Non-negative unbiased estimators for exact inference

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

ICMS – June 3rd, 2014

Outline

Outline

Exact inference

For a target probability distribution with unnormalised density π, a numerical method is "exact" if for any test function φ,

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

can be approximated with arbitrary precision.

- MCMC is exact but ABC is not.
- Particle filter is exact but Extended Kalman filter is not.

Exact inference

- Using exact methods, no systematic error. . .
- . . . but no guarantee that for a fixed computational budget, exact methods should be preferred over approximate methods.
- Settings for which exact methods are available?

Exact inference using Metropolis-Hastings

Assume we can compute the target density π pointwise.

Set some $\theta^{(1)}$.
for $i = 2$ to N_θ **do**
    Propose $\theta^\star \sim q(\cdot|\theta^{(i-1)})$.
    Compute the ratio:

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

    Set $\theta^{(i)} = \theta^\star$ with probability α, otherwise set $\theta^{(i)} = \theta^{(i-1)}$.
end for

Exact inference with unbiased estimators

Assume for each θ, 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_θ **do**
3: Propose $\theta^\star \sim q(\cdot|\theta^{(i-1)})$ and sample $Z(\theta^\star)$.
4: Compute the ratio:

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

5: Set $\theta^{(i)} = \theta^\star$, $Z(\theta^{(i)}) = Z(\theta^\star)$ with probability α, otherwise set $\theta^{(i)} = \theta^{(i-1)}$, $Z(\theta^{(i)}) = Z(\theta^{(i-1)})$.
6: **end for**

Pseudo-marginal approach, Andrieu & Roberts, 2009.

Exact inference with unbiased estimators

- Game-changer when one has access to efficient non-negative unbiased estimators of the target density, pointwise.
- Example: particle MCMC methods for hidden Markov models, based on particle filters to estimate the likelihood. (Andrieu, Doucet, Holenstein, 2010.)
- Is there a generic and efficient pseudo-marginal approach to sample from intractable distributions?

Exact inference with 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 ever computing the likelihood?

- Unbiased estimator of the *log-likelihood*

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

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

- It doesn't directly provide an unbiased estimator of the *likelihood*.

- Hence the pseudo-marginal approach is not applicable.

Exact inference with unbiased estimators

Doubly intractable distributions

Posterior density decomposable into

$$\pi(\theta \mid y) = \frac{1}{C(\theta)} p(y;\theta).$$

- One can typically get an unbiased estimator of $C(\theta)$ using importance sampling.
- It doesn't directly provide an unbiased estimator of $1/C(\theta)$.
- Hence the pseudo-marginal approach is not applicable.

Outline

Unbiased estimators

Von Neumann & Ulam (~ 1950), Kuti (~ 1980), Rychlik (~ 1990), McLeish (~ 2010), **Rhee & Glynn** ($\sim 2012, 2013$).

Removing the bias of consistent estimators

Introduce

- a random variable S with $\mathbb{E}(S) = \lambda \in \mathbb{R}$,
- a sequence $(S_n)_{n\geq 0}$ converging to S in L^2,
- N be an integer-valued random variable.

Then

$$Y = \sum_{n=0}^{N} \frac{1}{\mathbb{P}(N \geq n)} \times (S_n - S_{n-1})$$

where $S_{-1} = 0$, has expectation $\mathbb{E}(Y) = \mathbb{E}(S) = \lambda$.

Unbiased estimators

- If
$$\sum_{n=1}^{\infty} \frac{1}{\mathbb{P}(N \geq n)} \times \mathbb{E}(\,|S - S_{n-1}|^2\,) < \infty, \tag{1}$$
then the variance of Y is finite.

- Let $\bar{t}_n$ be the incremental expected time to compute S_n. Then the expected computing time of Y should preferably satisfy
$$\mathbb{E}(\bar{\tau}) = \sum_{n=0}^{\infty} \mathbb{P}(N \geq n) \times \bar{t}_n < \infty. \tag{2}$$

Debiasing MCMC estimators?

- A typical MCMC estimate S_n based on n draws would satisfy

$$\mathbb{E}\left(|S_n - S|^2\right) = \mathcal{O}(n^{-1}).$$

- Hence a finite variance of the debiased estimator requires selecting N such that

$$\sum_{n=1}^{\infty} \frac{1}{\mathbb{P}(N \geq n)} \times \frac{1}{n} < \infty$$

- The incremental cost of S_n given S_{n-1} is constant $= 1$. Hence a finite expected cost requires:

$$\mathbb{E}(\bar{\tau}) = \sum_{n=0}^{\infty} \mathbb{P}(N \geq n) < \infty.$$

Tricky!

Dealing with negative values

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

$$Y = \sum_{n=0}^{N} \frac{1}{\mathbb{P}(N \geq n)} \times (S_n - S_{n-1}),$$

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 using the absolute values of the estimators, noting

$$\int \varphi(x)g(x)dx = \frac{\int \varphi(x)\sigma(g(x))|g(x)|dx}{\int \sigma(g(x))|g(x)|dx}$$

which amounts to transferring the sign to the test function.

- This deteriorates with the dimension and is called the sign problem in lattice quantum chromodynamics.

Troyer & Wiese, 2005.
Application to the pseudo-marginal approach by Girolami, Lyne, Strathmann, Simpson, Atchade, 2013.

Dealing with negative values

- Can we remove the bias of consistent estimators AND guarantee almost sure non-negativity?

- Given an unbiased estimator of $\lambda > 0$, is there an algorithm producing non-negative unbiased estimators of λ?

- Let f be any function $f : \mathbb{R} \to \mathbb{R}^+$. Given an unbiased estimator of $\lambda \in \mathbb{R}$, is there an algorithm producing non-negative unbiased estimators of $f(\lambda)$?

  $f : x \mapsto \exp(x)$: big data.
  $f : x \mapsto 1/x$: doubly intractable distributions.

Outline

$\mathcal{X}$-algorithm

Let $\mathcal{X}$ be $\mathbb{R}$ or a subset of $\mathbb{R}$.

Definition

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

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

T specifies the rule to terminate the algorithm.

$$
\begin{aligned}
T_1(u, x_1) &= 0 \\
T_2(u, x_1, x_2) &= 0 \\
&\dots \\
T_\tau(u, x_1, \dots, x_\tau) &= 1.
\end{aligned}
$$

$\mathcal{X}$-algorithm

Exit time: $\tau = \inf\{n \geq 1 : T_n(u, x_1, \ldots, x_n) = 1\}$

When the algorithm terminates it returns the value given by φ_τ:

$$\mathcal{A}(u, x) = \varphi_\tau(u, x_1, \ldots, x_\tau).$$

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

Note that by definition if the output has an expectation then $\mathcal{A}$ terminates in finite time with probability one.

f-factory

Let $\mathcal{X}$ be $\mathbb{R}$ or a subset of $\mathbb{R}$ and $f : \mathrm{conv}(\mathcal{X}) \to \mathbb{R}^+$ a function.

Definition

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

- any i.i.d sequence $X = (X_n)_{n\geq 1}$ on $\mathcal{X}$ with mean $\lambda \in \mathbb{R}$,
- an auxiliary r.v. $U \sim \mathrm{Uniform}(0, 1)$ independent of $(X_n)_{n\geq 1}$,

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

General non-existence of f-factories

Here $\mathcal{X} = \mathbb{R}$.

Theorem

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

Consequences:

- given an unbiased estimator of $\lambda > 0$, impossible to generate a non-negative unbiased estimator of λ,
- given an unbiased estimator of $\lambda \in \mathbb{R}$, impossible to generate a non-negative unbiased estimator of $\exp(\lambda)$ or $1/\lambda$.

In other words, we need to "know more".

Proof

For the sake of contradiction, introduce

- a function $f : \mathbb{R} \to \mathbb{R}^+$, and $\lambda_X, \lambda_Y \in \mathbb{R}$ with $f(\lambda_X) > f(\lambda_Y)$,
- an f-factory (φ, T).

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

Then

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

is almost surely finite and

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

has expectation $f(\lambda_X)$.

Proof

Introduce Bernoulli variables $(B_n)_{n\geq 1}$, with $\mathbb{P}(B_n = 0) = \varepsilon$ and

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

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

Then

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

is almost surely finite and

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

has expectation $f(\lambda_Y) < f(\lambda_X)$.

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 ε. On the events

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

the algorithm has to "compensate" for the difference between $f(\lambda_Y)$ and $f(\lambda_X)$:

$$f(\lambda_Y) = \mathbb{E}[\mathcal{A}(U, Y)] < \mathbb{E}[\mathcal{A}(U, X)] = f(\lambda_X).$$

But the algorithm cannot output values lower than zero
$\Rightarrow$ for ε small enough it leads to a contradiction.

Other cases

- If we know more about the unbiased estimators used as input, we obtain different results.
- For the case where $\mathcal{X} \subset \mathbb{R}^+$ (lower bound on the estimator):
  - if f is decreasing, same non-existence result;
  - if f is increasing and real analytic, available constructions (Poisson estimator);
  - if f is increasing in general, no result yet.

Other cases

- The case where $\mathcal{X} = [a, b]$ (lower and upper bound on the estimator) 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

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

Bernoulli factory (Keane & O'Brien, 1994).

Outline

Back to statistics

- No f-factory for $\mathcal{X} = \mathbb{R}$ and any non-constant f.

  Big Data $\Rightarrow$ without lower bounds on the log-likelihood estimator, no non-negative unbiased likelihood estimators.

- No f-factory for decreasing functions f and $\mathcal{X} = \mathbb{R}^+$.

  Doubly intractable $\Rightarrow$ without lower and upper bounds on the estimator of $C(\theta)$, no non-negative unbiased estimators of $1/C(\theta)$.

Discussion

- We conjecture that the only functions $f : [a, +\infty) \to \mathbb{R}^+$ for which an f-factory exists are of the form

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

with non-negative coefficients $c_n \geq 0$.

- Should we tolerate negative values and come up with appropriate methodology?
- Is there any other trick than the "absolute value" trick?
- Should we aim for exact inference anyway?

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)