Supplementary Materials for
One Pseudo-Sample is Enough
in Approximate Bayesian Computation MCMC

BY LUKE BORNN, NATESH PILLAI
Department of Statistics, Harvard University
bornn@stat.harvard pillai@stat.harvard.edu

AARON SMITH
Department of Mathematics and Statistics, University of Ottawa
smith.aaron.matthew@gmail.com

AND DAWN WOODARD
Department of Operations Research and Information Engineering, Cornell University
woodard@cornell.edu

1. PROOFS

Proof of Theorem 2. For a reversible Markov kernel \( H \) with stationary distribution \( \mu \) and \( g \in L^2(\mu) \), we follow Tierney (1998, proof of Theorem 4) by writing the asymptotic variance as

\[
v(g, H) = \lim_{n \to \infty} \int_{-1}^{1} \left( 1 + 2 \sum_{i=1}^{n} \frac{n-i}{n} x^i \right) e_{g,H}(dx) = \int_{-1}^{1} \frac{1+x}{1-x} e_{g,H}(dx)
\]

(1)

where \( e_{g,H} \) is the spectral measure associated with \( g \) and \( H \), i.e., the finite positive measure on \([-1, 1]\) such that \( \langle g, H^i g \rangle_{\mu} = \int_{-1}^{1} x^i e_{g,H}(dx) \) for all \( i \in \mathbb{N} \). For nonnegative definite transition kernels \( H \) the spectral measure, and the integrals, are over the narrower range \([0, 1]\).

Denote by \( H_{2,\alpha} \) the transition kernel of the pseudo-marginal algorithm with proposal kernel \( q \), target marginal distribution \( \mu \), and estimator \( T_{2,x,\alpha} \) of the unnormalized target. We note that

\[
H_{2,\alpha} = \alpha I + (1 - \alpha) H_2,
\]

(2)

where \( I \) is the identity operator. If \( T_{1,x} \leq_{\text{ex}} T_{2,x,\alpha} \) then by Theorem 3 of Andrieu & Vihola (2014),

\[
v(f, H_1) \leq v(f, H_{2,\alpha})
\]

(3)

for any \( f \in L^2(\mu) \). Using (1)-(2) and the nonnegative-definiteness of \( H_2 \),

\[
v(f, H_{2,\alpha}) = \lim_{n \to \infty} \int_{-1}^{1} \left( 1 + 2 \sum_{i=1}^{n} \frac{n-i}{n} x^i \right) e_{f,H_{2,\alpha}}(dx)
\]

\[
= \lim_{n \to \infty} \left[ \langle f, f \rangle_{\mu} + 2 \sum_{i=1}^{n} \frac{n-i}{n} \langle f, H_{2,\alpha}^i f \rangle_{\mu} \right]
\]
by Theorem 2 it is sufficient to take condition (4) for normalized target \( \pi \). Combining this with (3) yields the desired result.  

\[ \int_{-1}^{1} \left( 1 + 2 \sum_{i=1}^{n} \left( \frac{n-i}{n} (\alpha + (1-\alpha)x)^i \right) e_{f,H_2}(dx) \right) = \frac{1 + \alpha}{1 - \alpha} \int_{0}^{1} \frac{1 + x}{1 - \alpha} e_{f,H_2}(dx) \]

\[ = \frac{1 + \alpha}{1 - \alpha} \nu(f,H_2). \]

Combining this with (3) yields the desired result.  

**Proof of Proposition 3.** For any \( M \geq 1 \), let \( T_{M,\theta} \) be the estimator \( \hat{\pi}_{K,M}(\theta | y_{\text{obs}}) \) of the un-normalized target \( \pi_K \), so that \( T_{M,\theta} \) is \( T_{M,\theta} \) handicapped by \( \alpha \) as defined in (5). To obtain (6), by Theorem 2 it is sufficient to take \( \alpha = 1 - \frac{1}{M} \) and show that \( T_{1,\theta} \leq c \) \( T_{M,\theta,\alpha} \). By Proposition 2 of Leskelä & Vihola (2014), it is furthermore sufficient to show that, for all \( c \in \mathbb{R} \),

\[
\mathbb{E}[|T_{1,\theta} - c|] \leq \mathbb{E}[|T_{M,\theta,\alpha} - c|]. \tag{4}
\]

Let \( \text{Bin}(n, q) \) denote the binomial distribution with \( n \) trials and success probability \( q \). For a given point \( \Theta \subset \Theta \), let \( p = p(\theta) = \mathbb{P}[T_{1,\theta} \neq 0] = \int \mathbb{1}_{\{\|y_{\text{obs}}\| - \nu(y) \mid \leq c\}} p(y(\theta)) dy. \) Noting that \( T_{1,\theta} \in \{0, 1\} \), we may then write \( T_{1,\theta} \) and \( T_{M,\theta,\alpha} \) as the following mixtures:

\[
\frac{T_{1,\theta}}{\pi(\theta)} \begin{cases} 
D = \text{Bin}(1, p) 
\end{cases}
\]

\[
\frac{T_{M,\theta,\alpha}}{\pi(\theta)} \begin{cases} 
D = \frac{M-1}{M} \delta_0 + \frac{1}{M} \text{Bin}(M, p). 
\end{cases}
\]

where \( \delta_0 \) is the unit point mass at zero. Denote \( T'_{1,\theta} = \frac{T_{1,\theta}}{\pi(\theta)} \) and \( T'_{M,\theta,\alpha} = \frac{T_{M,\theta,\alpha}}{\pi(\theta)} \). We will check condition (4) for \( T'_{1,\theta}, T'_{M,\theta,\alpha} \) and \( 0 \leq c \leq 1 \), then separately for \( c < 0 \) and \( c > 1 \). For \( 0 \leq c \leq 1 \), we compute:

\[
\mathbb{E}[|T'_{M,\theta,\alpha} - c|] = \left( 1 - \frac{1}{M} \right) c + \frac{1}{M} (1-p)^M c + \frac{1}{M} \left( \sum_{j=1}^{M} \frac{M!}{j!(M-j)!} p^j (1-p)^{M-j} (j-c) \right)
\]

\[
= p + c \left( 1 - \frac{2}{M} (1 - (1-p)^M) \right)
\]

\[
\geq p + c (1 - 2p)
\]

\[
= \mathbb{E}[|T'_{1,\theta} - c|].
\]

For \( c < 0 \), we have

\[
\mathbb{E}[|T'_{M,\theta,\alpha} - c|] = \mathbb{E}[T'_{M,\theta,\alpha}] - c = \mathbb{E}[T'_{1,\theta}] - c = \mathbb{E}[T'_{1,\theta} - c].
\]
and the analogous argument works for $c \geq M$. To prove the result for $1 < c < M$, note that
\begin{align}
\mathbb{E}[|T^t_{1, \theta} - 1|] &\leq \mathbb{E}[|T^t_{M, \theta, \alpha} - 1|] \\
\mathbb{E}[|T^t_{1, \theta} - M|] &\leq \mathbb{E}[|T^t_{M, \theta, \alpha} - M|].
\end{align}
(5)

Also, the functions $f_1(c) \equiv \mathbb{E}[|T^t_{1, \theta} - c|]$ and $f_2(c) \equiv \mathbb{E}[|T^t_{M, \theta, \alpha} - c|]$ are continuous and piece-wise linear. For $c \geq 1$, they satisfy
\[
\frac{d}{dc} f_1(c) = 1 \geq \frac{d}{dc} f_2(c)
\]
(6)

where the derivative of $f_2$ exists. Combining inequalities (5) and (6), we conclude that
\[
\mathbb{E}[|T^t_{1, \theta} - c|] \leq \mathbb{E}[|T^t_{M, \theta, \alpha} - c|]
\]
for all $1 < c < M$. Thus we have verified (4) and the first claim follows. Analogous calculations yield the second claim of Proposition 3.

\[\square\]

2. Analysis of an Alternative ABC-MCMC Method

Here we give a result analogous to Corollary 1 for an alternative version of ABC-MCMC proposed in Wilkinson (2013), given in Algorithm 1 below. The constant $c$ can be any value satisfying $c \geq \sup_y K(\eta(y_{\text{obs}}) - \eta(y))$.

Algorithm 1. Alternative ABC-MCMC Method

Initialize $\theta^{(0)}$;

for $t = 1$ to $n$ do

\begin{enumerate}
\item Generate $\theta'$ from $q(\cdot | \theta^{(t-1)})$;
\item Generate $y_{\theta'}$ from the model $p(\cdot | \theta')$;
\item Generate $u$ uniformly on $[0, 1]$ if $u \leq r(\theta^{(t-1)}, \theta' | y_{\theta'}) \equiv \frac{K(\eta(y_{\text{obs}}) - \eta(y_{\theta'}))}{c} \min \left\{ 1, \frac{p(\theta | y_{\text{obs}})}{p(\theta' | y_{\theta'})} \right\}$, then
\item Set $\theta^{(t)} = \theta'$
\item else $\theta^{(t)} = \theta^{(t-1)}$
\end{enumerate}

We compare this likelihood-free Metropolis-Hastings method to likelihood-based Metropolis-Hastings, i.e. a Metropolis-Hastings chain that has the same target density $\pi_K(\theta | y_{\text{obs}})$ and proposal density $q$ as Algorithm 1.

Lemma 1. For any function $f \in L^2(\pi_K)$, the normalized asymptotic variance (3) of Algorithm 1 is at least as large as the normalized asymptotic variance of the Metropolis-Hastings algorithm with target density $\pi_K$ and proposal distribution $q$.

Proof. Both Algorithm 1 and the Metropolis-Hastings algorithm form Markov chains on the state space $\Theta$ with the same limiting density; call their transition kernels $Q$ and $P$ respectively. To prove the result, it suffices to show that $Q(\theta, A \setminus \{\theta\}) \leq P(\theta, A \setminus \{\theta\})$ for every $\theta \in \Theta$ and every measurable $A \subset \Theta$ (see e.g. Tierney (1998), Theorem 4).

Since the two algorithms use the same proposal density $q(\theta' | \theta^{(t)})$, it is furthermore sufficient to show that for every $\theta^{(t)}$ and $\theta'$, the acceptance probability of Metropolis-Hastings is at least as large as that of ABC-MCMC, where computing the latter requires marginalizing over $y_{\theta'}$. 

\[\square\]
Recall that
\[ r(\theta^{(t)}, \theta'|y_\theta) = \frac{K(\eta(y_{\text{obs}}) - \eta(y_\theta))}{c} \min \left\{ 1, \frac{\pi(\theta')q(\theta'|\theta')}{\pi(\theta^{(t)})q(\theta'|\theta^{(t)})} \right\}. \]  
(7)

Since \( p(\cdot|\theta) \) is a probability density,
\[ \int K(\eta(y_{\text{obs}}) - \eta(y)) p(y|\theta) dy \leq \sup_y K(\eta(y_{\text{obs}}) - \eta(y)) \quad \forall \theta \in \Theta \]  
(8)

Combining (7)-(8), the acceptance probability of Algorithm 1, marginalizing over \( y_\theta \), is
\[ a_{\text{ABC}} = \mathbb{E}[r(\theta^{(t)}, \theta'|y_\theta)] = \int r(\theta^{(t)}, \theta'|y)p(y|\theta')dy \leq \min \left\{ 1, \frac{\int K(\eta(y_{\text{obs}}) - \eta(y)) p(y|\theta')dy}{\sup_y K(\eta(y_{\text{obs}}) - \eta(y))} \right\}. \]  
(9)

The acceptance probability of the Metropolis-Hastings algorithm is
\[ a_{\text{MH}} = \min \left\{ 1, \frac{\int K(\eta(y_{\text{obs}}) - \eta(y)) p(y|\theta')dy}{\int K(\eta(y_{\text{obs}}) - \eta(y)) p(y|\theta^{(t)})dy} \right\}. \]  
(10)

Using (8), (9), and (10), \( a_{\text{ABC}}/a_{\text{MH}} \leq 1 \).

3. Further Discounting Simulation

Fig. 1: Sensitivity of the relationship between the bias and \( M \) to changes in the likelihood. Each curve is normalized by dividing by its respective \( \epsilon \) for \( M = 1 \). (a): Varying \( y_{\text{obs}} = 2 \) (light grey), 4, 6, 8 (black). (b): Varying \( \sigma_y = 0.01 \) (black), 0.05, 0.1, 0.5, 1, 2 (light grey).

We further explore discounting in Figures 1(a) and 1(b), which use a discount of \( \delta = 8 \) and vary \( y_{\text{obs}} \) from 2 to 8 (Figure 1(a)) and \( \sigma_y \) from 0.01 to 2 (Figure 1(b)). In both cases the changes
are meant to induce a divergence between the prior and the likelihood, and hence the prior and posterior. In these Figures, the requisite $\epsilon$ are scaled such that at $M = 1$ all normalized $\epsilon$ are 1. Figure 1(a) shows that as $y_{\text{obs}}$ grows, the benefit associated with using higher values of $M$ shrinks and eventually disappears. This is because for large $y_{\text{obs}}$ a large value of $\epsilon$ is required in order to frequently get a nonzero value for the approximated likelihood (7) and thus a reasonable acceptance rate; for instance, the unnormalized value of $\epsilon$ is 0.08 when $y_{\text{obs}} = 2$ and $M = 1$, while $\epsilon = 7.64$ when $y_{\text{obs}} = 8$ and $M = 1$. As such, the increased diversity from multiple samples is dwarfed by the scale of $\epsilon$.

In Figure 1(b) we examine sensitivity of our conclusions to $\sigma_y$. For large $\sigma_y$, additional (discounted-cost) pseudo-samples provide a benefit, because they improve the accuracy of the approximated likelihood (7). However, for small values of $\sigma_y$, the variability of the pseudo-samples $y_{i, \theta'}$ is low and so additional pseudo-samples do not provide much incremental improvement to the likelihood approximation. In summary, we only find a benefit of increasing the number of pseudo-samples $M$ in cases where there is a discounted cost to obtain those pseudo-samples, and even then the benefit can be decreased or eliminated when $y_{\text{obs}}$ is extreme under typical proposed values of $\theta$, or when the variability of $y$ under the model is low.

REFERENCES


[Received April 2012. Revised September 2012]