On non-negative unbiased estimators

based on Jacob & Thiéry, arXiv:1309.6473

Dino Sejdinovic

Gatsby Unit, UCL

April 25, 2014

Non-negative unbiased estimators

- Estimators of non-negative quantities (distances, probabilities) that are **unbiased** and themselves **non-negative**.
- Non-negativity constraint: want to plug in estimates into *exact approximate* algorithms, e.g., simulate an event with estimated probability
- A generic problem: have unbiased estimators of Z, but need an unbiased estimator of $f(Z) \geq 0$.

Context: Pseudo-Marginal MCMC

- parameters θ, latent process $\mathbf{F}$, observations $\mathbf{y}$ with

$$p(\theta, \mathbf{F}, \mathbf{y}) = p(\theta)p(\mathbf{F}|\theta)p(\mathbf{y}|\mathbf{F}, \theta)$$

Context: Pseudo-Marginal MCMC

- parameters θ, latent process $\mathbf{F}$, observations $\mathbf{y}$ with

$$p(\theta, \mathbf{F}, \mathbf{y}) = p(\theta)p(\mathbf{F}|\theta)p(\mathbf{y}|\mathbf{F}, \theta)$$

- Interested in posterior

$$p(\theta|\mathbf{y}) = \frac{g(\theta; \mathbf{y})}{Z(\mathbf{y})} = \frac{p(\theta)p(\mathbf{y}|\theta)}{Z(\mathbf{y})}$$

Context: Pseudo-Marginal MCMC

- parameters θ, latent process $\mathbf{F}$, observations $\mathbf{y}$ with

$$p(\theta, \mathbf{F}, \mathbf{y}) = p(\theta)p(\mathbf{F}|\theta)p(\mathbf{y}|\mathbf{F}, \theta)$$

- Interested in posterior

$$p(\theta|\mathbf{y}) = \frac{g(\theta; \mathbf{y})}{Z(\mathbf{y})} = \frac{p(\theta)p(\mathbf{y}|\theta)}{Z(\mathbf{y})} = \frac{p(\theta)\int p(\mathbf{F}|\theta)p(\mathbf{y}|\mathbf{F}, \theta)d\mathbf{F}}{Z(\mathbf{y})}$$

- Often impossible to integrate out the latent process $\mathbf{F}$, i.e., unable to compute marginal likelihood $p(\mathbf{y}|\theta)$

Context: Pseudo-Marginal MCMC (2)

- Unable to compute correct Metropolis-Hasting acceptance probabilities:

$$\alpha(\theta, \theta') = \min\{1, \frac{g(\theta'; \mathbf{y}) q(\theta|\theta')}{g(\theta; \mathbf{y}) q(\theta'|\theta)}\}$$

Context: Pseudo-Marginal MCMC (2)

- Unable to compute correct Metropolis-Hasting acceptance probabilities:

$$\alpha(\theta, \theta') = \min\{1, \frac{g(\theta'; \mathbf{y}) q(\theta|\theta')}{g(\theta; \mathbf{y}) q(\theta'|\theta)}\}$$

- However, in some situations, we can obtain an unbiased Monte Carlo estimate of $p(\mathbf{y}|\theta)$ and thus of $g(\theta; \mathbf{y})$, e.g., by importance sampling the latent process:

$$\hat{p}(\mathbf{y}|\theta) = \sum_{j=1}^{m} p(\mathbf{y}|\mathbf{F}_j, \theta) \frac{p(\mathbf{F}_j|\theta)}{Q(\mathbf{F}_j)}$$

Context: Pseudo-Marginal MCMC (2)

- Unable to compute correct Metropolis-Hasting acceptance probabilities:

$$\alpha(\theta, \theta') = \min\{1, \frac{\hat{g}(\theta'; \mathbf{y}) q(\theta|\theta')}{\hat{g}(\theta; \mathbf{y}) q(\theta'|\theta)}\}$$

- However, in some situations, we can obtain an unbiased Monte Carlo estimate of $p(\mathbf{y}|\theta)$ and thus of $g(\theta; \mathbf{y})$, e.g., by importance sampling the latent process:

$$\hat{p}(\mathbf{y}|\theta) = \sum_{j=1}^{m} p(\mathbf{y}|\mathbf{F}_j, \theta) \frac{p(\mathbf{F}_j|\theta)}{Q(\mathbf{F}_j)}$$

- Remarkably, plugging in the unbiased estimates still leads to the correct invariant distribution $p(\theta|\mathbf{y})$ (Beaumont, 2003; Andrieu & Roberts, 2009)

Passing the unbiased estimator through a non-linearity

- In the latent process model, we were able to *unbiasedly and non-negatively* estimate $p(\theta)p(\mathbf{y}|\theta)$ directly, so can just plug in.

Passing the unbiased estimator through a non-linearity

- In the latent process model, we were able to *unbiasedly and non-negatively* estimate $p(\theta)p(\mathbf{y}|\theta)$ directly, so can just plug in.
- Often, $p(\theta)p(\mathbf{y}|\theta) = g(\theta; \mathbf{y}) f(Z(\theta; \mathbf{y}))$. Maybe we can have an unbiased estimator of $Z(\theta; \mathbf{y})$ but is that good for anything?

Example: Austerity in MCMC Land (Korattikara, Chen & Welling, 2014)

- "In today's Big Data world, we need to rethink our Bayesian inference algorithms."
- $\mathbf{y}$ is Big$^{\text{TM}}$: too expensive to compute $\log p(\mathbf{y}|\theta) = \sum_{i=1}^{n} \log p(y_i|\theta)$ so compute $\frac{n}{t}\sum_{j=1}^{t} \log p(y_j^*|\theta)$ instead with $t \ll n$.
- Gives us an unbiased estimator of $Z(\theta;\mathbf{y}) = \log p(\mathbf{y}|\theta)$, i.e., $f(z) = e^z$. **Can we transform it to an unbiased and non-negative estimator of $p(\mathbf{y}|\theta)$?**

Debiasing (Mc Leash, 2010; Rhee & Glynn 2012)

- True S, s.t. $\mathbb{E}_\pi S = \lambda$ parameter of interest
- A sequence of biased estimators $\{S_n\}_{n=0}^{\infty}$, with $\lim_{n\to\infty} \mathbb{E}_\pi [S_n] = \mathbb{E}_\pi [S] = \lambda$
- Assume: $\sum_{n=0}^{\infty} \mathbb{E}_\pi |S_n - S_{n-1}| < \infty$ **OR** $S_n \geq S_{n-1}$ a.s.

Debiasing (Mc Leash, 2010; Rhee & Glynn 2012)

- True S, s.t. $\mathbb{E}_\pi S = \lambda$ parameter of interest
- A sequence of biased estimators $\{S_n\}_{n=0}^{\infty}$, with $\lim_{n\to\infty} \mathbb{E}_\pi [S_n] = \mathbb{E}_\pi [S] = \lambda$
- Assume: $\sum_{n=0}^{\infty} \mathbb{E}_\pi |S_n - S_{n-1}| < \infty$ **OR** $S_n \geq S_{n-1}$ a.s.

Lemma

Let T be an integer-valued random variable (independent of everything else) with $\mathbb{P}[T \geq n] > 0\ \forall n$. Then

$$S_T^* = \sum_{n=0}^{T} \frac{S_n - S_{n-1}}{\mathbb{P}[T \geq n]}$$

is unbiased.

Debiasing

- Assume: $\sum_{n=0}^{\infty} \mathbb{E}_\pi |S_n - S_{n-1}| < \infty$ **OR** $S_n \geq S_{n-1}$ a.s.

Proof.

$$\begin{aligned}
\mathbb{E}S_T^* &= \mathbb{E}\sum_{n=0}^{\infty} \frac{\mathbf{1}_{\{T \geq n\}}}{\mathbb{P}[T \geq n]}(S_n - S_{n-1}) \\
&= \sum_{n=0}^{\infty} \frac{\mathbb{E}_T \mathbf{1}_{\{T \geq n\}}}{\mathbb{P}[T \geq n]} \mathbb{E}_\pi (S_n - S_{n-1}) \\
&= \sum_{n=0}^{\infty} (\mathbb{E}_\pi S_n - \mathbb{E}_\pi S_{n-1}) \\
&= \lambda.
\end{aligned}$$

□

Debiasing

- Assume: $\sum_{n=0}^{\infty} \mathbb{E}_\pi \left| S_n - S_{n-1} \right| < \infty$ **OR** $S_n \geq S_{n-1}$ a.s.

Proposition

Let T be an integer-valued random variable (independent of everything else) with $\mathbb{P}\left[T \geq n\right] > 0 \ \forall n$. Then

$$S_T^* = \sum_{n=0}^{T} \frac{S_n - S_{n-1}}{\mathbb{P}\left[T \geq n\right]}$$

is unbiased.

- Achtung! $\frac{1}{\mathbb{P}[T \geq n]} \to \infty$, so variance *can be infinite*. Need $\mathbb{E}\left[(S - S_n)^2\right] \to 0$ faster.

Debiasing

- Assume: $\sum_{n=0}^{\infty} \mathbb{E}_\pi |S_n - S_{n-1}| < \infty$ **OR** $S_n \geq S_{n-1}$ a.s.

Proposition

Let T be an integer-valued random variable (independent of everything else) with $\mathbb{P}[T \geq n] > 0\ \forall n$. Then

$$S_T^* = \sum_{n=0}^{T} \frac{S_n - S_{n-1}}{\mathbb{P}[T \geq n]}$$

is unbiased.

- Achtung! $\frac{1}{\mathbb{P}[T \geq n]} \to \infty$, so variance *can be infinite*. Need $\mathbb{E}\left[(S - S_n)^2\right] \to 0$ faster.
- Lebensgefahr! Even if all $S_n \geq 0$ a.s., S_T^* *can be negative*. Fine if $S_n \geq S_{n-1}$ a.s.

Example: Russian Roulette (Girolami et al, 2013)

- In this case: $p(\theta)p(\mathbf{y}|\theta) = \frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})}$, i.e., $f(z) = 1/z$. Introduce an auxiliary variable $v \sim Exp\left(Z(\theta;\mathbf{y})\right)$.

$$p(\theta, v|\mathbf{y}) \quad \propto \quad Z(\theta;\mathbf{y})e^{-vZ(\theta;\mathbf{y})}\frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})} = e^{-vZ(\theta;\mathbf{y})}g(\theta;\mathbf{y})$$

Example: Russian Roulette (Girolami et al, 2013)

- In this case: $p(\theta)p(\mathbf{y}|\theta) = \frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})}$, i.e., $f(z) = 1/z$. Introduce an auxiliary variable $v \sim Exp\left(Z(\theta;\mathbf{y})\right)$.

$$p(\theta, v|\mathbf{y}) \quad \propto \quad Z(\theta;\mathbf{y})e^{-vZ(\theta;\mathbf{y})}\frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})} = e^{-vZ(\theta;\mathbf{y})}g(\theta;\mathbf{y})$$

- Taylor expand $e^{-vZ} = \sum_{k=0}^{\infty} \frac{(-v)^k}{k!} Z^k$ and use biased but asymptotically unbiased estimators of $e^{-vZ(\theta;\mathbf{y})}$:

$$S_n = \sum_{k=0}^{n} \frac{(-v)^k}{k!} \prod_{i=1}^{k} \hat{Z}_i(\theta;\mathbf{y})$$

Example: Russian Roulette (Girolami et al, 2013)

- In this case: $p(\theta)p(\mathbf{y}|\theta) = \frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})}$, i.e., $f(z) = 1/z$. Introduce an auxiliary variable $v \sim Exp\left(Z(\theta;\mathbf{y})\right)$.

$$p(\theta, v|\mathbf{y}) \quad \propto \quad Z(\theta;\mathbf{y})e^{-vZ(\theta;\mathbf{y})}\frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})} = e^{-vZ(\theta;\mathbf{y})}g(\theta;\mathbf{y})$$

- Taylor expand $e^{-vZ} = \sum_{k=0}^{\infty} \frac{(-v)^k}{k!} Z^k$ and use biased but asymptotically unbiased estimators of $e^{-vZ(\theta;\mathbf{y})}$:

$$S_n = \sum_{k=0}^{n} \frac{(-v)^k}{k!} \prod_{i=1}^{k} \hat{Z}_i(\theta;\mathbf{y})$$

- Apply debiasing lemma. We turned a sequence of i.i.d. unbiased estimators $\left\{\hat{Z}_i(\theta;\mathbf{y})\right\}_{i\geq 1}$ of $Z(\theta;\mathbf{y})$ into an unbiased estimator of $e^{-vZ(\theta;\mathbf{y})}$. Happy days!

Example: Russian Roulette (Girolami et al, 2013)

- In this case: $p(\theta)p(\mathbf{y}|\theta) = \frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})}$, i.e., $f(z) = 1/z$. Introduce an auxiliary variable $v \sim Exp\left(Z(\theta;\mathbf{y})\right)$.

$$p(\theta, v|\mathbf{y}) \quad \propto \quad Z(\theta;\mathbf{y})e^{-vZ(\theta;\mathbf{y})}\frac{g(\theta;\mathbf{y})}{Z(\theta;\mathbf{y})} = e^{-vZ(\theta;\mathbf{y})}g(\theta;\mathbf{y})$$

- Taylor expand $e^{-vZ} = \sum_{k=0}^{\infty} \frac{(-v)^k}{k!} Z^k$ and use biased but asymptotically unbiased estimators of $e^{-vZ(\theta;\mathbf{y})}$:

$$S_n = \sum_{k=0}^{n} \frac{(-v)^k}{k!} \prod_{i=1}^{k} \hat{Z}_i(\theta;\mathbf{y})$$

- Apply debiasing lemma. We turned a sequence of i.i.d. unbiased estimators $\left\{\hat{Z}_i(\theta;\mathbf{y})\right\}_{i\geq 1}$ of $Z(\theta;\mathbf{y})$ into an unbiased estimator of $e^{-vZ(\theta;\mathbf{y})}$. Happy days!
- Errm, but $S_n - S_{n-1}$ can be negative. **Is it possible to ensure non-negativity?**

Formal definition of an algorithm

- Input:
  - A sequence $\mathbf{X} = \{X_k\}_{k\geq 1}$ of $\mathcal{X}$-valued r.v's marginally following identical law π and $\mathbb{E}_\pi X = Z$ (e.g., unbiased estimators of $Z(\theta; \mathbf{y})$)
  - Auxiliary source of randomness $U \sim \text{Uniform}(0, 1)$

Formal definition of an algorithm

- Input:
  - A sequence $\mathbf{X} = \{X_k\}_{k \geq 1}$ of $\mathcal{X}$-valued r.v's marginally following identical law π and $\mathbb{E}_\pi X = Z$ (e.g., unbiased estimators of $Z(\theta; \mathbf{y})$)
  - Auxiliary source of randomness $U \sim \text{Uniform}(0, 1)$

- Ingredients of the algorithm:
  - A sequence of functions $T_n : (0, 1) \times \mathcal{X}^n \to \{0, 1\}$ (1 is the stopping criterion for algorithm)
  - A sequence of functions $\varphi_n : (0, 1) \times \mathcal{X}^n \to \mathbb{R}^+$

Formal definition of an algorithm

- Input:
  - A sequence $\mathbf{X} = \{X_k\}_{k \geq 1}$ of $\mathcal{X}$-valued r.v's marginally following identical law π and $\mathbb{E}_\pi X = Z$ (e.g., unbiased estimators of $Z(\theta; \mathbf{y})$)
  - Auxiliary source of randomness $U \sim \mathrm{Uniform}(0, 1)$
- Ingredients of the algorithm:
  - A sequence of functions $T_n : (0, 1) \times \mathcal{X}^n \to \{0, 1\}$ (1 is the stopping criterion for algorithm)
  - A sequence of functions $\varphi_n : (0, 1) \times \mathcal{X}^n \to \mathbb{R}^+$
- Output:

$$\mathcal{A}(U, \mathbf{X}) = \varphi_\tau (u, x_1, \ldots, x_\tau), \quad \tau = \inf \{n \geq 0 \ : T_n (u, x_1, \ldots, x_n) = 1\}$$

Formal definition of an algorithm

- Input:
  - A sequence $\mathbf{X} = \{X_k\}_{k\geq 1}$ of $\mathcal{X}$-valued r.v's marginally following identical law π and $\mathbb{E}_\pi X = Z$ (e.g., unbiased estimators of $Z(\theta; \mathbf{y})$)
  - Auxiliary source of randomness $U \sim \text{Uniform}(0,1)$
- Ingredients of the algorithm:
  - A sequence of functions $T_n : (0,1) \times \mathcal{X}^n \to \{0,1\}$ (1 is the stopping criterion for algorithm)
  - A sequence of functions $\varphi_n : (0,1) \times \mathcal{X}^n \to \mathbb{R}^+$
- Output:

$$\mathcal{A}(U, \mathbf{X}) = \varphi_\tau (u, x_1, \ldots, x_\tau), \quad \tau = \inf\{n \geq 0 \ : T_n(u, x_1, \ldots, x_n) = 1\}$$

- For $f : \mathcal{X} \to \mathbb{R}^+$, we say there exists an $(f, \mathcal{X})$-algorithm if τ is finite a.s. and $\mathcal{A}(U, \mathbf{X})$ is an unbiased estimator of $f(Z)$.

Negative results

Lemma

If $f : \mathbb{R} \to \mathbb{R}^+$ *is not constant, no* $(f, \mathbb{R})$*-algorithm exists.*

Lemma

If $f : [a, \infty) \to \mathbb{R}^+$ *is continuous, and* $(f, [a, \infty))$*-algorithm exists,* f *is non-decreasing.*

Lemma

If $f : (-\infty, b] \to \mathbb{R}^+$ *is continuous, and* $(f, (-\infty, b])$*-algorithm exists,* f *is non-increasing.*

Proof

- Idea: Construct $\mathbf{X} = \{X_k\}_{k\geq 1}$ and $\mathbf{Y} = \{Y_k\}_{k\geq 1}$ with different means but which agree in almost all terms.

Proof

- Idea: Construct $\mathbf{X} = \{X_k\}_{k\geq 1}$ and $\mathbf{Y} = \{Y_k\}_{k\geq 1}$ with different means but which agree in almost all terms.
- Take z_1 and z_2 s.t. $f(z_1) > f(z_2)$. Let $\mathbb{E}X = z_1$. Take $\varepsilon > 0$ and $\{B_k\}_{k\geq 1} \overset{i.i.d.}{\sim} \text{Bern}(1-\varepsilon)$, and set

$$Y_k = B_k X_k + (1 - B_k)\frac{z_2 - z_1(1-\varepsilon)}{\varepsilon}$$

so that $\mathbb{E}Y = z_2$.

Proof

- Idea: Construct $\mathbf{X} = \{X_k\}_{k\geq 1}$ and $\mathbf{Y} = \{Y_k\}_{k\geq 1}$ with different means but which agree in almost all terms.
- Take z_1 and z_2 s.t. $f(z_1) > f(z_2)$. Let $\mathbb{E}X = z_1$. Take $\varepsilon > 0$ and $\{B_k\}_{k\geq 1} \overset{i.i.d.}{\sim} \text{Bern}(1-\varepsilon)$, and set

$$Y_k = B_k X_k + (1 - B_k)\frac{z_2 - z_1(1-\varepsilon)}{\varepsilon}$$

  so that $\mathbb{E}Y = z_2$.
- Assume $\mathcal{A}(U, \mathbf{X}) = z_1$ and $\mathcal{A}(U, \mathbf{Y}) = z_2$. Recall the stopping time:

$$\tau_X = \inf\{n \geq 0 \ : \ T_n(u, x_1, \ldots, x_n) = 1\}$$

Proof

- Define events $M_n = \{B_1 = \ldots = B_n = 1\}$, $L_n = \{\tau_X \leq n\}$. Clearly, $\mathcal{A}(U, \mathbf{X})\mathbf{1}_{M_n \cap L_n} = \mathcal{A}(U, \mathbf{Y})\mathbf{1}_{M_n \cap L_n}$. Pick $\delta < f(z_1) - f(z_2)$.

Proof

- Define events $M_n = \{B_1 = \ldots = B_n = 1\}$, $L_n = \{\tau_X \le n\}$. Clearly, $\mathcal{A}(U, \mathbf{X})\mathbf{1}_{M_n \cap L_n} = \mathcal{A}(U, \mathbf{Y})\mathbf{1}_{M_n \cap L_n}$. Pick $\delta < f(z_1) - f(z_2)$.

$$
\begin{array}{rcl}
f(z_2) & = & \mathbb{E}\left[\mathcal{A}(U, \mathbf{Y})\right] \\
\text{(since } \mathcal{A} \text{ is non-negative)} & \ge & \mathbb{E}\left[\mathcal{A}(U, \mathbf{Y})\mathbf{1}_{M_n \cap L_n}\right] \\
(X\text{'s and } Y\text{'s are the same before stopping)} & = & \mathbb{E}\left[\mathcal{A}(U, \mathbf{X})\mathbf{1}_{M_n \cap L_n}\right] \\
(\{B_n\} \text{ are independent of everything)} & = & (1-\varepsilon)^n \mathbb{E}\left[\mathcal{A}(U, \mathbf{X})\mathbf{1}_{L_n}\right] \\
\text{(for } n = n(\delta)\text{large enough since } \lim \mathbb{P}\left(L_n\right) = 1) & > & (1-\varepsilon)^n \left(f(z_1) - \delta\right) \\
\text{(for } \varepsilon \text{ small enough since } f(z_1) > f(z_2)) & > & f(z_2).
\end{array}
$$

Positive results

Lemma

If $f : [a, \infty) \to \mathbb{R}^+$ can be expressed as $f(x) = \sum_{k=0}^{\infty} c_k (x-a)^k$, with $c_k \geq 0$, then $(f, [a, \infty))$-algorithm exists.

Proof.

Simply use debiasing lemma on

$$S_n = \sum_{k=0}^{n} c_k \prod_{i=1}^{k} (X_i - a),$$

where $S_{n+1} \geq S_n$ a.s. $\square$

Positive results

Lemma

Let $f : [a, b] \to \mathbb{R}^+$ *be continuous such that* $\exists m, n \in \mathbb{N}$ *and* $\delta > 0$*, s.t.*

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

Then $(f, [a, b])$*-algorithm exists.*

Proof.

Since $f(x)/((x-a)^m (b-x)^n)$ is bounded away from zero on (a, b), can approximate it arbitrarily well from below in terms of Bernstein polynomials with non-negative coefficients. Then apply debiasing lemma to these approximations. □

Positive results

Lemma

Let $f : [a,b] \to \mathbb{R}^+$ be continuous such that $\exists m, n \in \mathbb{N}$ and $\delta > 0$, s.t.

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

Then $(f, [a,b])$-algorithm exists.

Proof.

Since $f(x)/((x-a)^m (b-x)^n)$ is bounded away from zero on (a,b), can approximate it arbitrarily well from below in terms of Bernstein polynomials with non-negative coefficients. Then apply debiasing lemma to these approximations. $\square$

Can be viewed as a consequence of the Bernoulli factory theorem from (Keane & O'Brien 1994).

Positivation Lemma

Lemma

Let $a < 0 < b$ and assume that $\mathbf{X} = \{X_k\}_{k \geq 1}$ are $[a, b]$-valued. If $\mathbb{E}_\pi X = Z$ is bounded away from zero, i.e., $Z > \eta > 0$, there exists an algorithm for which $\mathcal{A}(U, \mathbf{X})$ is a non-negative unbiased estimator of Z.

Proof.

It's an algorithm with $f(z) = \max(\eta, z)$ which satisfies the polynomial lower bound. And obviously $f(Z) = Z$. □

Summary

- Non-negative unbiased estimators of $f(\mathbb{E}X)$ for a non-constant f based on X-samples are impossible without domain restrictions on X.
- We can get exact approximate austerity in MCMC if and only if we can bound $\log p(\mathbf{y}|\theta)$ from below (!)
- Unbiased estimators of positive quantities bounded away from zero can be positivised.
- Close relation to the *Bernoulli Factory* problem: get an $f(p)$ coin from a p-coin.