Practical unbiased Monte Carlo for intractable models

Sergios Agapiou

Department of Statistics, University of Warwick

Joint work with Gareth Roberts (Warwick) and Sebastian Vollmer (Oxford)

Greek Stochastics ζ', Athens 2014

http://www.sergiosagapiou.com/

S. Agapiou, G. O. Roberts and S. J. Vollmer, *Unbiased Monte Carlo: posterior estimation for intractable/infinite dimensional models*, http://arxiv.org/abs/1411.7713

C. H. Rhee, *Unbiased estimation with biased samples*, PhD thesis, Stanford University, 2013 (supervisor P. W. Glynn).

Problem overview

Want to estimate expectations wrt measure μ, available as limit of distributions.

e.g. μ is limit of:

- approximations corresponding to time-discretizations of SDE's
- basis expansion (Karhunen-Loeve)
- finite-time distributions of Markov chains (MCMC)

Standard methods truncate $\rightarrow$ bias:

- time-discretization bias in SDEs (GR13)
- discretization bias for measures in function space (ARV14)
- burn-in time for MCMC (GR13, ARV14)
- burn-in time and discretization bias for MCMC in function space (ARV14)

Outline

Unbiasing procedure of Glynn and Rhee

Aim: unbiasedly estimate $\mathbb{E}Y$, where simulating Y has infinite cost.

(think $Y = f(X)$, $X \sim \mu$ for $f : \mathcal{X} \to \mathbb{R}$)

- Use approximations Y_i of Y.
- Assume $\mathbb{E}Y_i \to \mathbb{E}Y$, hence

$$\mathbb{E}Y = \sum_{i=0}^{\infty} \mathbb{E}(Y_i - Y_{i-1}).$$

- Let $\Delta_i := Y_i - Y_{i-1}$. If Fubini applies

$$\mathbb{E}Y = \mathbb{E}\sum_{i=0}^{\infty} \Delta_i$$

$\sum_{i=0}^{\infty} \Delta_i$ unbiased but has infinite cost.

Unbiasing procedure of Glynn and Rhee

Idea (von Neumann, Ulam): use random truncation and correct for introduced bias.

- $$Z := \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)},$$
  N integer-valued r.v. independent of Δ_i, s.t. $\mathbb{P}(N \geq i) > 0, \forall i$.
- If Fubini applies
  $$\mathbb{E}(Z) = \mathbb{E}\left(\sum_{i=0}^{\infty} \frac{\mathbb{1}_{\{N\geq i\}}\Delta_i}{\mathbb{P}(N \geq i)}\right) = \sum_{i=0}^{\infty} \frac{\mathbb{E}(\mathbb{1}_{\{N\geq i\}}\Delta_i)}{\mathbb{P}(N \geq i)} = \sum_{i=0}^{\infty} \mathbb{E}\Delta_i = \mathbb{E}Y.$$
- Z is unbiased and has (almost) finite cost.
- To be practical, Z needs to have finite variance and finite expected computing time.

Unbiasing procedure of Glynn and Rhee

- Write $\|h\|_2 := (\mathbb{E}(h^2))^{\frac{1}{2}}$.

Proposition 1 (GR13)

Assume

$$\sum_{i \leq \ell} \frac{\|\Delta_i\|_2 \|\Delta_\ell\|_2}{\mathbb{P}(N \geq i)} < \infty.$$

Let $\tilde{\Delta}_i$ copy of Δ_i s.t. $\{\tilde{\Delta}_i\}$ mutually independent.

Then $\tilde{Z} := \sum_{i=0}^{N} \frac{\tilde{\Delta}_i}{\mathbb{P}(N \geq i)}$ is an unbiased estimator for $\mathbb{E}Y$ with finite variance.

Unbiasing procedure of Glynn and Rhee

- t_i expected cost of generating Δ_i. Expected computing time of Z

$$\mathbb{E}(\tau) = \mathbb{E}\sum_{i=0}^{N} t_i = \sum_{i=0}^{\infty} t_i \mathbb{P}(N \geq i).$$

- $\mathbb{P}(N \geq i)$ needs to decay *fast enough* for $\mathbb{E}(\tau) < \infty$ *but not too fast* for $\mathrm{Var}(Z) < \infty$.
- Suffices to generate Δ_i's with correct expectation.

Removing discretization bias

- $\mu = N(0, \mathcal{C}_0)$ Gaussian measure in separable Hilbert space $(\mathcal{X}, \langle\cdot,\cdot\rangle, \|\cdot\|)$.
- $\{\ell^{-2a}, \varphi_\ell\}$, $a > \frac{1}{2}$ orthonormal eigenpairs of $\mathcal{C}_0$.
- Karhunen-Loeve expansion: $u \sim \mu$ written as

$$u = \sum_{\ell=1}^{\infty} \ell^{-a} \xi_\ell \varphi_\ell$$

ξ_ℓ i.i.d. $N(0,1)$.

Aim: unbiasedly estimate $\mathbb{E}_\mu[f]$ for $f : \mathcal{X} \to \mathbb{R}$ Lipschitz.

Removing discretization bias

- Approximate by truncating

$$u_i = \sum_{\ell=1}^{j_i} \ell^{-a} \xi_\ell \varphi_\ell$$

- $\Delta_i = f(u_i) - f(u_{i-1})$.
- Expected cost of Δ_i is $t_i = \mathcal{O}(j_i)$ (number of $N(0,1)$ draws).
- Bound

$$\begin{aligned} \|\Delta_i\|_2^2 &\le \|f'\|_\infty^2 \, \mathbb{E}(\|u_i - u_{i-1}\|^2) \\ &= c\,\mathbb{E}\Big(\sum_{\ell=j_{i-1}+1}^{j_i} \ell^{-2a} \xi_\ell^2\Big) = c \sum_{\ell=j_{i-1}+1}^{j_i} \ell^{-2a} = \mathcal{O}(j_{i-1}^{1-2a} - j_i^{1-2a}). \end{aligned}$$

Removing discretization bias

Proposition 2 (ARV14)

Assume $a > 1$. Then $\exists$ choices j_i and $\mathbb{P}(N \geq i)$, s.t. Z is unbiased estimator of $\mathbb{E}_\mu[f]$ with finite variance and finite expected computing time.

Proof.

- Use Prop 1. Consider $j_i = 2^i$.
- $\|\Delta_i\|_2^2 = \mathcal{O}(2^{i(1-2a)})$

$$\sum_{i\leq\ell} \frac{\|\Delta_i\|_2\|\Delta_\ell\|_2}{\mathbb{P}(N \geq i)} \leq c\sum_{i=0}^{\infty} \frac{2^{\frac{i(1-2a)}{2}}}{\mathbb{P}(N \geq i)} \sum_{\ell=i}^{\infty} 2^{\frac{\ell(1-2a)}{2}} \leq c\sum_{i=0}^{\infty} \frac{2^{i(1-2a)}}{\mathbb{P}(N \geq i)}.$$

- $t_i = \mathcal{O}(2^i)$, $\quad \sum_{i=0}^{\infty} t_i \mathbb{P}(N \geq i) \leq c\sum_{i=0}^{\infty} 2^i\, \mathbb{P}(N \geq i)$.
- Can choose $\mathbb{P}(N \geq i)$ s.t. both sums finite since $2^{i(1-2a)}$ decays faster than 2^i blows-up.

□

Removing burn-in time bias

- $\mathcal{X}$ general state space, d distance-like function.
- X Markov chain with transition P and invariant distribution μ.

Aim: unbiasedly estimate $\mathbb{E}_{\mu}[f]$ for $f : \mathcal{X} \to \mathbb{R}$ Lipschitz wrt d.

Removing burn-in time bias

- Approximate using finite-time distributions.
- Finite-time distributions converge weakly but not enough. Need couplings s.t. $f(X_i)$ comes close in L^2.
- GR13: use tricks which turn weak convergence to a.s. convergence/coalescence.
- ARV14: suffices to have simulatable coupling K between chains started at different states which contracts wrt d.

Assumption

- $K^n d^2 \leq c r^n d^2$ for some $r < 1$;
- $\exists x_0 \in \mathcal{X}$ s.t. $sup_n P^n d(x_0, \cdot) < \infty$.

Removing burn-in time bias

- $\{a_i\}$ increasing sequence of positive integers.
- To generate Δ_i, use top approximation level chain $\mathcal{T}_\cdot^i$ running for a_i steps and bottom approximation level chain $\mathcal{B}_\cdot^i$ running for a_{i-1} steps, both started at x_0.

Coupled contraction for unbiased estimation

For $i = 0$

- set $\mathcal{T}^0_{-a_0} = x_0$ and run chain until $\mathcal{T}^0_0$;
- set $\Delta_0 = f(\mathcal{T}^0_0)$.

For $i \geq 1$

- set $\mathcal{T}^i_{-a_i} = x_0$ and run chain until $\mathcal{T}^i_{-a_{i-1}}$;
- set $\mathcal{B}^i_{-a_{i-1}} = x_0$;
- evolve $\mathcal{B}^i_k$ and $\mathcal{T}^i_k$ jointly according to K upto time 0;
- set $\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0)$.

Removing burn-in time bias

- Estimate

$$
\begin{aligned}
\|\Delta_i\|_2^2 \leq & \|f'\|_\infty^2 \,\mathbb{E} d^2\left(\mathcal{T}_0^i, \mathcal{B}_0^i\right) \\
\leq & c\, \mathbb{E}\mathbb{E}\left(d^2\left(\mathcal{T}_0^i, \mathcal{B}_0^i\right) | \mathcal{F}_{-a_{i-1}}\right) \\
\leq & c\, \mathbb{E}\left(K^{a_{i-1}} d^2(\mathcal{T}_{-a_{i-1}}^i, x_0)\right) \\
\leq & c\, r^{a_{i-1}} \mathbb{E} d^2(\mathcal{T}_{-a_{i-1}}^i, x_0) \\
\leq & c\, r^{a_{i-1}}.
\end{aligned}
$$

- Cost of Δ_i, $t_i = \mathcal{O}(a_i)$ (number of steps).

Removing burn-in bias

Proposition 3 (ARV14)

$\exists$ choices a_i and $\mathbb{P}(N \geq i)$, s.t. Z is unbiased estimator of $\mathbb{E}_\mu[f]$ with finite variance and expected computing time.

Proof.

- Use Prop 1. Note $a_i \geq i$.
- Since $r < 1$

$$\sum_{i \geq \ell} \frac{\|\Delta_i\|_2 \|\Delta_\ell\|_2}{\mathbb{P}(N \geq i)} \leq c \sum_{i=0}^{\infty} \frac{r^{\frac{i}{2}}}{\mathbb{P}(N \geq i)} \sum_{\ell=i}^{\infty} r^{\frac{\ell}{2}} \leq c \sum_{i=0}^{\infty} \frac{r^i}{\mathbb{P}(N \geq i)}$$

- $\sum_{i=0}^{\infty} t_i \mathbb{P}(N \geq i) \leq c \sum_{i=0}^{\infty} a_i \, \mathbb{P}(N \geq i)$.
- Possible to choose $\mathbb{P}(N \geq i)$ s.t. both sums finite, under mild growth condition on a_i.

□

Comparison of unbiased estimator (UE) vs ergodic average (EA)

- 1d Gaussian autoregression

$$X_{n+1} = \rho X_n + \sqrt{1-\rho^2}\,\xi_{n+1},$$

  $\rho \in (0,1)$, ξ_n i.i.d. $N(0,1)$.

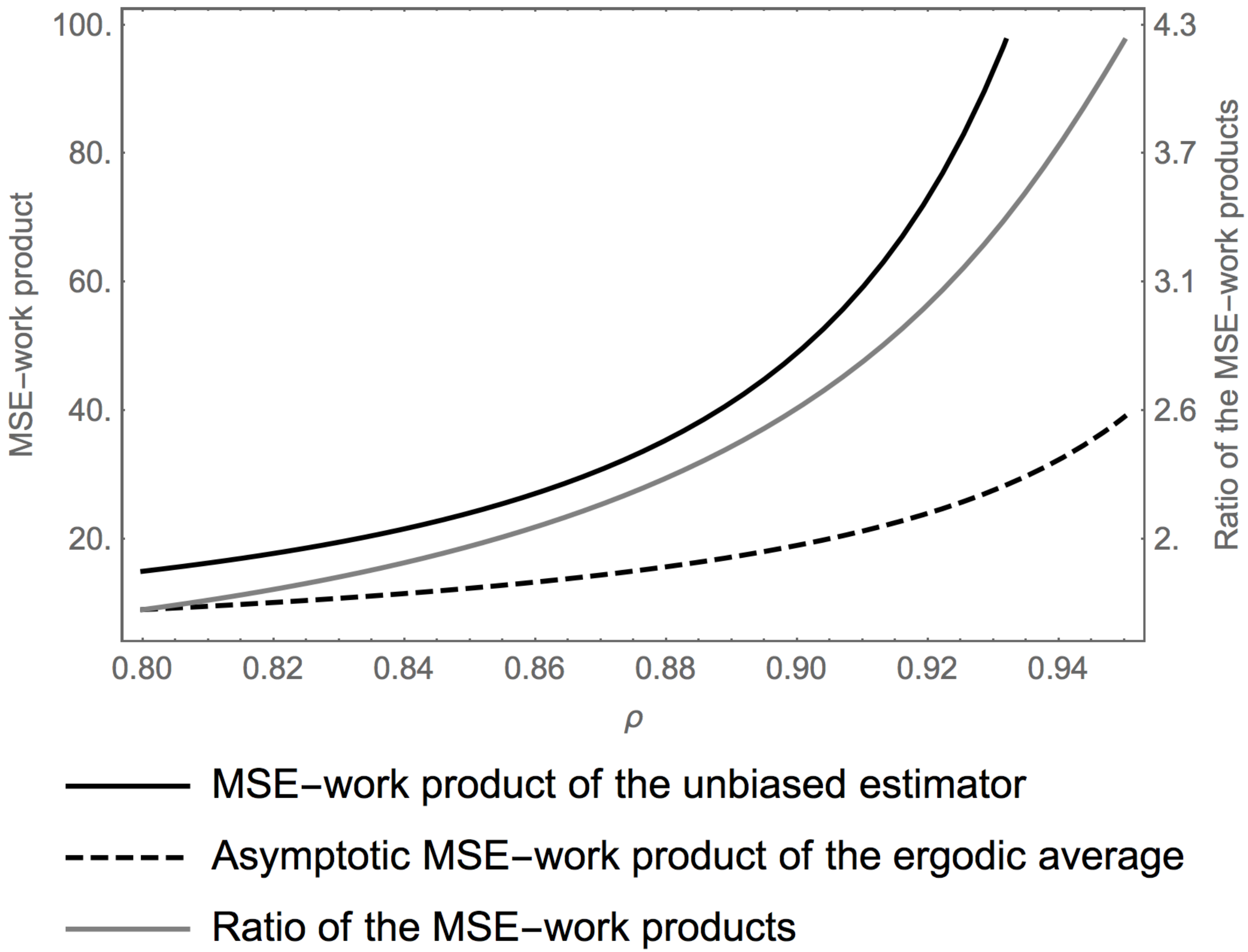
- Ergodic with invariant distribution $\mu = N(0,1)$. Estimate $\mathbb{E}_\mu(\mathrm{Id})(=0)$.

- UE constructed by coupling chains started at different points using same randomness.

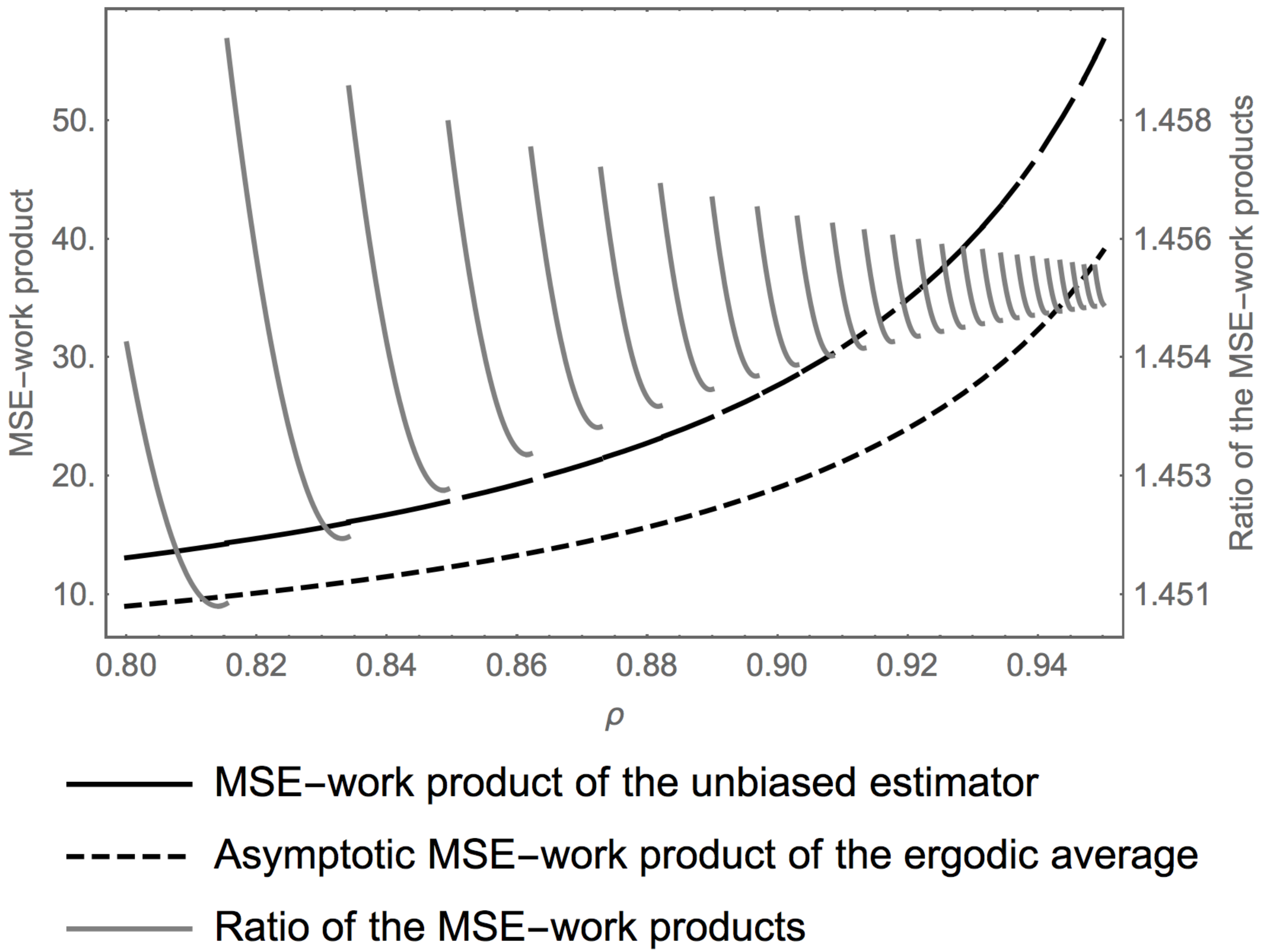
Comparison of unbiased estimator (UE) vs ergodic average (EA)

- Compare MSE-work product of MC estimator based on UE vs EA.
- For EA
$$\lim_{n\to\infty} \text{MSE-work} = \frac{1+\rho}{1-\rho} T_{\text{step}}.$$
- For UE
$$\text{MSE-work} = \left(\sum_{i=1}^{\infty} \frac{\rho^{2a_{i-1}} \left(1-\rho^{2(a_i - a_{i-1})}\right)}{\mathbb{P}(N \geq i)} + 1 - \rho^{2a_0} \right) \sum_{i=0}^{\infty} a_i \, \mathbb{P}(N \geq i).$$
- Can optimize performance of UE by minimizing wrt a_i and $\mathbb{P}(N \geq i)$.
  Hard optimization problem!

Optimized $\mathbb{P}(N \geq i)$, fixed $a_i = 4(i+1)$



Optimized $\mathbb{P}(N \geq i)$ and a_i over subclass $a_i = m(i+1)$



Conclusions - further work

- UE is often feasible.
- Combining described techniques can perform UE using MCMC in function space.
  - Approximation using finite-time distributions and discretizing space.
  - Challenge to get good transdimensional couplings.
  - In ARV14, use independence sampler and pCN algorithm to perform UE wrt posteriors arising in Bayesian inverse problems in function space.
- Optimization wrt parameters is crucial especially in function space setting.
- UE easily parallelizable: a) use independent copies of Z, b) Δ_i's independent.
- UE seems competitive. Looking forward to comparisons in problems of higher complexity.

http://www.sergiosagapiou.com/

- S. Agapiou, G. O. Roberts and S. J. Vollmer, *Unbiased Monte Carlo: posterior estimation for intractable/infinite dimensional models*, arXiv:1411.7713
- C. H. Rhee, *Unbiased estimation with biased samples*, PhD thesis, Stanford University, 2013, (supervisor P. W. Glynn).
- J. G. Propp and D. B. Wilson, *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Random Structures and Algorithms, 1996.
- M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, arXiv:1302.6989.
- M. Hairer, A. M. Stuart and S. J. Vollmer *Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions*, arXiv:1112.1392