Practical unbiased Monte Carlo for intractable models

Sergios Agapiou

Department of Statistics, University of Warwick

Statistics Seminar
20th March 2015, Imperial College London

Enabling Quantification of
EQUIP
Uncertainty for Inverse Problems

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).

Outline

Outline

Problem overview

Want to estimate expectations of functions f wrt an intractable measure μ,

$$\mathbb{E}_\mu[f] := \mathbb{E}_\mu[f(\cdot)].$$

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)

Problem overview

- Would like to use Monte Carlo estimator: for $X^{(m)} \overset{iid}{\sim} \mu$ let

$$R_M := \frac{1}{M} \sum_{m=1}^{M} f(X^{(m)}).$$

For all M

$$\mathbb{E}[R_M] = \mathbb{E}_\mu[f] \qquad (R_M \text{ unbiased})$$

and

$$R_M \xrightarrow{M} \mathbb{E}_\mu[f], \text{ almost surely} \qquad (R_m \text{ consistent})$$

Problem overview

Intractability of μ forces the use of approximations μ_i introducing 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)

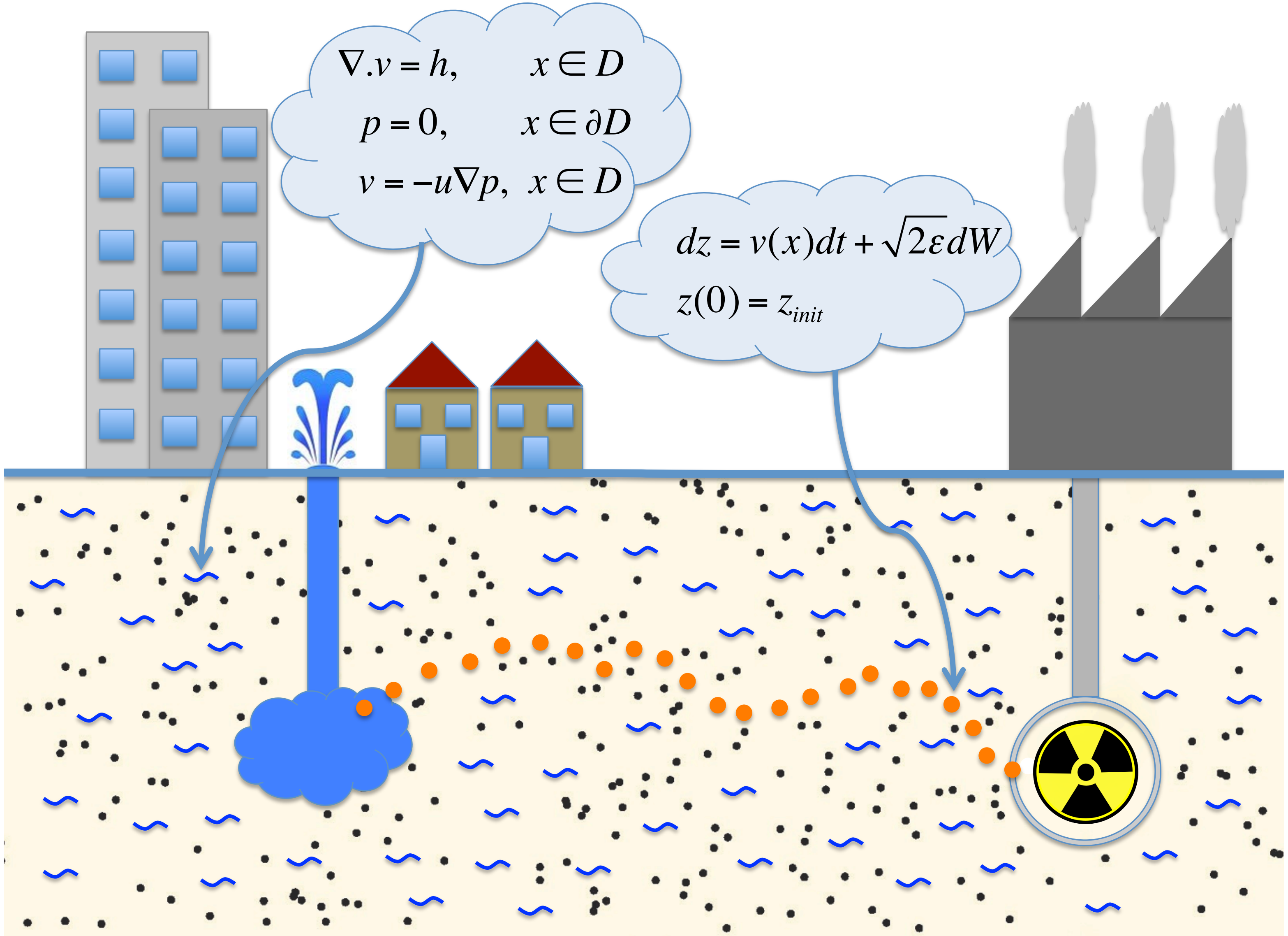
Bias typically leads to sub-optimal convergence rates of MC estimator (ergodic average) in infinite computational budget limit.

(even MLMCMC loses at least a log, see KST13)

Outline

Example - Contamination scenario

- u permeability field
- p pressure
- v Darcy velocity
- $p = G(u)$



Quantity of interest: $f(u) = \mathbb{E}[\inf_{t \geq 0} \{|z(t)| > R\}]$

Example - UQ in contamination scenario

Permeability field u unknown, have prior information $u \sim \mu_0$.

- **Vanilla-UQ**: probe $\mu_0 \circ f^{-1}$, e.g. estimate $\mathbb{E}_{\mu_0}[f(u)]$.
- Have noisy indirect measurements of pressure at J locations: data model in $\mathbb{R}^J$

$$y = \mathcal{G}(u) + \eta, \ \eta \sim N(0, \mathrm{I}).$$

  Formulate Bayesian inverse problem (see DS13), μ^y posterior on $u|y$

$$\frac{d\mu^y}{d\mu_0}(u; y) \propto \exp\left(-\frac{1}{2}\|y - \mathcal{G}(u)\|^2\right).$$

- **BIP-UQ**: probe $\mu^y \circ f^{-1}$, e.g. estimate $\mathbb{E}_{\mu^y}[f(u)]$.

Example - UQ sources of bias

- **Vanilla-UQ**:
  - μ_0 is ∞-dim, needs to be approximated by $\mu_{0,i}$ in $\mathbb{R}^i$ introducing discretization bias.
- **BIP-UQ**:
  - cannot sample μ^y directly, construct Markov chain targeting μ^y, use finite-time distributions $\mu^{y,k}$ burn-in time issues.
  - to implement in computer construct Markov chain targeting approximation μ_i^y in $\mathbb{R}^i$, use finite-time distributions $\mu_i^{y,k}$ introducing discretization bias and burn-in time issues.

Outline

Debiasing idea - John von Neumann, Stanislaw Ulam

- We study unbiased estimation of $\mathbb{E}_\mu[f]$ using biased samples, $X_i \sim \mu_i$.
- Assume $\mathbb{E}_{\mu_i}[f] \xrightarrow{i} \mathbb{E}_\mu[f]$.
- Define $\Delta_i := f(X_i) - f(X_{i-1})$.
- If Fubini applies

$$\mathbb{E}_\mu[f] = \sum_{i=1}^{\infty}(\mathbb{E}_{\mu_i}[f] - \mathbb{E}_{\mu_{i-1}}[f]) = \sum_{i=1}^{\infty}\mathbb{E}\Delta_i \stackrel{?}{=} \mathbb{E}\sum_{i=1}^{\infty}\Delta_i.$$

- $\sum_{i=1}^{\infty}\Delta_i$ is unbiased but requires infinite computing time.

Debiasing idea - John von Neumann, Stanislaw Ulam

$$Z := \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)},$$

N integer-valued r.v. independent of Δ_i, s.t. $\mathbb{P}(N \geq i) > 0, \forall i$.

- If Fubini applies

$$\mathbb{E}[Z] = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{\mathbb{1}_{\{N \geq i\}}\Delta_i}{\mathbb{P}(N \geq i)}\right] \stackrel{?}{=} \sum_{i=0}^{\infty} \frac{\mathbb{E}[\mathbb{1}_{\{N \geq i\}}\Delta_i]}{\mathbb{P}(N \geq i)} = \sum_{i=0}^{\infty} \mathbb{E}\Delta_i = \mathbb{E}_\mu[f].$$

- Z unbiased and requires finite computing time (almost).

Debiasing idea - John von Neumann, Stanislaw Ulam

- To be practical, Z needs to have finite variance and finite expected computing time.
- $Z^{(m)}$ independent copies of Z, $Z_M := \frac{1}{M}\sum_{m=1}^{M} Z^{(m)}$.
- $Z_{M(c)}$ MC estimator with computational budget c.
- GW92, as $c \to \infty$

$$\sqrt{c}\,\left(Z_{M(c)} - \mathbb{E}_\mu[f]\right) \Rightarrow \sqrt{\mathbb{E}(\tau).\mathrm{Var}(Z)}\,N(0,1).$$

- Optimal rate of convergence $c^{-\frac{1}{2}}$.
- Optimize by minimizing $\mathbb{E}(\tau).\mathrm{Var}(Z)$ (refer to this as MSE-work product).

Unbiasing theory of Glynn and Rhee

Proposition (GR13)

Assume

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

Then $Z := \sum_{i=0}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$ is an unbiased estimator for $\mathbb{E}_\mu[f]$ with finite variance.

Can use $\tilde{\Delta}_i$ copies of Δ_i s.t. $\{\tilde{\Delta}_i\}$ mutually independent.

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

$$\mathbb{E}(\tau) = \mathbb{E}\sum_{i=0}^{N} t_i = \mathbb{E}\sum_{i=1}^{\infty} t_i \mathbb{1}_{\{N\geq i\}} = \sum_{i=0}^{\infty} t_i \mathbb{P}(N \geq i).$$

- To be possible to choose $\mathbb{P}(N \geq i)$ s.t. Z practical, suffices to generate Δ_i's with correct expectation s.t. $\|\Delta_i\|_2^2$ decays sufficiently faster than t_i blows-up.

Outline

Removing discretization bias

- $\mathcal{X} = L^2[0,1]$, $\{\varphi_\ell\}$ complete orthonormal basis.
- μ Gaussian measure in $\mathcal{X}$ given via the Karhunen-Loeve expansion:

$$\mu = \mathcal{L}\left(\sum_{\ell=1}^{\infty} \ell^{-a} \xi_\ell \varphi_\ell\right), \qquad \xi_\ell \overset{iid}{\sim} N(0,1), \quad a > \frac{1}{2}.$$

- To estimate $\mathbb{E}_\mu[f]$, need to truncate introducing discretization bias in MC estimators. (Vanilla-UQ example)

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

Removing discretization bias

- Approximations

$$\mu_i = \mathcal{L}\left(\sum_{\ell=1}^{j_i} \ell^{-a}\xi_\ell\varphi_\ell\right), \quad j_i \text{ increasing.}$$

- $\Delta_i = f(u_i) - f(u_{i-1})$, $u_i \sim \mu_i$ using same random seeds.
- Bound

$$\|\Delta_i\|_2^2 = \mathbb{E}(|f(u_i) - f(u_{i-1})|^2) \leq \|f'\|_\infty^2 \, \mathbb{E}(\|u_i - u_{i-1}\|^2) = \mathcal{O}(j_{i-1}^{1-2a} - j_i^{1-2a}).$$

- Cost of Δ_i, $t_i = \mathcal{O}(j_i) \quad (\# \; N(0,1)$ draws).

Removing discretization bias

Theorem 1 (ARV14)

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

Proof.

- Consider $j_i = 2^i$. Use Proposition from GR13.
- $t_i = \mathcal{O}(2^i)$, $\|\Delta_i\|_2^2 = \mathcal{O}(2^{i(1-2a)})$.
- For $a > 1$, $\|\Delta_i\|_2^2$ decays sufficiently faster than t_i blows-up.
- Can choose $\mathbb{P}(N \geq i)$ s.t. $\mathbb{E}(\tau), \mathrm{Var}(Z) < \infty$.

□

Removing burn-in time issues

- $\mathcal{X}$ general state space, d distance in $\mathcal{X}$.
- Measure μ intractable, cannot be sampled directly but can construct Markov chain $\mathbb{X} = (X_n)_{n\in\mathbb{N}}$ with transition kernel P and stationary distribution μ.
- a_i increasing positive integers.
- To estimate $\mathbb{E}_\mu[f]$, use finite-time distributions $\mu_i = \mathcal{L}(X_{a_i})$ introducing burn-in issues.

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

Removing burn-in time issues

- Finite-time distributions converge weakly. This is not enough for $f(X_i)$ to come close in L^2, i.e. for convergence of Δ_i.
- GR13: use tricks which turn weak convergence to a.s. convergence/coalescence (coupling from the past Propp & Wilson). Require uniform ergodicity.
- ARV14: suffices to have simulatable coupling K between chains started at different states which contracts wrt d.

Assumption

- $K^n(d^2(x,y)) \leq cr^n d^2(x,y)$ 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 issues

- To generate Δ_i, use top level chain $\mathcal{T}^i_\cdot$ running for a_i steps and bottom level chain $\mathcal{B}^i_\cdot$ running for a_{i-1} steps, coupled as follows:

Coupled contraction for unbiased estimation

- 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 $\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0)$.

$$
\begin{array}{ccccccccccc}
 & & & x_0 = & \mathcal{B}^i_{-a_{i-1}} & \ldots & \mathcal{B}^i_{-a_0} & \ldots & \mathcal{B}^i_0 & & \\
 & & & & | & | & | & | & | & \} & \Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0) \\
x_0 = & \mathcal{T}^i_{-a_i} & \ldots & & \mathcal{T}^i_{-a_{i-1}} & \ldots & & \ldots & \mathcal{T}^i_0 & &
\end{array}
$$

Removing burn-in time issues

- 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 time issues

Theorem 2 (ARV14)

$\exists$ choices a_i and $\mathbb{P}(N \geq i)$, s.t. $Z = \sum_{i=1}^{N} \frac{\Delta_i}{\mathbb{P}(N \geq i)}$ is unbiased estimator of $\mathbb{E}_\mu[f]$ with finite variance and finite expected computing time.

Proof.

- Use Proposition from GR13.
- $t_i = \mathcal{O}(a_i)$, $\|\Delta_i\|_2^2 = \mathcal{O}(r^{a_i})$.
- $\|\Delta_i\|_2^2$ decays sufficiently faster than t_i blows-up.
- Can choose $\mathbb{P}(N \geq i)$ s.t. $\mathbb{E}(\tau), \mathrm{Var}(Z) < \infty$.

$\square$

Removing burn-in time issues - remarks

- Genuine generalization of GR13.
- Algebraic contraction rate of the coupling is sufficient for UE to work

$$K^n d^2 \leq C n^{-2r} d^2, \quad r > \frac{1}{2}.$$

- Many couplings available from e.g. stochastic control and coupling from the past.

Removing both burn-in and discretization bias

- Combining can perform UE of $\mathbb{E}_\mu[f]$ for μ both ∞-dim and only accessible in the limit of a Markov chain (BIP-UQ example).
- $\mathcal{X}$ ∞-dim state space, d distance in $\mathcal{X}$.
- Approximation using finite-time distributions and discretizing space.

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

Removing both burn-in and discretization bias

- a_i, j_i increasing sequences of integers.
- Top chain $\mathcal{T}^i_\cdot$ more steps and higher discretization level than bottom chain $\mathcal{B}^i_\cdot$

$$
\begin{array}{lccccccccc}
j_{i-1}: & & & x_0 = \mathcal{B}^i_{-a_{i-1}} & \dots & \mathcal{B}^i_{-a_0} & \dots & \mathcal{B}^i_0 & \\
 & & & | & | & | & | & | & \}\Delta_i = f(\mathcal{T}^i_0) - f(\mathcal{B}^i_0) \\
j_i: & x_0 = & \mathcal{T}^i_{-a_i} \quad \dots & \mathcal{T}^i_{-a_{i-1}} & \dots & & \dots & \mathcal{T}^i_0 &
\end{array}
$$

Removing both burn-in and discretization bias - strategy

$$\|\Delta_i\|_2^2 \leq \|f'\|_\infty^2 \mathbb{E} d^2(\mathcal{T}_0^i, \mathcal{B}_0^i)$$

- Need good couplings between chains started at different initial states and at neighbouring discretization levels.
- d bdd distance. Suppose MCMC has fixed-state space contracting coupling s.t.

$$\mathbb{E} d\left(\mathcal{T}_n^i(x_1), \mathcal{T}_n^i(x_2)\right) \leq r^n d(x_1, x_2). \qquad \text{(artificial)}$$

Removing both burn-in and discretization bias - strategy

- $\mathcal{I}_k^i$ intermediate steps evolving $\mathcal{B}_{k-1}^i$ according to top level kernel P_{j_i}.

$$\begin{aligned}\mathbb{E}d(\mathcal{T}_0^i, \mathcal{B}_0^i) &\leq \mathbb{E}d(\mathcal{T}_0^i, \mathcal{I}_0^i) + \mathbb{E}d(\mathcal{I}_0^i, \mathcal{B}_0^i) \\ &\leq rd(\mathcal{T}_{-1}^i, \mathcal{B}_{-1}^i) + C_{j_{i-1}, j_i} \\ &\dots \\ &\leq r^{a_{i-1}} + C_{j_{i-1}, j_i} \frac{1 - r^{a_{i-1}}}{1 - r}.\end{aligned}$$

- $C_{j_{i-1}, j_i} = \mathcal{O}(i^{-p}) \xrightarrow{i} 0$ provided acceptance behaviour of $P_{j_{i-1}}$ and P_{j_i} similar for large i.
- Optimize by choosing $j_i = j_i(a_i)$ to balance terms.
- Get convergence $\|\Delta_i\|_2^2 \lesssim r^{a_i}$ as $i \to \infty$, sufficient for unbiased estimation if e.g. $t_i \lesssim a_i j_i^{\theta}$.

Removing both burn-in and discretization bias

In ARV14, show this works:

1. in non-linear Bayesian inverse problem setting with uniform priors, using independence sampler under assumptions securing uniform ergodicity;
2. for targets μ which have Lipschitz log-density wrt Gaussian, using pCN algorithm (MH with proposal $X_{k+1} = \lambda X_k + \sqrt{1-\lambda^2}\xi$).

   Use fixed-state space, dimension independent coupling contraction results from HSV11, DM14.

Outline

Toy model - Gaussian autoregression

- 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.
- Coupling contracts geometrically with rate $r = \rho$ for $d(x,y) = |x-y|$.

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

- Compare MSE-work product of MC estimator based on UE vs EA.
- For EA

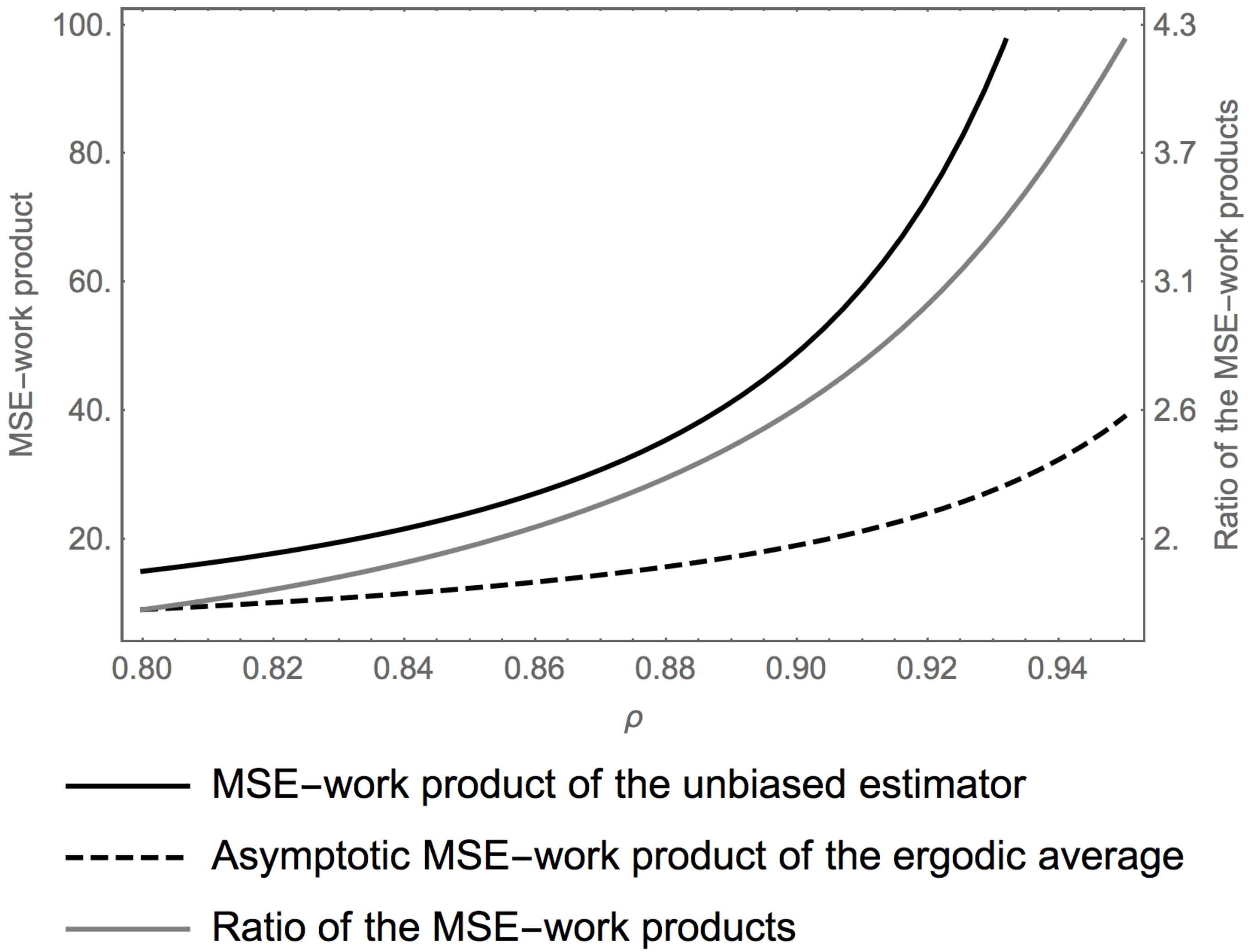
$$\lim_{n\to\infty} \text{MSE-work} = \frac{1+\rho}{1-\rho} T_{\text{step}}.$$

- For UE

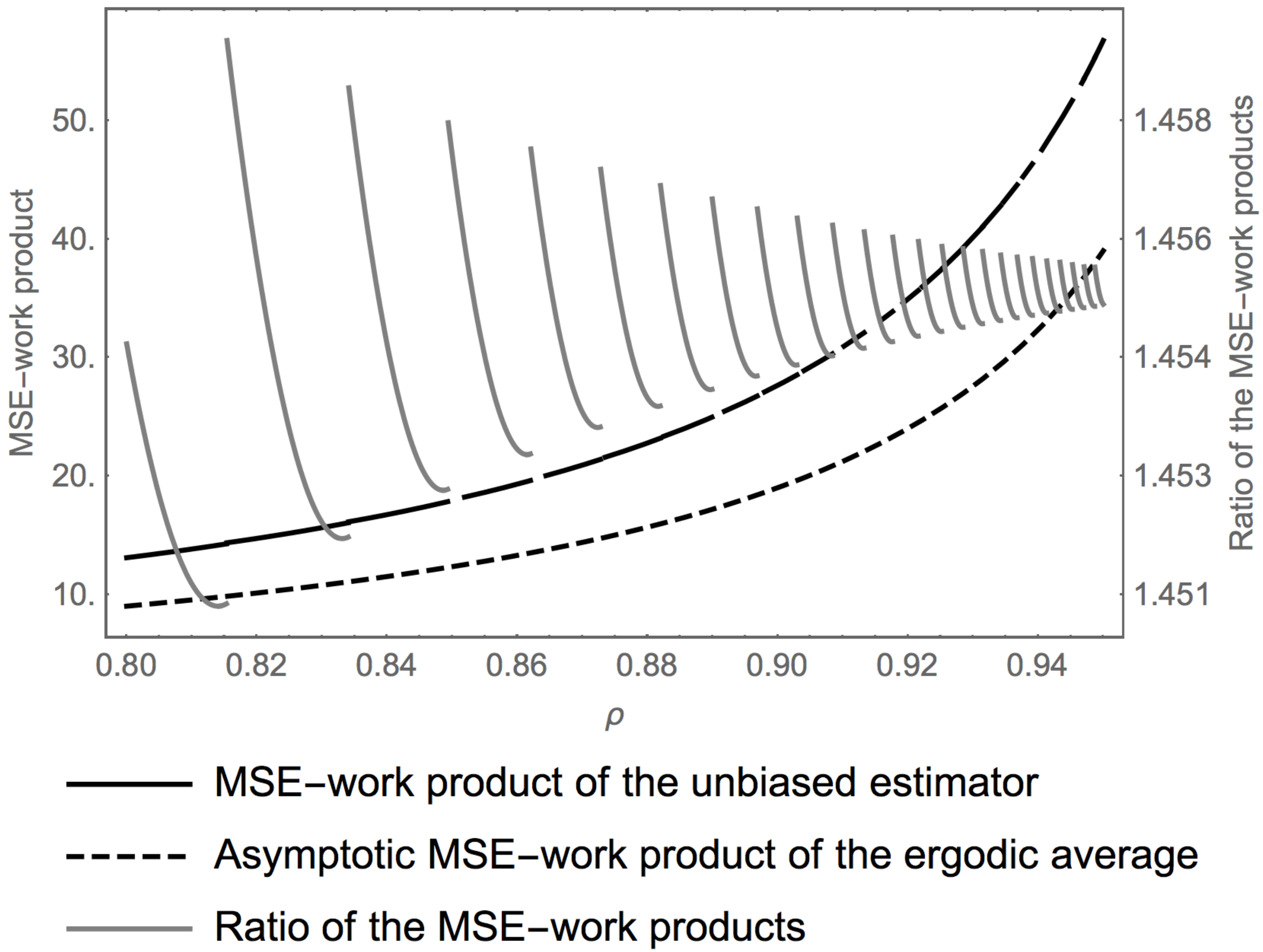
$$\text{MSE-work} = \left(\sum_{i=1}^{\infty} \frac{\rho^{2a_{i-1}} \left(1 - \rho^{2(a_i - a_{i-1})}\right)}{\mathbb{P}(N \geq i)} + 1 - \rho^{2a_0} \right) \sum_{i=0}^{\infty} a_i \, \mathbb{P}(N \geq i).$$

- Can optimize performance of UE by minimizing wrt a_i and $\mathbb{P}(N \geq i)$. Hard!
- In GR13 consider only $a_i = i$, optimize over $\mathbb{P}(N \geq i)$.

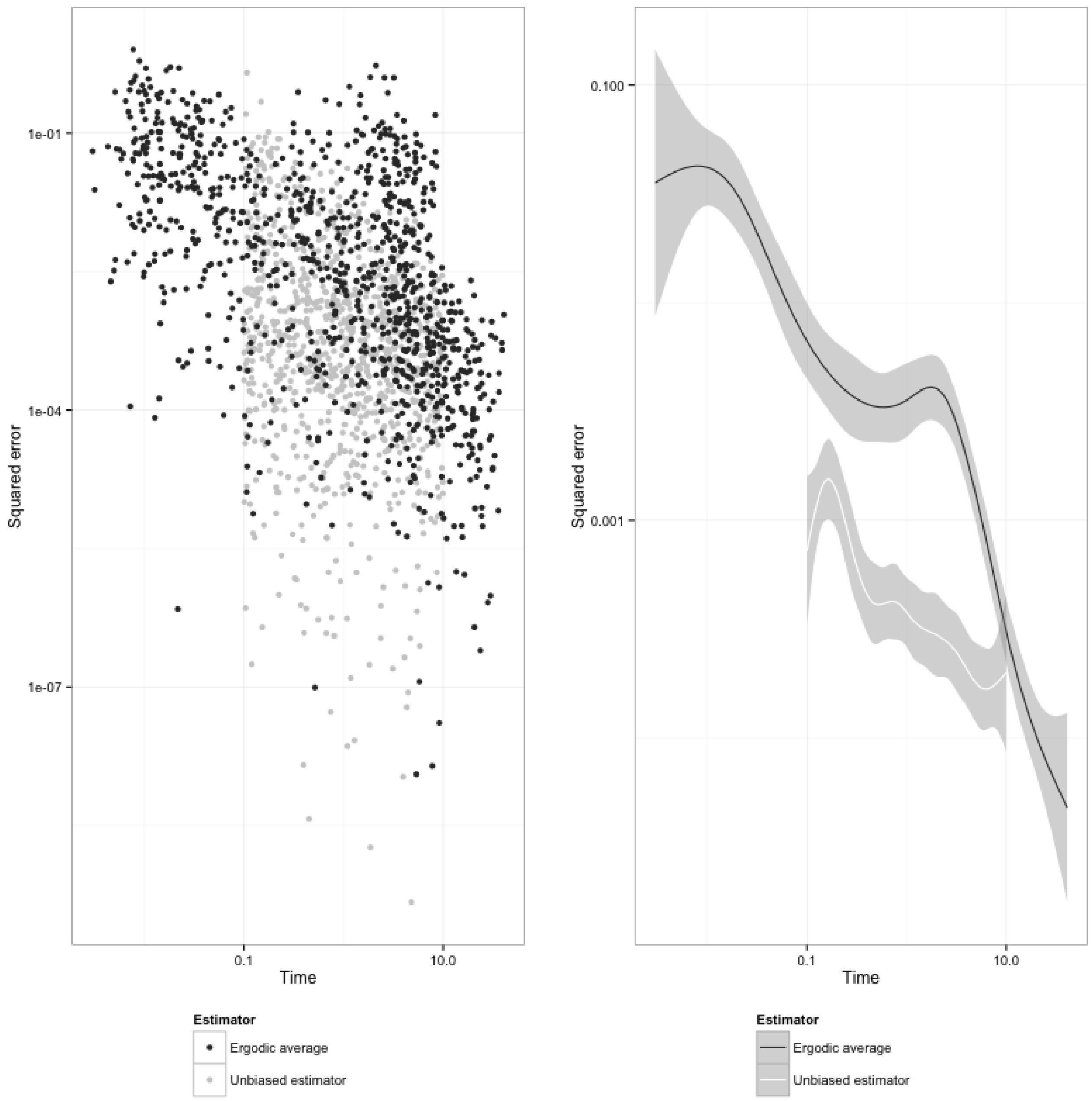
Optimized $\mathbb{P}(N \geq i)$, fixed $a_i = 4(i+1)$



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



10-core parallel setting, $\rho = 0.8$



Outline

Conclusions - further work

- UE is often feasible.
- Optimization wrt parameters is crucial especially in function space setting (although performance not overly sensitive on knowledge of the coupling).
- 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*, The Annals of Applied Probability, 2014.
- C. Ketelsen, R. Scheichl and A. K. Teckentrup *A hierachical Multilevel Markov Chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow*, arXiv:1303.7343.