Error analysis for parallel Monte Carlo estimation from many short Markov chains

Austin Brown [1]

Postdoc at the University of Toronto,
Toronto, Ontario

September 23, 2024

[1] ad.brown@utoronto.ca

Motivation

Goal in MCMC

The goal is to estimate integrals with respect to a probability measure via simulation with modern computation. This is crucial for modern Bayesian inference.

For some real-valued function $\varphi : \mathbf{X} \to \mathbb{R}$, estimate

$$\int_{\mathbf{X}} \varphi d\Pi$$

MCMC in the single-chain regime

Simulate realizations of a *single* Markov chain $X_0, \ldots, X_t, \ldots$ with unique invariant distribution Π until convergence (approximate).

The marginal realizations stabilize to realizations from Π. Use the empirical average to estimate an expectation:

$$\frac{1}{T}\sum_{s=1}^{T}\varphi(X_{t+s}) \approx \int_{\mathbf{X}} \varphi d\Pi.$$

Single-chain regime theoretical properties

- Theoretically solid foundation
- The dependence in the Markov chain restricts many advantages of modern parallel computation
- Knowing when the Markov chain converged is an extremely challenging problem
- Obtaining explicit and useful convergence rates is challenging

A generally long serial Markov chain path simulation

The many-short-chains regime

Simulate realizations from *multiple* independent *short* Markov chain paths $X_0^m, \ldots, X_t^m$ for $m = 1, \ldots, M$ and ensemble them.

Ensemble examples:

- Take the average of $X_t^1, \ldots, X_t^m$
- Take the average of averages

Many-short-chains regime properties

- Short Markov chain length = fast simulation
- Parallel simulations able to utilize modern parallel compute

A possibly fast, parallel simulation

- The bias can be large or even unknown due to the short simulation length = estimate incorrectly
- Many-short-chains has been used [Gelman and Rubin, 1992] and debated, and criticized due to its weaknesses [Geyer, 1992].

Contribution of this work

Construct an ensemble MCMC estimator with an **error guarantee** in the **many-short-chains** regime.

i.e. **short** Markov chain simulations in **parallel**.

Many-short-chains (MSC) estimator

MSC estimator requirements

- Let $Y_1, \ldots, Y_N$ independently from an importance sampling proposal
- Construct an initial distribution for the Markov chains using the self-normalized importance weights
- Use independent Markov chains $X_0^m, X_1^m, \ldots, X_t^m, \ldots$ for $m = 1, \ldots, M$ initialized with X_0^m from this initial distribution.
- An explicit set C (we will use a drift condition for this later).

Note: This initialization does not require normalizing constant of Π.

MSC estimator construction

τ_C is the first return time to the set C and

$$S_{\tau_C}(\varphi) = \begin{cases} \sum_{k=1}^{\tau_C} \varphi(X_k) & X_0 \in C \\ 0 & X_0 \notin C. \end{cases}$$

The *MSC estimator* is the average over these independent sums of Markov chains, that is,

$$\overline{S_{\tau_C}(\varphi)} = \frac{1}{M} \sum_{m=1}^{M} S_{\tau_C}^{m}(\varphi).$$

Mean squared error analysis

Geometric drift condition

There is a function $V \geq 1$ and constants $\gamma \in (0,1)$ and $K > 0$ such that
$$\mathbb{E}\left[V(X_t) \mid X_{t-1}\right] \leq \gamma V(X_{t-1}) + K.$$

A geometric drift condition ensures the Markov chain visits

$$C = \{x \in \mathbf{X} : V(x) \leq R\} \tag{1}$$

for any $R > K/(1-\gamma)$.

Mean squared error analysis

Theorem

Under the geometric drift condition,

$$\sup_{|\varphi| \leq \sqrt{V}} \mathbb{E}\left[\overline{S_{\tau_C}(\varphi)} - \int \varphi d\Pi\right]^2 \lesssim \frac{R+K}{M(2-\gamma_R)^2} + \frac{(R+K)^2 \int w d\Pi}{N(1-\gamma_R)^2}$$

where $\gamma_R > \gamma$.

- The actual constants are explicit and non-asymptotic
- If $V(\cdot) \geq \|\cdot\|_2^2$ can estimate the posterior mean
- Does not require convergence analysis of the Markov chain

MSC estimator properties

- The drift condition ensures the Markov chain simulation is short in length. Not always available e.g. Metropolis-Hastings.
- Requires importance sampling proposal and drift condition, but does not require convergence analysis

MSC estimator concentration

Under uniformly bounded importance weights and a multiplicative drift condition, there is a stronger sub-Gaussian concentration inequality.

Applications

Toy example

Consider the autoregressive process

$$X_t = \rho X_{t-1} + \sqrt{1-\rho^2}\xi_t \qquad \xi_t \sim N(0, I) \text{ independent}$$

has invariant distribution $N(0, I)$.

MSC construction

- Use the autoregressive process as the Markov chain
- Use importance sampling proposal $N(0, (1/2 + h)I_d)$.
- For any $r > 1$, define the set

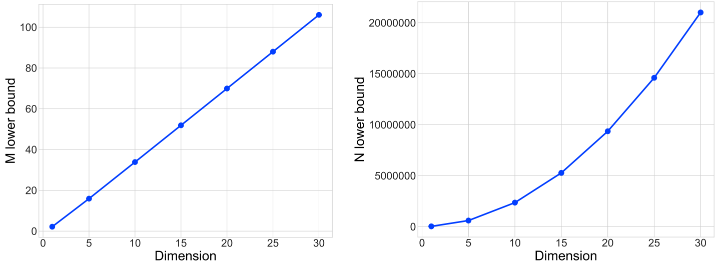
$$C_r = \{\|x\|_2^2 \leq rd\}$$

MSE error bound

If N, M sufficiently large, then for $|\varphi| \leq \|\cdot\|_2$,

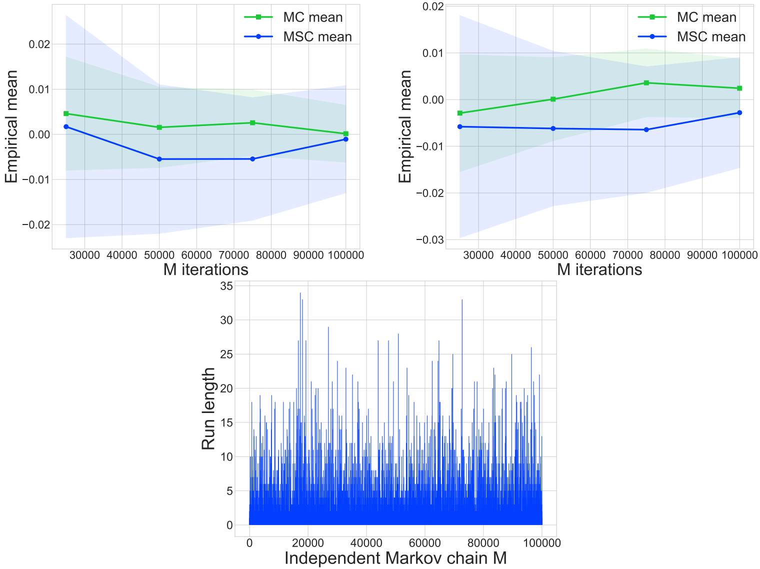
$$\overline{S_{\tau_{C_r}}(\varphi)} \approx \mathbb{E}[\varphi(Z)] \qquad\qquad Z \sim N(0, I)$$

with high probability.



Simulation result

Use $N = 10^6$ and $M = 10^5$ in dimension 2 to estimate the mean.



Predicting cardiovascular disease

The aim is to predict cardiovascular disease from a data set provided by the Cleveland Clinic [Detrano et al., 1989]. The data consists of 303 patients with binary responses determining if cardiovascular disease is present and 21 covariates on patient characteristics.

Bayesian logistic regression

Consider Bayesian logistic regression with a Gaussian prior

$$\beta \sim N(0, \Sigma)$$

The Pólya-Gamma Gibbs sampler [Nicholas G. Polson and Windle, 2013] is popular but the convergence rate [Choi and Hobert, 2013] can be problematic in moderate dimensions.

MSC construction

- Importance sampling proposal $N(\beta_n^*, (1/2+h)\Sigma)$ with $h \in (0, 1/2]$.
- Marginal Pólya-Gamma Gibbs Markov chain $(\beta_t)_t$.
- For any $r > 1$,

$$C_r = \{\|\beta\|_2^2 \leq rL\}. \tag{2}$$

where $L = \|\Sigma\|_2^2 \left\|X^T(Y - 1/21_d)\right\|_2^2$.

MSC error analysis

Using the MSC estimator with this Pólya-Gamma Gibbs sampler:

Proposition

Then for $|\varphi| \leq \|\cdot\|_2$,

$$MSE[S_{\tau_{C_r}}(\varphi)] \lesssim \frac{rL}{M(2-\gamma_r)^2} + \frac{[rL]^2 \left(\frac{1}{2\sqrt{2h}} + \sqrt{\frac{h}{2}}\right)^d}{N(1-\gamma_r)^2}$$

where $\gamma_r \in (0,1)$.

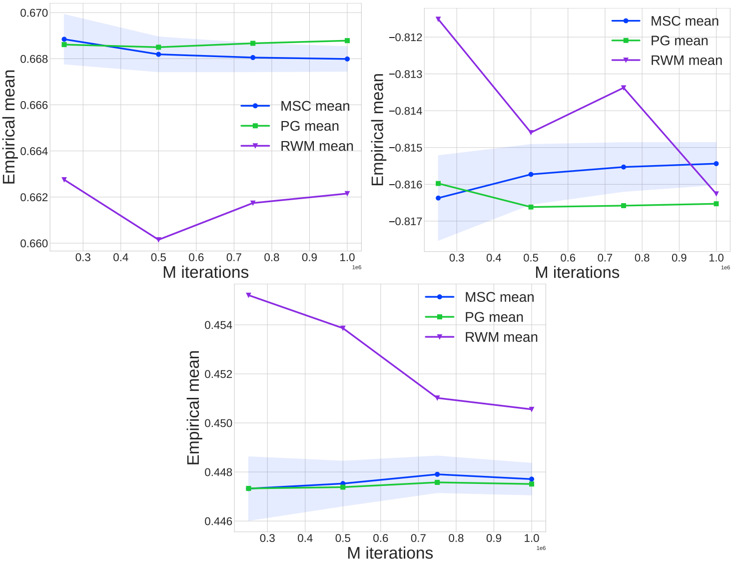
- The bound is computable
- Since the drift condition could be improved this will result 1 step runs of the Markov chain.

Simulation setup

Use $N = 10^7$ and $M = 10^6$ to estimate the posterior mean for the cardiovascular data. Use some reasonable tuning parameters and prior choice.

Coefficients β_1 indicating male versus female patients, β_2 for the number of major blood vessels colored by flouroscopy, and β_3 for resting blood pressure in mm/Hg on admission to the hospital.

Simulation result



Summary

- Developed an estimator guaranteed in the *many-short-chains* regime and error analysis that does not depend on convergence of the Markov chain.
- The bounds can have issues scaling to high dimensional problems
- The bounds appear useful compared to importance sampling bounds for unbounded functions [Agapiou et al., 2017, Theorem 2.3].

References I

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

Hee Min Choi and James P. Hobert. The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7(none):2054–2064, 2013.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. ISSN 0002-9149.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4): 457–472, 1992.

References II

Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

James G. Scott Nicholas G. Polson and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.