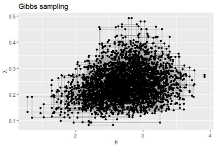
Unbiased Langevin Monte Carlo

Neil Chada
neilchada123@gmail.com

Algorithms and Inference Seminar
University of Warwick, 3rd May 2024

HERIOT WATT UNIVERSITY
MAXWELL INSTITUTE FOR MATHEMATICAL SCIENCES

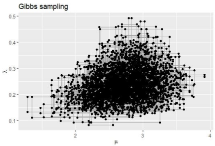
Motivation



- Aim: unbiased estimator - $\mathbb{E}_{\pi_h}[f(X)] = \mathbb{E}_{\pi}[f(X)]$.
- Produce samples from a distribution π

$$\mathbb{E}_{\pi}[f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x)dx, \quad \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_{\pi}[f(X)]\right) \to N(0, \sigma^2).$$

Motivation

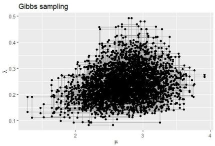


- Aim: unbiased estimator - $\mathbb{E}_{\pi_h}[f(X)] = \mathbb{E}_{\pi}[f(X)]$.
- Produce samples from a distribution π

$$\mathbb{E}_{\pi}[f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x)dx, \quad \sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_{\pi}[f(X)]\Big) \to N(0,\sigma^2).$$

- Issues: (i) MCMC bias (ii) discretization bias (iii) scalability $\sim \mathcal{O}(d^{\cdots})$

Motivation



- Aim: unbiased estimator - $\mathbb{E}_{\pi_h}[f(X)] = \mathbb{E}_{\pi}[f(X)]$.
- Produce samples from a distribution π

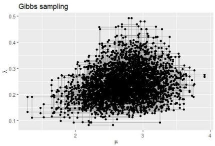
$$\mathbb{E}_{\pi}[f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x)dx, \quad \sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_{\pi}[f(X)]\Big) \to N(0, \sigma^2).$$

- Issues: (i) MCMC bias (ii) discretization bias (iii) scalability $\sim \mathcal{O}(d^{\cdots})$

$\vdots$

$\vdots$

Motivation



- Aim: unbiased estimator - $\mathbb{E}_{\pi_h}[f(X)] = \mathbb{E}_{\pi}[f(X)]$.
- Produce samples from a distribution π

$$\mathbb{E}_{\pi}[f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x)dx, \quad \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_{\pi}[f(X)]\right) \to N(0,\sigma^2).$$

- Issues: (i) MCMC bias (ii) discretization bias (iii) scalability $\sim \mathcal{O}(d^{\cdots})$

$\vdots$

$\vdots$

Exploit (Kinetic) Langevin methods to handle all issues!

Part I: Unbiased Estimation with ULD

H. Ruzaquat (KAUST), NKC, and A. Jasra (CUHK-SZ) [SISC 23]

Biased MCMC

- MCMC algorithms define π-invariant Markov kernel K.
- Initialize $X_0 \sim \pi_0 \neq \pi$ & iterate

$$X_t \sim K(X_{t-1}, \cdot), \quad t = 1, \ldots, T.$$

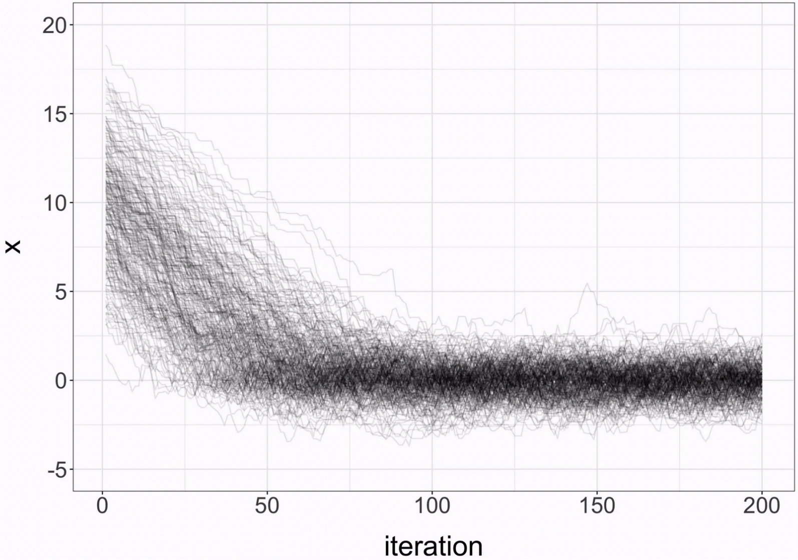
- Compute

$$\frac{1}{T-b+1} \sum_{t=b}^{T} f(X_t) - \mathbb{E}_\pi[f(X)], \quad T \to \infty,$$

where $b \geq 0$ are discarded as burn-in.
- Estimator is biased since $\pi_0 \neq \pi$.

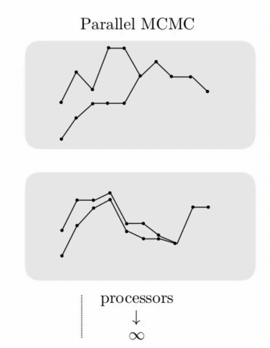
Averaging of independent copies does not provide an unbiased estimator

$\pi = \mathcal{N}(0,1)$, $\pi_0 = \mathcal{N}(10, 3^2)$, $K =$ RWMH with proposal std 0.5



Proposed Methodology

- Each processor runs two coupled chains $X = (X_t)$ and $Y = (Y_t)$.
- Terminate at some random time, i.e. meeting time.
- Returns unbiased estimator $H_{k:m}$ of $\mathbb{E}_\pi[f(X)]$.
- "Independent averaging" to estimate $\mathbb{E}_\pi[f(X)]$, as copies $\longrightarrow \infty$.
- **Efficiency** depends on expected cost and variance.



Debiasing Ideas

Glynn, Rhee. *Exact estimation for Markov chain equilibrium expectation.* (2014).

$$\mathbb{E}_\pi[f(X)] = \lim_{t\to\infty} \mathbb{E}_\pi[f(X_t)] = \mathbb{E}[f(X_k)] + \mathbb{E} \sum_{t=k+1}^{\infty} f(X_t) - f(X_{t-1})$$

- Truncate series, since $X_t = Y_{t-1}$, for $t \geq \tau$

$$\mathbb{E}_\pi[f(X)] = \mathbb{E}\left[f(X_k) + \sum_{t=k+1}^{\tau-1} f(X_t) - f(Y_{t-1}) \right].$$

- Unbiased estimator: for any $k \geq 0$

$$F_k(X, Y) = f(X_k) + \sum_{t=k+1}^{\tau-1} f(X_t) - f(Y_{t-1})\}$$

1st term is biased, 2nd term corrects the bias!

Unbiased Estimator

We consider

$$\mathbb{E}[\xi_{l_*}] = \pi_{l_*}(\varphi)$$
$$\mathbb{E}[\xi_l] = \pi_l(\varphi) - \pi_{l-1}(\varphi) =: [\pi_l - \pi_{l-1}](\varphi).$$

Our unbiased estimator is

$$\widehat{\pi}(\varphi) = \frac{\xi_l}{\mathbb{P}_L(l)}.$$

Moreover, if

$$\sum_{l \in \mathbb{N}_{l_*}} \frac{\mathbb{E}[\xi_l^2]}{\mathbb{P}_L(l)} < +\infty,$$

the estimator $\widehat{\pi}(\varphi)$ has finite variance. [Vihola, OR, 2018]

Recap: Unbiased MCMC

- Debiasing + couplings $\implies$ unbiased MCMC
- However what issues can arise?

---

1. Require complex couplings of Markov chains
2. This is by no means trivial!
3. multimodal densities $\rightarrow$ inefficiency, increased variance
4. Difficulty on more general models
5. Relationship between α and d

Motivates the use of simple coupling schemes of Markov chains!

Underdamped Langevin Dynamics

We propose the use of the ULD

$$\begin{aligned} dX_t =& V_t dt, \\ dV_t =& -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma} dW_t, \end{aligned}$$

with invariant measure

$$\pi(x, v) \propto \exp\left\{ -U(x) - \frac{\|v\|^2}{2} \right\}.$$

- Relatively easy to implement.
- Weak conditions for invariant distribution π.
- Euler-discretization well understood.

Additional Bias

- Issue: Such methods discussed $\implies$ additional bias:

$$
\begin{aligned}
x_{(k+1)h_l} &= x_{kh_l} + v_{kh_l} h_l \\
v_{(k+1)h_l} &= v_{kh_l} + \left(b(X_{kh_l}) - \gamma v_{kh_l}\right) h_l + \sigma \left(W_{(k+1)h_l} - W_{kh_l}\right).
\end{aligned}
$$

- Remedy: Exact methods (simulate exactly)?
- Actual remedy: Debiasing again!

⋮

We can exploit MLMC to gain "**good couplings**"
of unbiased estimators of

$$\pi^l(\varphi^l) - \pi^{l-1}(\varphi^{l-1})$$

⋮

We use maximal coupling.

Theory

We require various assumptions (not all stated)

- Geometric ergodicity of ULD.
- Lipschitz continuity of the kernel.
- Rates of convergence, i.e.

$$|[\pi_l - \pi](\varphi)| \leq C\|\varphi\|\Delta_l^{\beta_1}.$$

Theorem [HCJ22]; Given above assumptions, there exists a choice of PMF $\mathbb{P}_L$, such that for the metric $\tilde{d}$ in and any $\varphi \in \mathcal{B}_b(\mathsf{X}) \cap Lip_{\tilde{d}}(\mathsf{X})$, $\widehat{\pi}(\varphi)$ is an unbiased and finite variance estimator of $\pi(\varphi)$.

$\Longrightarrow$ unbiased and finite-variance estimator.

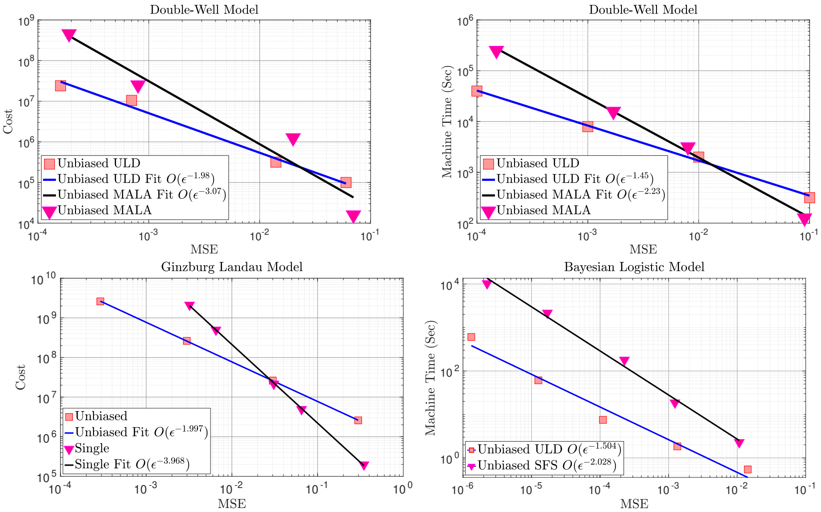
Cost for 'SL' is $\mathcal{O}(\epsilon^{-3})$ to target MSE $\mathcal{O}(\epsilon^{2})$
Cost for "U-ULD " is $\mathcal{O}(\epsilon^{-2})$

Numerical experiments

We test our ULD estimator on 3 examples:

(i) Logistic regression, (ii) Double well potential, (iii) Gindzburg-Landau

- Compare with the MALA
- Compare MSE (ϵ^2) vs Cost (MLMC framework)



Double-Well Model
Cost
MSE
Unbiased ULD
Unbiased ULD Fit $O(\epsilon^{-1.98})$
Unbiased MALA Fit $O(\epsilon^{-3.07})$
Unbiased MALA
Double-Well Model
Machine Time (Sec)
MSE
Unbiased ULD
Unbiased ULD Fit $O(\epsilon^{-1.45})$
Unbiased MALA Fit $O(\epsilon^{-2.23})$
Unbiased MALA
Ginzburg Landau Model
Cost
MSE
Unbiased
Unbiased Fit $O(\epsilon^{-1.997})$
Single
Single Fit $O(\epsilon^{-3.968})$
Bayesian Logistic Model
Machine Time (Sec)
MSE
Unbiased ULD $O(\epsilon^{-1.504})$
Unbiased SFS $O(\epsilon^{-2.028})$

Part II: Unbiased Kinetic Langevin Monte Carlo

NKC, B. Leimkhuler, D. Paulin and P. Whalley (UoE) [ArXiv 23]

Issues/Improvements

- We have 4 chains (2 chains within the telescoping sum)
- Can **exploit** higher order numerical schemes
- Gain more theoretical insights
- Extension to stochastic gradients

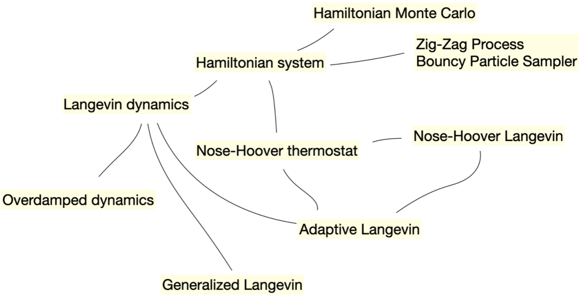


Figure 2: a menagerie of sampling methods

Discretization Schemes

Kinetic Langevin dynamics have many discretizations:

- Euler-Maruyama (EM)
- BAOAB, OBABO, OABAO [Matthew, Leimkhuler 13]
- Stochastic Euler scheme [Buckholz 80]
- BBK Scheme [Brunget et al. 84]
- **UBU/BUB** [Zapatero 21] $\sim \mathcal{O}(h^2)$.

$$\begin{pmatrix} dx \\ dv \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ -\nabla U(x)dt \end{pmatrix}}_{\mathcal{B}} + \underbrace{\begin{pmatrix} vdt \\ -\gamma vdt + \sqrt{2\gamma}dW_t \end{pmatrix}}_{\mathcal{U}},$$

We present a new unbiased method called: **UBUBU**

Role of Metropolization

Kinetic Langevin dynamics have many discretizations:

- Discretization of SDEs do not exactly converge to the correct π (require MH step)
- Examples: MALA, HMC, RHMC
- Curse of dimensionality: dim-dep step-size restrictions (for α)

Algorithm	**Gradient Evaluations**	**Conditions**	**Reference**
MALA	$\mathcal{O}(d^{1/2})$	$h = \mathcal{O}(d^{-1/2})$	Lee 21
HMC	$\mathcal{O}(d^{1/4})$	$h = \mathcal{O}(d^{-1/4})$, warm start	Chen 23
RHMC	$\mathcal{O}(d^{1/4})$	$h = \mathcal{O}(d^{-1/4})$, warm start, Gaussian target	Apers 22
UBUBU	$\mathcal{O}(d^{1/4})$	$h_0 = \mathcal{O}(d^{-1/4})$	**this work**

Stochastic Gradients

- We have looked at the stochastic gradient variant:

$$G(x,\omega) = \nabla U_0(x) + \sum_{i=1}^{N_0} \nabla U_i(\hat{x}) + \frac{N}{b} \sum_{i \in \omega} (\nabla U_i(x) - \nabla U_i(\hat{x})).$$

  where x^* is the minimizer, and $\omega = (\omega_1, \ldots, \omega_b)$, are uniform i.i.d.

- Another possibility is the use an approximate gradient:

$$G(x) = \nabla U(\hat{x}) + \nabla^2 U(x^*)(x - \hat{x}).$$

We consider a different telescoping sum,

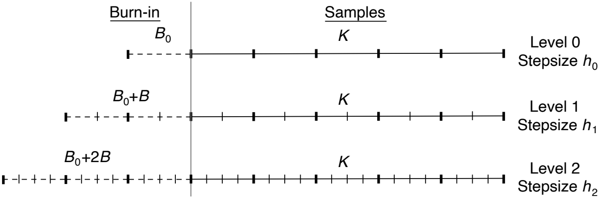
$$\mu(f) = \tilde{\mu}_{h_0}(f) + \sum_{l=0}^{\infty} (\tilde{\mu}_{h_{l+1}}(f) - \tilde{\mu}_{h_l}(f)),$$

where

$$D_{l,l+1} := \frac{1}{K} \sum_{i=1}^{K} [f({z'}_i^{(l,l+1)}) - f(z_i^{(l,l+1)})],$$

$$S_{l,l+1} = \frac{1}{\mathbb{E}(N_{l,l+1})} \sum_{r=1}^{N_{l,l+1}} D_{l,l+1}^{(r)}.$$

From the definitions, it follows that $\mathbb{E}S_{l,l+1}$ is an unbiased estimator.



UBU $\mathcal{W}$-contraction

Theorem: Suppose that U is m-strongly convex and M-∇Lipschitz. Let

$$a = \frac{1}{M}, \quad b = \frac{1}{\gamma}, \quad c_2(h) = \frac{mh}{4\gamma}, \quad c(h) = \frac{mh}{8\gamma}.$$

Let P_h denote the transition kernel for a step of UBU with stepsize h. For all $\gamma \geq \sqrt{8M}$, $h < \frac{1}{2\gamma}$, $1 \leq p \leq \infty$, $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^{2d})$, $n \in \mathbb{N}$,

$$\mathcal{W}_{p,a,b}\left(\nu P_h^n, \mu P_h^n\right) \leq (1 - c_2(h))^{n/2}\, \mathcal{W}_{p,a,b}\left(\nu, \mu\right) \leq (1 - c(h))^n\, \mathcal{W}_{p,a,b}\left(\nu, \mu\right).$$

P_h has a unique invariant measure π_h satisfying
that $\pi_h \in \mathcal{P}_p(\mathbb{R}^{2d})$ for all $1 \leq p \leq \infty$.

Analysis (Some...)

- CLT, finite variance/unbiased estimator

Theorem: Suppose various assumptions hold, &

$$\gamma \geq \sqrt{8M}, \quad h_0 \leq \frac{1}{\gamma} \cdot \frac{m}{264M}, \quad B \geq \frac{16\log(4)\gamma}{mh_0}, \quad B_0 \geq \frac{16\gamma}{mh_0} \log\left(\frac{c_{\mu_0}+1}{\sqrt{M}\gamma h_0^2}\right).$$

Then UBUBU is a finite variance and unbiased estimator. Moreover, it satisfies a CLT as $N \to \infty$, with asymptotic variance bound

$$\sigma_S^2 \leq \frac{C(m, M, M_1, \gamma, c_N, \phi_N)}{Kh_0}\left(1 + \frac{1}{h_0 K} + dh_0^4\right).$$

Analysis (Some...)

- Dimension-indepdent result for production distributions

Theorem: Given similar assumptions, and f is of the form

$$f(x,v) = g(\langle w^{(1)}, x\rangle, \ldots \langle w^{(r)}, x\rangle),$$

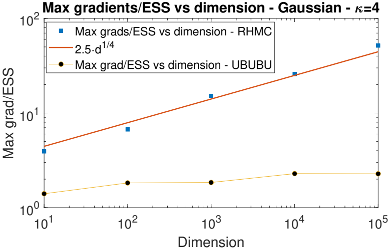
where $g : \mathbb{R}^r \to \mathbb{R}$ is 1-Lipschitz, and $w^{(1)}, \ldots, w^{(r)} \in \mathbb{R}^d$. Moreover, it satisfies a CLT as $N \to \infty$, with asymptotic variance bound

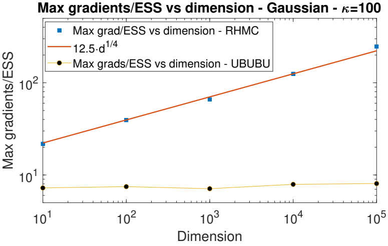
$$\sigma_S^2 \leq \frac{C(m, M, M_1, \gamma, r, c_N, \phi_N)}{Kh_0} \sum_{1 \leq i \leq r} \|w^{(i)}\|^2.$$

- Also show (i) big data limit (ii) extensions to SG/approx grad

Gaussian Target Example

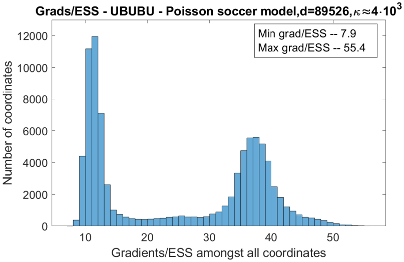
$$\text{Gaussian distribution}: \quad \pi(x) = \prod_{i=1}^{d} \pi_0(x) \frac{e^{-v_i^2/2}}{\sqrt{2\pi}}.$$

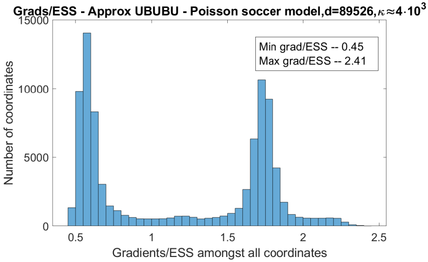


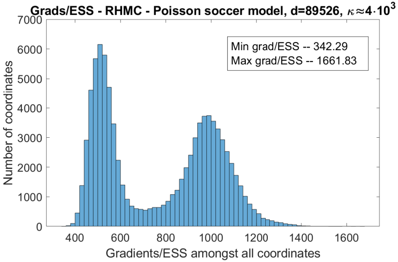


Poisson Football Model

This example is from [Koopman and Lit, 15], a Poisson random effect model.







Summary and Outlook

- Proposed new Unbiased estimator for sampling.
- Focus was on the use of ULD (Kinetic).
- Provided theorem and tested applications with comparisons.

- Unbiased estimation for constrained domains?
- Extension to non-convex setting which is natural.
- Other Bayesian applications, based on point above.

More details in:
Unbiased estimation with underdamped Langevin dynamics,
H. Ruzaquat, N. K. C and A. Jasra.
arXiv e-prints, 2022. [arXiv:2206.07202] (Accepted by SISC)

Unbiased kinetic Langevin Monte Carlo,
N. K.C, B. Leimkuhler, D. Paulin and P. Whalley.
arXiv e-prints, 2023.