( \mathcal{X} \) be a subset of \( \mathbb{R} \) and \( f : \text{conv}(\mathcal{X}) \to \mathbb{R}^+ \) a function.

**Definition**

An \( \mathcal{X} \)-algorithm \( \mathcal{A} \) is an \( f \)-factory if, given as inputs

- any i.i.d sequence \( X = (X_k)_{k \geq 1} \) with expectation \( \lambda \in \mathbb{R} \),
- an auxiliary random variable \( U \sim \text{Uniform}(0, 1) \) independent of \( (X_k)_{k \geq 1} \),

then \( Y = \mathcal{A}(U, X) \) is a non-negative unbiased estimator of \( f(\lambda) \).
Let $\mathcal{X}$ be a subset of $\mathbb{R}$.

**Definition**

An $\mathcal{X}$-algorithm $\mathcal{A}$ is a pair $(T, \varphi)$ where

- $T = (T_n)_{n \geq 0}$ is a sequence of $T_n : (0, 1) \times \mathcal{X}^n \to \{0, 1\}$,
- $\varphi = (\varphi_n)_{n \geq 0}$ is a sequence of $\varphi_n : (0, 1) \times \mathcal{X}^n \to \mathbb{R}^+$.

$\mathcal{A} \equiv (T, \varphi)$ takes $u \in (0, 1)$ and $x = (x_i)_{i \geq 1} \in \mathcal{X}^\infty$ as inputs and produces as output

- exit time: $\tau = \tau(u, x) = \inf\{n \geq 0 : T_n(u, x_1, \ldots, x_n) = 1\}$
- exit value: $\mathcal{A}(u, x) = \varphi_\tau(u, x_1, \ldots, x_\tau)$

Set $\mathcal{A}(u, x) = \infty$ if $T_n$ never gives 1.
Theorem

For any non constant function $f : \mathbb{R} \rightarrow \mathbb{R}^+$, no $f$-factory exists.

Lemma

Given i.i.d copies of an unbiased estimator of $\lambda > 0$ and a uniform random variable $U$, there is no algorithm producing a non-negative unbiased estimator of $\lambda$.

The lemma is not directly implied by the theorem but the proof is very similar.
Proof

For the sake of contradiction, introduce

- a non-constant function $f : \mathbb{R} \to \mathbb{R}^+$, and $\lambda_1, \lambda_2 \in \mathbb{R}$ with $f(\lambda_1) > f(\lambda_2)$,
- an $f$-factory $(\varphi, T)$.

Consider an i.i.d sequence $X = (X_n)_{n \geq 1}$ with expectation $\lambda_1$. Then

$$A(U, X) = \varphi_{\tau_X}(U, X_1, \ldots, X_{\tau_X})$$

has expectation $f(\lambda_1)$, and

$$\tau_X = \inf\{n : T_n(U, X_1, \ldots, X_n) = 1\}$$

is almost surely finite.
An $f$-factory should work for any input sequence. Introduce Bernoulli variables $(B_n)_{n \geq 1}$, with $\mathbb{P}(B_n = 0) = \varepsilon$ and

$$Y_n = B_n X_n + \frac{\lambda_2 - \lambda_1(1 - \varepsilon)}{\varepsilon} (1 - B_n)$$

so that $\mathbb{E}(Y_n) = \lambda_2$.

Then

$$A(U, Y) = \varphi_{\tau_Y}(U, Y_1, \ldots, Y_{\tau_Y})$$

has expectation $f(\lambda_2) < f(\lambda_1)$, and

$$\tau_Y = \inf\{n : T_n(U, Y_1, \ldots, Y_n) = 1\}$$

is almost surely finite.
Proof

By construction we can tune the probability $(1 - \varepsilon)^n$ of

$$M_n = \{(Y_1, \ldots, Y_n) = (X_1, \ldots, X_n)\},$$

by changing $\varepsilon$. On the events

$$\{(Y_1, \ldots, Y_n) \neq (X_1, \ldots, X_n)\}$$

the algorithm has to “compensate”, so that

$$f(\lambda_2) = \mathbb{E}[A(U, Y)] < \mathbb{E}[A(U, X)] = f(\lambda_1).$$

But the algorithm cannot output values lower than zero

$\Rightarrow$ for $\varepsilon$ small enough it leads to a contradiction.
Other cases

- By putting more restrictions on $\mathcal{X}$ we get different results.

- The case where $\mathcal{X} \subset \mathbb{R}^+$ and $f$ is decreasing also leads to a non-existence result.

- The case where $\mathcal{X} \subset \mathbb{R}^+$ and $f$ is increasing allows some constructions, for instance for real analytic functions.
Other cases

- No full characterisation of increasing functions allowing $f$-factories for $\mathcal{X} \subset \mathbb{R}^+$, yet.

- The case where $\mathcal{X} = [a, b]$ is related to the Bernoulli factory. Necessary and sufficient condition: $f$ continuous and there exist $n, m \in \mathbb{N}$ and $\varepsilon > 0$ such that

\[
\forall x \in [a, b] \quad f(x) \geq \varepsilon \min ((x - a)^m, (b - x)^n)
\]
Case $\mathcal{X} = [a, b]$

Assume

$$\forall x \in [a, b] \quad f(x) \geq \varepsilon \min ((x - a)^m, (b - x)^n).$$

Introduce

$$g : x \mapsto f(x) / \{(x - a)^m (b - x)^n\}$$

bounded away from zero.

Hence $g$ can be approximated from below by polynomials.

Introduce

$$P_1(x) = \sum_{(i,j) \in I_1} \alpha_{i,j}^{(1)} (x - a)^i (b - x)^j$$

with non-negative coefficients, and such that $P_1(x) \leq g(x)$.

Then approximate $g - P_1$ from below by $P_2$, $g - P_1 - P_2$ by $P_3$, etc.
Case $\mathcal{X} = [a, b]$

We obtain a sum of polynomials $\sum_{k=0}^{n} P_k(x)$ converging to $g(x)$ when $n \to \infty$.

We multiply by $(x - a)^m (b - x)^n$ to estimate $f(x)$ instead. This leads to a sequence of estimators

$$S_n = \sum_{k=0}^{n} \sum_{(i,j) \in I_k} a_{i,j}^{(k)} \left\{ \prod_{p=1}^{i} (X_p - a)^i \right\} \left\{ \prod_{q=1}^{j} (b - X_{i+q})^j \right\}$$

for which $\mathbb{P}(S_{n-1} \leq S_n) = 1$, yielding a non-negative unbiased estimator of $f(\lambda)$. 

Pierre Jacob

Non-negative unbiased estimators
Outline

1. Exact inference
2. Unbiased estimators and sign problem
3. Existence and non-existence results
4. Discussion
No $f$-factory for $\mathcal{X} = \mathbb{R}$ and any non-constant $f$.  
\[\Rightarrow\] without lower bounds on the log-likelihood estimator, no non-negative unbiased likelihood estimators.

No $f$-factory for decreasing functions $f$ and $\mathcal{X} = \mathbb{R}^+$.  
\[\Rightarrow\] without lower and upper bounds on the estimator of $C(\theta)$, no non-negative unbiased estimators of $1/C(\theta)$.

For the reference prior, it seems hopeless unless $\mathcal{X} = [a, b]$. 
Discussion

- No answer for the case $f$ “slowly” increasing and $\mathcal{X} \subseteq \mathbb{R}^+$. 

- We only considered the transformation of an unbiased estimator of $\lambda$ to an unbiased estimator of $f(\lambda)$. 

- Should we tolerate negative values and come up with appropriate methodology? 

- Should we aim for exact inference?

Thanks!
References

- On non-negative unbiased estimators, Jacob, Thiéry, 2014 (arXiv)

- Playing Russian Roulette with Intractable Likelihoods, Girolami, Lyne, Strathmann, Simpson, Atchade, 2013 (arXiv)

- Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations, Troyer, Wiese, 2005 (Phys. rev. let. 94)

- Unbiased Estimation with Square Root Convergence for SDE Models, Rhee, Glynn, 2013 (arXiv)