Unbiased Multilevel Monte Carlo methods for intractable distributions: MLMC meets MCMC

Guanyang Wang[*] and Tianze Wang[†]

December 14, 2021

Abstract

[GW: To be added]

1 Introduction

Monte Carlo methods provide unbiased estimators for the expected value of a distribution. In practice, however, the distribution may be infeasible to sample from, and the quantity of interest may not be the expected value. Examples include the ratio of normalizing constants of two distributions in statistical physics; the quantiles of a distribution or the mode of a density in statistics; the minimum of the expected loss over a family of probability measures in stochastic optimization, and so on. [GW: Expand this] In all the aforementioned applications, the distributions can be intractable, and the interested quantities are functionals of measures rather than expectations.

Generally, most inference problems can be viewed as estimating a quantity of the form $\mathcal{T}(\pi)$, where π stands for one or a family of distributions, and $\mathcal{T}$ is a scalar or verctor-valued functional of π. In particular, we are looking for unbiased estimators as unbiasedness paves a convenient way to parallel computation[GW: Expand this].

Computational challenges appear in both components of the pair $(\mathcal{T}, \pi)$. Fortunately, recent works provide promising solutions when one component is easy while the other is difficult. When $\mathcal{T}(\pi) := \int f(x)\pi(\mathrm{d}x)$ is an integral but π is difficult to sample from, Jacob, O'Leary, and Atchadé, 2020 propose a general method for designing unbiased estimators using coupling of Markov chains. This unbiased MCMC method has been

[*] Department of Statistics, Rutgers University, New Brunswick, USA. Email: guanyang.wang@rutgers.edu

[†] Department of Statistics, Rutgers University, New Brunswick, USA. Email: tw522@scarletmail.rutgers.edu

Alphabetical authorship.

successfully generlized and applied in different contexts. We refer the readers to Heng and Jacob [2019], Middleton et al. [2020], Biswas et al. [2019], Ruiz et al. [2020], Wang et al. [2021] and the references therein for recent progresses. When π can be sampled perfectly, but $\mathcal{T}(\pi) := g(\int f(x)\pi(\mathrm{d}x))$ is a functional of the expectation, the state of the art debiasing technique is the unbiased Multilevel Monte Carlo (MLMC) developed by McLeish, Glynn, Rhee, and Blanchet [Blanchet and Glynn, 2015, Rhee and Glynn, 2015, McLeish, 2011] which is a randomized version of the celebrated MLMC methods pioneered by Heinrich and Giles [Heinrich, 2001, Giles, 2008, 2015]. Unbiased MLMC methods have also found many applications including gradient estimation [Shi and Cornish, 2021], optimal stopping [Zhou et al., 2021], robust optimization [Levy et al., 2020]. In summary, the unbiased MCMC method assumes easy $\mathcal{T}$ (an integral operator) but difficult π, and the unbiased MLMC method assumes easy π (perfectly simulable) but difficult $\mathcal{T}$. Both assumptions can be vialoted in many applications as discussed at the beginning, which motivates this work.

In this article, we present a step towards designing unbiased estimators of $\mathcal{T}(\pi)$ for general $(\mathcal{T}, \pi)$ pair by combining the ideas of the unbiased MCMC and MLMC methods. We propose unbiased estimators of functional of expectations, i.e., $T(\pi) = g(m(\pi)) := g(\mathbb{E}_\pi[f(X)])$ where π is a d-dimensional probability measure, $f : \mathbb{R}^d \to \mathbb{R}^m$ is a deterministic map, and $g : \mathbb{R}^m \to \mathbb{R}$ is a deterministic function [1]. Other technical assumptions including the domain and smoothness of g, the existence of moments will be made clear in the subsequent sections. The unbiased estimator is easily parallelable. It has both finite variance and finite computational cost for a general class of problems, which implies a 'square root convergence rate' that matches the oracle rate given by the Central Limit Theorem. [GW: Discuss something on the domain and the algorithm trick] Moreover, some technical assumptions on g relax the standard 'linear growth' assumption in Blanchet and Glynn [2015] and Blanchet et al. [2019], which may be of independent interest.

The rest of this paper is organized as follows. Section 1.1 introduces the notations. In Section 2 we discribe the high-level idea behind our method without touching much details. The connections between unbiased MCMC and MLMC methods will also be clear in Section 2. In Section 3 we formally propose our estimator, state the assumptions, and prove its theoretical properties. In Section 4 we implement our method on two [GW: several?] examples to study its empirical performance. We conclude this paper in Section ??. Technical details of some proofs are deferred to the Appendix [GW: add appendix].

1.1 Notations

Throughout this article, we fix the notation g to denote a function from its domain $\mathcal{D} \subset \mathbb{R}^m$ to $\mathbb{R}$ of our interest. The domain $\mathcal{D}$ plays an important role in both algorithm design and theoretical analysis, which will be addressed in Section [GW: add]. We write π as a d-dimensional probability measure, and $\pi_1, \cdots, \pi_d$ for its marginal distributions. We denote by $m_f(\pi) := \mathbb{E}_\pi[f(X)]$ the expected value/vector of f under π, and write it as $m(\pi)$ when it is unlikely to cause confusion. The L^p norm of a vector $v \in \mathbb{R}^d$ is written as $\|v\|_p := \left(\sum_{i=1}^d v_i^p\right)^{1/p}$. For the L^2 norm, we simply write

[1] For simplicity, we only consider scalar-valued g in this paper, though our method can be naturally generalized to vector-valued functions.

$\|v\| := \|v\|_2 = \sqrt{\sum_{i=1}^d v_i^2}$. The geometric distribution with success probability r is denoted by $\mathsf{Geo}(r)$, and $p_n = p_n(r) := \mathbb{P}(\mathsf{Geo}(r) = n) = (1-r)^{n-1}r$. The uniform distribution on $[0,1]$ is denoted by $\mathsf{U}[0,1]$. The multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ is denoted by $\mathsf{N}(\mu, \Sigma)$. We adopt the convention that $\sum_{i=m}^n a_i = 0$ if $m > n$. Given a set $A \subset \mathbb{R}^d$, we denote by A° all the interior points of A. For a differentiable function $h : \mathbb{R}^d \to \mathbb{R}$, we denote by $Dh := (\frac{\partial h}{\partial x_1}, \frac{\partial h}{\partial x_2}, \cdots, \frac{\partial h}{\partial x_d})$ the gradient of h. [GW: to be finished, add tv distance]

2 A Simple Identity: Unbiased MCMC meets MLMC

Consider the task of designing unbiased estimators of $g(m(\pi)) = g\Big(\mathbb{E}_\pi[f(X)]\Big)$. The problem is extensitively studied in the literature when one can draw independent and identically distributed (*i.i.d.*) samples from π. Unbiased estimators are known to exist, or not exist under different contexts [Keane and O'Brien, 1994, Jacob and Thiery, 2015]. Different debiasing techniques including Bernstein polynomials [Nacu and Peres, 2005], Taylor polynomials [Blanchet et al., 2015], and the MLMC method [Blanchet and Glynn, 2015, Vihola, 2018] have been proposed and analyzed. The core idea in all these techniques is to write $g(m(\pi))$ as an infinite series $\sum_{k=1}^\infty a_k$. One can then choose a random level k with probability p_k and construct the importance sampling-type estimator $\hat{a}_k/p_k$. Suppose each $\hat{a}_k$ is unbiased for a_k, then $\hat{a}_k/p_k$ is generally unbiased for $\sum_{k=1}^\infty a_k$. Among the existing methods, the unbiased MLMC framework seems to work with the greatest generality, as it does not require the knowledge on the derivatives of g.

When π is infeasible to sample from, our first observation is based on the following simple identity:

$$g(m(\pi)) = g(m(\tilde{\pi})) \tag{1}$$

for every $\tilde{\pi}$ which satisfies $m(\tilde{\pi}) = m(\pi)$. In other words, if we are able to *i.i.d.* sample unbiased estimators of $m(\pi)$, then the unbiased MLMC methods (and all the debiasing tricks mentioned above) can be directly applied.

After observing (1), it suffices to construct unbiased estimators of $m(\pi)$ provided that π cannot be directly simulated. The unbiased MCMC framework provides us natural solutions. Suppose there exists a Markov chain with transition kernel P that targets π as stationary distribution. It is often possible to construct a pair of coupled Markov chains $(Y, Z) = (Y_t, Z_t)_{t=1}^\infty$ that both evolve according to P. By design, if the pair (Y_t, Z_{t-1}) meets at some random time τ and stay together after meeting, then the Jacob-O'Leary-Atchadé (JOA) estimator is unbiased for $m(\pi)$. Therefore, we are able to unbiasedly estimate $g(m(\pi))$ using the following two-step simulation strategy. The overall workflow is described in the Figure 1 below. The unbiased MCMC algorithm is used here as a generator for random variables with expectation $m(\pi)$. Therefore we can use the outputs of the unbiased MCMC algorthm as inputs for the unbiased MLMC approach, and eventually construct an unbiased estimator of $g(m(\pi))$.

3 Unbiased estimators for functions of expectation

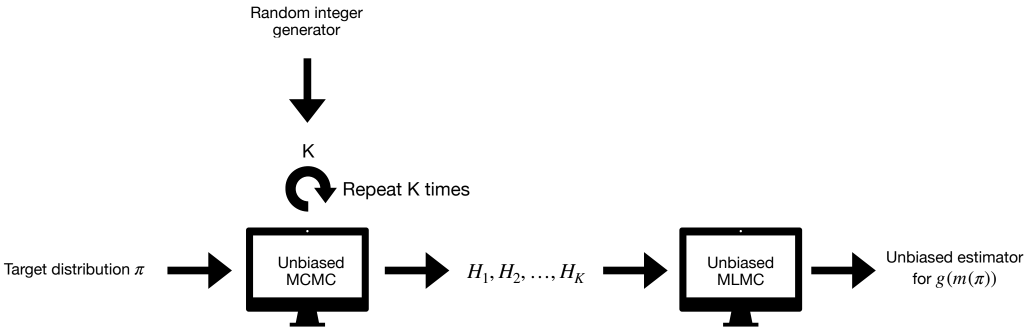


Figure 1: The workflow for constructing an unbiased estimator of $g(m(\pi))$.

3.1 Constructing the unbiased estimator

3.1.1 The Jacob-O'Leary-Atchadé (JOA) estimator of $m(\pi)$

Let Ω be a Polish space equipped with the standard Borel σ-algebra $\mathcal{F}$. Let $P : \Omega \times \mathcal{F} \to [0, 1]$ be the Markov transition kernel that leaves π as stationary distribution. The Jacob-O'Leary-Atchadé (JOA) estimator uses a coupled pair of Markov chains that both has transition kernel P. Formally, the coupled pair $(Y, Z) = (Y_t, Z_t)_{t=1}^{\infty}$ can be viewed as a Markov chain on the produce space $\Omega \times \Omega$. The transition kernel $\bar{P}$, which is also called the coupling of (Y, Z), satisfies

$$\bar{P}((x, y), A \times \Omega) = P(x, A), \bar{P}((x, y), \Omega \times B) = P(y, B)$$

for every $x, y \in \Omega$ and $A, B \in \mathcal{F}$. The coupled chain starts with $Y_0 \sim \pi_0, Y_1 \sim P(Y_0, \cdot)$ and $Z_0 \sim \pi_0$ independently. Then at each step $t \geq 2$, one samples $(Y_t, Z_{t-1}) \sim \bar{P}((Y_{t-1}, Z_{t-2}), \cdot)$. Suppose the coupling $\bar{P}$ is 'faithful' [Rosenthal, 1997], meaning that there is a random but almost surely finite time τ such that $Y_\tau = Z_{\tau-1}$, and $Y_t = Z_{t-1}$ for every $t \geq \tau$. Then for every k,

$$H_k(Y, Z) := f(Y_k) + \sum_{i=k+1}^{\tau-1} (f(Y_i) - f(Z_{i-1})) \tag{2}$$

is an unbiased estimator. For any fixed integer $m \geq k$, the following 'time-averaged' estimator $H_{k:m}(Y, Z) := (m - k + 1)^{-1} \sum_{l=k}^{m} H_l(Y, Z)$ is the average of $m - k + 1$ unbiased estimators, which retains unbiasedness and reduces the variance. The unbiasedness of $H_k(Y, Z)$ is justified by the following informal calculation in Jacob et al. [2020]:

$$
\begin{aligned}
m(\pi) = \lim_{n\to\infty} \mathbb{E}[f(Y_n)] &= \mathbb{E}[f(Y_k)] + \sum_{n=k+1}^{\infty} (\mathbb{E}[f(Y_n)] - \mathbb{E}[f(Y_{n-1})]) \\
&= \mathbb{E}[f(Y_k)] + \sum_{n=k+1}^{\infty} \mathbb{E}[f(Z_n) - f(Y_{n-1})] \\
&= \mathbb{E}[f(Y_k)] + \sum_{n=k+1}^{\tau-1} \mathbb{E}[f(Z_n) - f(Y_{n-1})] = \mathbb{E}[H_k(Y, Z)].
\end{aligned}
$$

The rigorous proof requires assumptions on the target π, and the distribution of τ. We refer the readers to Jacob et al. [2020], Middleton et al. [2020] and our appendix for more details. More sophisticated unbiased estimators using L-lag coupled chains are discussed in Biswas et al. [2019], but the main idea remains the same.

3.1.2 Unbiased estimator of $g(m(\pi))$

Suppose we get access to a routine $\mathcal{S}$ such as the JOA estimator in Section 3.1.1 which outputs unbiased estimators of $m(\pi)$. The estimator of $g(m(\pi))$ can then be constructed by the randomized MLMC method. Let $H_1, H_2, \cdots, H_{2m}$ be a sequence of *i.i.d.* random variables. We let

$$S_H(2m) := \sum_{k=1}^{2m} H_i \tag{3}$$

be the summation of all the $2m$ terms, and let

$$S_H^{\mathsf{O}}(m) := \sum_{k=1}^{m} H_{2k-1} \qquad S_H^{\mathsf{E}}(m) := \sum_{k=1}^{m} H_{2k} \tag{4}$$

be the summation of all the odd and even terms, respectively. The estimator is described by Algorithm 1 below.

Algorithm 1 Unbiased Multilevel Monte-Carlo estimator

```
Input:
  • A subroutine $\mathcal{S}$ for generating unbiased estimators of $m(\pi)$
  • A function $g : \mathcal{D} \to \mathbb{R}$
  • The parameter $p$ for geometric distribution
Output: Unbiased estimator of $g(m(\pi))$
  1. Sample $N$ from the geometric distribution $\mathsf{Geo}(p)$
  2. Call $\mathcal{S}$ for $2^N$ times and label the outputs by $H_1, ..., H_{2^N}$
  3. Calculate the quantities $S_H(2^N)$, $S_H^{\mathsf{O}}(2^{N-1})$ and $S_H^{\mathsf{E}}(2^{N-1})$ by (3),(4)
  4. Calculate $\Delta_N = g\left(S_H(2^N)/2^N\right) - \frac{1}{2}\left(g\left(S_H^{\mathsf{O}}(2^{N-1})/2^{N-1}\right) + g\left(S_H^{\mathsf{E}}(2^{N-1})/2^{N-1}\right)\right)$
Return: $W = \Delta_N/p_N + g(H_1)$.
```

Again, the following informal calculation justifies the unbiasedness of W.

$$\begin{aligned}\mathbb{E}\left[W\right] &= \mathbb{E}\left[g(H_1)\right] + \mathbb{E}\left[\Delta_N/p_N\right] \\ &= \mathbb{E}\left[g(H_1)\right] + \mathbb{E}[\mathbb{E}\left[\Delta_N/p_N \mid N\right]] \\ &= \mathbb{E}\left[g(H_1)\right] + \sum_{n=1}^{\infty}\mathbb{E}[\Delta_n] \\ &= \mathbb{E}\left[g(H_1)\right] + \sum_{n=1}^{\infty}(\mathbb{E}[g(S_H(2^n)/2^n)] - \mathbb{E}[g(S_H(2^{n-1})/2^{n-1})]) \\ &= \lim_{n\to\infty}\mathbb{E}[g(S_H(2^n)/2^n)] = g(m(\tilde{\pi})) = g(m(\pi)),\end{aligned}$$

where $\tilde{\pi}$ is the distribution of each H_i.

Algorithm 1 differs from the original unbiased MLMC design Blanchet and Glynn [2015] as it relaxes the assumption '$i.i.d.$ samples from π' by 'unbiased estimator of $m(\pi)$'. It also provides practical methods for finding the estimator via the JOA estimator Jacob et al. [2020] described in Section 3.1.1. It is worth mentioning that, any unbiased estimator of $m(\pi)$, including but not limited to the JOA estimator (see also Glynn and Rhee [2014]), can be feed into Algorithm 1 as a subroutine. On the other hand, we find the JOA estimator is by far the most general framework for constucting unbiased estimators of $m(\pi)$ given intractable π. We will implicitly assume the subroutine $\mathcal{S}$ we choose is the JOA estimator in the subsequent sections. [GW: Check if this sentence can be deleted after writing the article.]

Now we have discussed the main ideas of our design, but several important problems remain. In Section 3.2 we discuss the problem regarding the domain of g. We provide a transformation which avoids the domain problem for a general class of functions. In Section 3.3 we give theoretical justifications of our method. [GW: Change/move this paragraph.]

3.1.3 Unbiased estimator of polynomials and other special functions

Section 3.1.2 provides us a relatively general framework for unbiased estimators of $g(m(x))$. In some situations where the target function g has certain nice properties, the unbiased estimators can be easily obtained without resorting to the unbiased MLMC framework. For example, if $g(x) = x^k$ is a univariate monomial function, one can simply call the unbiased MCMC algorithm k times and obtain $H_1, \cdots, H_k$. The estimator $\prod_{l=1}^{k} H_l$ will then be unbiased for $m(\pi)^k$. The argument above can be naturally extended to the case where g is a multivariate polynomial function. We use the multi-index $k = (k_1, \cdots, k_m)$ with $\sum_{i=1}^{m} k_i \leq n$, and $x^k = x_1^{k_1}x_2^{k_2}\cdots x_m^{k_m}$. Let $g(x) = \sum_{k\leq n}\alpha_k x^k$ denote a multivariate polynomial with degree at most n, the unbiased estimator of $g(m(\pi))$ can be constructed as follows. First, we call the unbiased MCMC subroutine $\mathcal{S}$ for n times and label the outputs by $H_1, \cdots, H_n$, each is an independent unbiased estimator of $m(\pi)$. Then for each $k = (k_1, \cdots, k_m)$ we calculate the quantity

$$\hat{H}(k) = \prod_{l_1=1}^{k_1} H_{l_1,1} \prod_{l_2=k_1+1}^{k_1+k_2} H_{l_2,2} \cdots \prod_{l_m=k_1+\cdots+k_{m-1}+1}^{k_1+\cdots+k_m} H_{l_m,m},$$

where $H_{a,b}$ stands for the b-th coordinate of $H_a \in \mathbb{R}^d$. It is then clear from the independence of $H_1, \cdots, H_n$ that $\mathbb{E}[\hat{H}(k)] = m(\pi)^k$. Finally we output $\sum_k \alpha_k \hat{H}(k)$ which is unbiased for $g(m(\pi))$ by the linearity of expectation. It is different from Algorithm 1 as it requires a fixed number of calls for $\mathcal{S}$, and it does not require calculating the consecutive difference Δ_k.

When $g : \mathbb{R} \to \mathbb{R}$ is a real analytic function on $\mathcal{D}$, i.e., $g(x) = \sum_{n=0}^{\infty} a_i(x-a)^n$ for some real number a. Suppose $\tilde{N}$ is a non-negative integer-valued random variable with $\mathbb{P}(\tilde{N} = k) = q_k$. The unbiased estimator for $g(m(\pi))$ can be constructed by first generating $\tilde{N}$, and then call the unbiased MCMC subroutine $\mathcal{S}$ for $\tilde{N}$ times with outputs $H_1, \cdots H_{\tilde{N}}$. The final estimator can be expressed by

$$\frac{a_{\tilde{N}}}{q_{\tilde{N}}} \cdot \frac{\prod_{j=1}^{\tilde{N}}(H_j - a)}{\tilde{N}!}$$

which follows from the idea in Blanchet et al. [2015]. In particular, when $g(x) = e^x$ and $\tilde{N}$ follows from the Poisson distribution, the estimator is known as the 'Poisson estimator' which is used in both physics and statistics, see Wagner [1987], Papaspiliopoulos [2009], Fearnhead et al. [2010] for more discussions. Albeit useful in many cases, the power-series-type estimators generally have strong assumptions on the smoothness of the target function. It also requires the knowledge of all the higher-order derivatives of g, which is generally infeasible when g is either a complicated univariate function, or a multivariate function.

Therefore, thoughout this paper we will mostly focus on the unbiased MLMC framework which has the greatest generality. This subsection intends to remind our readers that there may be easier choices when the target function g is 'nice' enough.

3.2 The domain problem and the δ-transformation

There is an extra subtly in implementing Algorithm 1. Besides requiring H to be an unbiased estimator of $m(\pi)$, Algorithm 1 implicitly requires the range of $S_H(m)/m$ is a subset of the domain of g, as otherwise the algorithm cannot be implemented. This constraint is naturally satisfied when $g : \mathcal{D} \to \mathbb{R}$ has domain $\mathcal{D} = \mathbb{R}^m$, such as $g(x) = e^x$, or $g(x_1, x_2) = \max\{x_1, x_2, 1\}$ and so on. However, many natural functions are not defined on the whole space, such as $g(x) = 1/x$, or $g(x_1, x_2) = x_1/x_2$. These functions arise in statistical applications such as inference on doubly-intractable problems Murray et al. [2006], Lyne et al. [2015], estimating the ratio of normalizing constants Meng and Wong [1996]. Unfortunately, Algorithm 1 cannot be implemented if $S_H(m)/m$ falls outside the domain of g.

Consider a concrete problem of estimating $g(m(\pi))) = 1/m(\pi)$ where π is a probability measure on some state space Ω. The problem can be naturally avoided if $S_H(m)/m \neq 0$ almost surely, which is often the case for continuous state-space Ω. However, the algorithm fails for discrete state spaces. Even if Ω only contain positive numbers that are far from 0, the resulting JOA estimator may still take 0 with positive probability, as it is constructed by the difference of two chains, see formula (2). The same problem only gets worse if the domain of g is of the form $\{x \mid \|x\| \geq c\}$, where both continuous and discrete Markov chains may fail.

To address this issue, we add an extra δ-transformation when necessary. Suppose $\mathcal{D} \supset \mathbb{R}^d \backslash B_\delta := \{x \mid \|x\| \geq \delta\}$ contains everything in $\mathbb{R}^d$ except for a compact set. Let H be the output of the unbiased MCMC subroutine $\mathcal{S}$, then the transformation $H \to H1_{H \geq \delta} + (H + 2\delta B)1_{H < \delta}$ outputs a random variable supporting on $\mathbb{R}^d \backslash B_\delta$ while maintaining the expectation, where B is uniform random variable on $\{-1, 1\}$ independent with H. The procedure is formally described below.

Algorithm 2 The δ-transformation

```
Input:
  • A subroutine $\mathcal{S}$ for generating unbiased estimators of $m(\pi)$.
  • A positive constant $\delta$.
Output: An unbiased estimator of $m(\pi)$ supporting on $\mathbb{R}^d \setminus B_\delta$.
Call $\mathcal{S}$ once and label the outputs by $H$.
If $H \geq \delta$:
    Return $H$.
Else:
    Flip a fair coin, Return $H + 2\delta$ if the coin comes up heads. Return $H - 2\delta$ otherwise.
```

The following proposition gives the validity proof of Algorithm 2.

Proposition 1. *Let $\tilde{H}$ be the output of Algorithm 2, then $\tilde{H} \geq \delta$ and $\mathbb{E}[\tilde{H}] = \mathbb{E}[H] = m(\pi)$.*

Proof. It suffices to show the unbiasedness of $\tilde{H}$. Notice that $\tilde{H} = H1_{H \geq \delta} + (H + 2\delta B)1_{H < \delta}$ where $B \sim \mathsf{U}\{-1, 1\}$ is independent with H. Therefore,

$$\mathbb{E}[\tilde{H}] = \mathbb{E}[H1_{H \geq \delta}] + \mathbb{E}[(H + 2\delta B)1_{H < \delta}] = \mathbb{E}[H1_{H \geq \delta}] + \mathbb{E}[H1_{H < \delta}] = \mathbb{E}[H].$$

□

The δ-transformation is used as a post-processing technique for the outputs of the unbiased MCMC algorithm. The whole algorithm for $g(m(\mu))$ is described in Algorithm 3 below. Algorithm 3 is the same as Algorithm 1 execpt for the post-processing procedure (Algorithm 2).

Although the δ-transformation trick allows one to work with functions not defined on the entire space, the assumption $\mathcal{D} \supset \mathbb{R}^d \setminus B_\delta$ still excludes many important functions. One typical example is $g(x) = \log(x)$ which is defined on $(0, \infty)$. In order to apply Algorithm 1, it is necessary to design a non-negative unbiased estimator of $m(\pi)$, which is in general quite difficult, and sometimes impossible as discussed in Jacob and Thiery [2015]. For example, the JOA estimator cannot be directly applied even if Ω contains only positive numbers – as it is the difference between two Markov chains. The domain constraint of g can be viewed as a limitation of our method, and we hope to report further progresses on relaxing this constaint in future works.

Algorithm 3 Unbiased Multilevel Monte-Carlo estimator after δ-transformation

```
Input:
  • A subroutine $\mathcal{S}$ for generating unbiased estimators of $m(\pi)$.
  • A function $g : \mathbb{R}^d \to \mathbb{R}$.
  • The parameter $p$ for geometric distribution.
Output: Unbiased estimator of $g(m(\pi))$.
  1. Sample $N$ from the geometric distribution $\mathsf{Geo}(p)$
  2. Call $\mathcal{S}$ for $2^N$ times and label the outputs by $H_1, \ldots, H_{2^N}$
  3. Call Algorithm 2 for each $H_i$ and label the outputs by $\tilde{H}_1, \ldots, \tilde{H}_{2^N}$
  4. Calculate the quantities $S_{\tilde{H}}(2^N)$, $S^{\mathsf{O}}_{\tilde{H}}(2^{N-1})$ and $S^{\mathsf{E}}_{\tilde{H}}(2^{N-1})$ by (3),(4)
  5. Calculate $\Delta_N = g\left(S_{\tilde{H}}(2^N)/2^N\right) - \frac{1}{2}\left(g\left(S^{\mathsf{O}}_{\tilde{H}}(2^{N-1})/2^{N-1}\right) + g\left(S^{\mathsf{E}}_{\tilde{H}}(2^{N-1})/2^{N-1}\right)\right)$
     Return: $W = \Delta_N/p_N + g(\tilde{H}_1)$.
```

3.3 Theoretical results

With all the notations as above, we are ready to state our technical assumptions and prove the theoretical results. Our theoretical analysis will focus on the unbiased estimator described in Algorithm 1. All the results still go through if the δ-transformation is needed. Recall that g is a function from $\mathcal{D}$ to $\mathbb{R}$, and $H_1, H_2, \cdots$ are *i.i.d.* unbiased estimators of $m(\pi)$. Now we denote by $V_n \subset \mathbb{R}^d$ the range of $(H_1 + \cdots + H_n)/n$ for every n and $V := \cup_{n=1}^{\infty} V_n$. Our assumptions are posed on both g and H_i:

Assumption 3.1 (Domain)**.** *The function $g : \mathcal{D} \to \mathbb{R}$ satisfies $V \subset \mathcal{D}$. Moreover, $m(\pi)$ is in the interior of $\mathcal{D}$, i.e., $m(\pi) \in \mathcal{D}^{\circ}$.*

Assumption 3.2 (Consistency)**.** $\mathbb{E}[g(\frac{S_H(n)}{n})] \to g(m(\pi))$ *as* $n \to \infty$.

Assumption 3.3 (Smoothness)**.** *The function g is continuously differentiable in a neighborhood of $m(\pi)$, and $Dg\,(\cdot)$ is locally Hölder continuous with exponent $\alpha > 0$. In other words, there exists $\varepsilon > 0$, $\alpha > 0$ and $c = c(\epsilon) > 0$ such that the following inequality holds for every $x, y \in (m(\pi) - \epsilon, m(\pi) + \epsilon)$:*

$$\|Dg(x) - Dg(y)\| \leq c\|x - y\|^{\alpha}.$$

Assumption 3.4 (Moment)**.** *There exists some $l > 2 + \alpha$ such that H has finite l-th moments, i.e.,*

$$\mathbb{E}[\|H_1\|_l^l] = \sum_{i=1}^{m} \mathbb{E}[H_{1,i}^l] < \infty.$$

Assumption 3.5 (Smoothness–Moment Tradeoff)**.** *There exist constants $s > 1$, $\alpha_s \in \mathbb{R}$, and $\mathcal{C}_s > 0$ such that $2\alpha_s + (s-1)l > 2s$ and $\mathbb{E}(|\Delta_n|^{2s}) \leq \mathcal{C}_s 2^{-\alpha_s n}$ for every $n \geq 0$*[2]*, where*

$$\Delta_n = g\left(S_H(2^n)/2^n\right) - \frac{1}{2}\left(g\left(S^{\mathsf{O}}_H(2^{n-1})/2^{n-1}\right) + g\left(S^{\mathsf{E}}_H(2^{n-1})/2^{n-1}\right)\right).$$

Now we briefly comment on the Assumptions 3.1 – 3.5. The descriptions below are mostly pedagogical, and the detailed proofs are deferred to the Appendix (Section 6).

[2] When $n = 0$, we define $\Delta_0 := g(H_1)$.

The Domain Assumption 3.1 guarantees Algorithm 1 can be implemented. When g does not directly satisfy this assumption, but $\mathcal{D} \supset \mathbb{R}^d \setminus B_\delta$, then we apply the δ-transformation (Algorithm 2 and 3) to enforce the first half of Assumption 3.1 holds. All the theoretical results still hold as long as the second half of Assumption 3.1 holds.

The consistency Assumption 3.2 is also expected and somewhat necessary. It appears in related works Vihola [2018], Blanchet and Glynn [2015], Zhou et al. [2021] explicitly or implicitly. The Law of Large Numbers guarantees $S_H(n)/n \to m(\pi)$ almost surely, therefore $g(S_H(n)/n) \to g(m(\pi))$ almost surely due to the continuity of g. Assumption 3.2 requires $\mathbb{E}[g(S_H(n)/n)] \to \mathbb{E}[g(m(\pi))]$, which is generally satisfied using the dominated convergence theorem.

The Smoothness Assumption 3.3 guarantees both g is smooth enough at a neighborhood of $m(\pi)$, and the derivative of g is Hölder continuous. When g is infinitely differentiable and there is no singularity on a neighborhood of μ, then we expect Assumption 3.3 to hold with $\alpha \geq 1$.

The Moment Assumption 3.4 requires more than $(2+\alpha)$-th moment of the unbiased estimator H_i. When the JOA estimator is used for generating H_i, Assumption 3.4 generally holds when π has more than l-th moment and the coupling time τ has a very light tail. We recall that a π-stationary Markov chain with transition kernel P is said to be geometrically ergodic if there is a $\gamma \in (0,1)$ and a function $C : \Omega \to (0,\infty)$ such that

$$\|P^n(x,\cdot) - \pi\|_{\mathsf{TV}} \leq C(x)\gamma^n,$$

for π–a.s. x. Our result guaranteeing Assumption 3.4 is the following.

Proposition 2 (Verifying Assumption 3.4, informal)**.** *Suppose the Markov chain P is π-stationary and geometrically ergodic, and f is a measurable function with finite p-th moment under π for any $p > l$. Suppose also there exists a set $\mathcal{S} \subset \Omega$, a constant $\tilde{\epsilon} \in (0,1)$ such that*

$$\inf_{(x,y)\in\mathcal{S}\times\mathcal{S}} \bar{P}((x,y),\mathcal{D}) \geq \tilde{\epsilon},$$

where $\mathcal{D} := \{(x,x) : x \in \Omega\}$ is the diagonal of $\Omega \times \Omega$. Then the JOA estimator $H_k(Y,Z) := f(Y_k) + \sum_{i=k+1}^{\tau-1}(f(Y_i) - f(Z_{i-1}))$ has a finite l-th moment, and therefore satisfies Assumption 3.4.

The formal description of the above proposition and the detailed proofs will be deferred to Appendix 6.3. Proposition 2 shows the existence of existence of the l-th moment of the JOA estimator for $l > 2$. The proof uses very similar techniques as Jacob et al. [2020]. It can be viewed as a slightly stronger version of Proposition 3.1 in Jacob et al. [2020], where the authors establishe the finite second order moment of the JOA estimator.

The Tradeoff Assumption 3.5 bounds the magnitude of $\mathbb{E}(\|\Delta_n\|^{2s})$. The condition $2\alpha_s+(s-1)l > 2s$ reflects the tradeoff between the smoothness of g and the moment assumption on H_i. Consider the following scenarios:

- Suppose g is at least twice continuously differentiable, and the derivative Dg is Lipschitz continuous. Then we have $\Delta_n = \mathcal{O}((S_H(2^n)/2^n)^2)$ by Taylor expansion. Meanwhile, the Central Limit Theorem (CLT) gives us $S_H(2^n)/2^n$, $S_H^{\mathsf{O}}(2^{n-1})/2^{n-1}$, and $S_H^{\mathsf{E}}(2^{n-1})/2^{n-1}$ are all of the magnitude $\mathcal{O}_p(2^{-n/2})$, which in turn shows $\Delta_n = \mathcal{O}_p(2^{-n})$.

Therefore we expect to choose $\alpha_s = s$ and therefore Assumption 3.5 is true for positive l. In this case Assumption 3.5 is weaker than Assumption 3.4.

- Suppose g is at most of linear growth, i.e., $|g(x)| \leq c(1 + \|x\|)$. When g does not have higher-order derivatives but still have some global control on the growth rate, we can not directly use Taylor expansion to cancel out the linear terms. In this case we can only bound Δ_n by $\mathcal{O}(S_H(2^n)/2^n)$, which is $\mathcal{O}_p(2^{-n/2})$ again by the CLT. We expect to choose $\alpha_s = s/2$ and it thus requires $l > s/(s-1)$. This is also the assumption in Blanchet and Glynn [2015], Blanchet et al. [2019].
- Suppose there is no special smoothness assumption on g, but $\mathbb{E}[\|\Delta_n\|^{2s}]$ is uniformly bounded. Then we expect to choose $\alpha_s = 0$, and therefore $l > 2s/(s-1)$.

As we can see from the above discussions, stronger smoothness requirements on g result in weaker assumptions on the moment of H_i, and vice versa. Theoretically, Assumption 3.5 even holds when $\alpha_s < 0$, provided that l is large enough.

Our main theoretical result is as follows.

Theorem 1. *Under Assumption 3.1 – 3.5, let* $\gamma := \min\{\alpha, \frac{\alpha_s}{s} + \frac{(s-1)l}{2s} - 1\} > 0$. *if* $N \in \{1, 2, \ldots\}$ *is geometrically distributed with success parameter* $p \in \left(\frac{1}{2}, 1 - \frac{1}{2^{(1+\gamma)}}\right)$, *then the estimator* $W := \Delta_N/p_N + g(H_1)$ *described in Algorithm 1 satisfies:*

1. $\mathbb{E}[W] = g(m(\pi))$,
2. *There exists a constant* C *such that*
$$\mathsf{Var}(W) \leq \mathbb{E}[W^2] \leq Cp^{-1} \frac{2^{-(1+\gamma)}}{1 - \left((1-p)2^{1+\gamma}\right)^{-1}} < \infty.$$
3. *The expected computational cost of Algorithm 1 is finite.*

The proof of Theorem 1 relies on the following key lemma to upper bound the second moment of Δ_n. The idea of bounding $\mathbb{E}[|\Delta_n|^2]$ appears in many relevant literatures such as Blanchet and Glynn [2015], Blanchet et al. [2019], Vihola [2018], Rhee and Glynn [2015] under different technical assumptions. Our tradeoff assumption 3.5 seems to be novel to our best knowledge.

Lemma 1. *We have*
$$\mathbb{E}[|\Delta_n|^2] = C2^{-(1+\gamma)n},$$
where
$$\gamma = \{\alpha, \frac{\alpha_s}{s} + \frac{(s-1)l}{2s} - 1\} > 0,$$
and $C = C(m, l, \epsilon, s, \alpha)$ *is a constant provided that Assumption 3.1 – 3.5 are satisfied.*

Though the detailed proof will be deferred to the Appendix 6.2, we sketch the proof idea here to provide some insights. For now, we temporarily assume g is smooth and has bounded second order derivative. Since
$$\Delta_n = g\left(S_H(2^n)/2^n\right) - \frac{1}{2}\left(g\left(S_H^{\mathsf{O}}(2^{n-1})/2^{n-1}\right) + g\left(S_H^{\mathsf{E}}(2^{n-1})/2^{n-1}\right)\right), \tag{5}$$
each term in the above summation is $\mathcal{O}_p(2^{-n/2})$ by the Central Limit Theorem. Therefore, using the triangle inequality will give an $\mathcal{O}(2^{-n/2})$ upper bound for $\mathbb{E}[|\Delta_n|^2]$, which is

strictly weaker than Lemma 1. The key observation is the antithetic design of Δ_n reduces the variance by a factor of $2^{-\Omega(1)n}$. More precisely, recall that

$$\frac{S_H(2^n)}{2^n} = \frac{1}{2}\Big(\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} + \frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\Big).$$

By Taylor expansion we have:

$$g(a) = g(\frac{a+b}{2}) + g'\Big(\frac{a+b}{2}\Big)\Big(\frac{a-b}{2}\Big) + \mathcal{O}((a-b)^2),$$

and

$$g(b) = g(\frac{a+b}{2}) + g'\Big(\frac{a+b}{2}\Big)\Big(\frac{b-a}{2}\Big) + \mathcal{O}((a-b)^2).$$

Therefore,

$$g\Big(\frac{a+b}{2}\Big) - \frac{1}{2}\Big(g(a) + g(b)\Big) = \mathcal{O}((a-b)^2),$$

so the anthithesic difference cancels out the constant and linear terms in the Taylor expansion, leaving the second order term as the dominating term. Taking $a = \frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}$ and $b = \frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}$, we know $\Delta_n = \mathcal{O}((\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}})^2) = \mathcal{O}_p(2^{-n})$. Therefore $\mathbb{E}[\Delta_n^2] = \mathcal{O}(2^{-2n})$, corresponding to the case $\gamma = 1$ in Lemma 1.

This gives the intuition of the proof idea. Our real technical assumptions (3.3) are more general than the idealized assumption above. It only requires α-Hölder continuity of Dg on a neighborhood of $m(\pi)$. Therefore, we discuss the behavior of $|\Delta_n|^2$ when $\frac{S_H(2^n)}{2^n}$ is closed to, or far away from its expectation. In both cases, we show the expected value of $|\Delta_n|^2$ is $\mathcal{O}(2^{-(1+\Omega(1))n})$. The details of the proof can be found in Appendix 6.2.

With Lemma 1 in hand, we are ready to show Theorem 1.

Proof of Theorem 1. We will first show 1 assuming 2 holds. Then we show 2 and 3 both holds. The proof of 2 depends heavily on Lemma 1.

Proof of Statement 1: Suppose W has a finite second moment, then the conditional expectation $\mathbb{E}[W \mid N]$ is well defined (see Section 4.1 of Durrett [2019] for details). The law of iterated expectation yields

$$\mathbb{E}[W] = \mathbb{E}\Big[\mathbb{E}[W \mid N]\Big] = \mathbb{E}[g(H_1)] + \mathbb{E}\Big[\frac{\mathbb{E}[\Delta_n \mid N]}{p_N}\Big] = \mathbb{E}[g(H_1)] + \mathbb{E}\Big[\frac{d_N}{p_N}\Big],$$

where $d_n = \mathbb{E}[g(S_H(2^n)/2^n)] - \mathbb{E}[g(S_H(2^{n-1})/2^{n-1})]$. We can further calculate $\mathbb{E}\Big[\frac{d_N}{p_N}\Big]$:

$$\mathbb{E}\Big[\frac{d_N}{p_N}\Big] = \sum_{i=1}^{\infty} \frac{d_i}{p_i} p_i = \sum_{i=1}^{\infty} d_i.$$

Therefore

$$\mathbb{E}[W] = \lim_{n\to\infty} \mathbb{E}[g(S_H(2^n)/2^n)] = g(m(\pi)),$$

as desired. The last equality uses Assumption 3.2.

Proof of Statement 2: Now we show $\mathbb{E}[W^2] < \infty$. We can directly calculate:

$$\mathbb{E}[W^2] \le 2\Big(\mathbb{E}[g(H_1)^2] + \mathbb{E}\Big[\frac{\Delta_N^2}{p_N^2}\Big]\Big).$$

It suffices to show $\mathbb{E}\Big[\frac{\Delta_N^2}{p_N^2}\Big] < \infty$. We have

$$\mathbb{E}\Big[\frac{\Delta_N^2}{p_N^2}\Big] = \sum_{n=1}^{\infty} \frac{\mathbb{E}[\Delta_n^2]}{p_n} = \sum_{n=1}^{\infty} \mathbb{E}[\Delta_n^2](1-p)^{-n+1}p^{-1}.$$

By Lemma 1,

$$\begin{aligned}\mathbb{E}\Big[\frac{\Delta_N^2}{p_N^2}\Big] \le Cp^{-1}(1-p)\sum_{n=1}^{\infty} 2^{-(1+\gamma)n}(1-p)^{-n} &= Cp^{-1}(1-p)\sum_{n=1}^{\infty}\Big((1-p)2^{1+\gamma}\Big)^{-n} \\ &= Cp^{-1}\frac{2^{-(1+\gamma)}}{1-\Big((1-p)2^{1+\gamma}\Big)^{-1}} < \infty,\end{aligned}$$

where the last inequality follows from $(1-p) > 2^{-(\gamma+1)}$.

Proof of Statement 3: We are now in the position of bounding the computation cost of Algorithm 1. Let C_H be the computation cost for implementing the unbiased MCMC subroutine $\mathcal{S}$ once. It is shown in Jacob et al. [2020] that $C_H < \infty$ given [GW: Add a little on Markov chain]. The computation cost for implementing Algorithm 1 essentially comes from 2^N calls of the subroutine $\mathcal{S}$, where $N \sim \mathsf{Geo}(p)$. Therefore it suffices to show 2^N has a finite expectation. We calculate

$$\mathbb{E}[2^N] = \sum_{n=1}^{\infty} 2^n p(n) = 2p\sum_{n=1}^{\infty}\Big(2(1-p)\Big)^{n-1} = \frac{2p}{2p-1} < \infty,$$

where the last inequality follows from $p > \frac{1}{2}$. □

The proof of Theorem 1 also suggests the following strategy to choose the success parameter p. One can define the work-normalized variance ... minimize the [GW: normalized variance]

4 Numerical examples

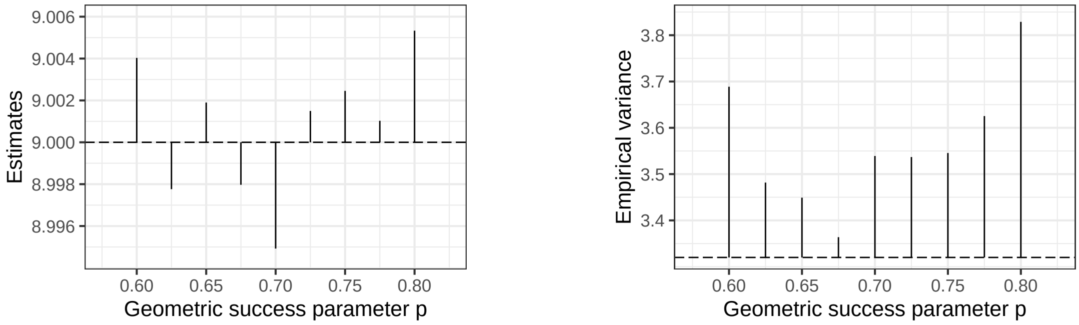
In this section, we investigate the empirical performance of the proposed method with [TW: several?] examples . We first inplement the algorithm on a multivariate distribution to show the corretness of our estimator and study the algorithm's sensitivity to the success parameter p in geometric distribution. [GW: have we studied the sensitivity of p in the first or second example?] Then we present the effect of meeting time in JOA estimator on 2D Ising model with periodic boundaries, and show numerical results of two statistics of interest.

4.1 Multivarite distribution

We begin with a toy model to illustrate the performance of our method. Let the multivariate random variable be $X = (X_1, \cdots, X_d) \in \mathbb{R}^d$ with $X_i \sim \mathsf{Beta}(i, 1)$ independently, and we want to estimate

$$g_K\left(\mathbb{E}(X_1), \cdots, \mathbb{E}(X_d)\right) = \prod_{i=1}^{K} \frac{1}{\mathbb{E}(X_i)} = \prod_{i=1}^{K} \frac{i+1}{i} = K + 1,$$

where $K = 1, \cdots, d$. First, for $K = 8$ and same settings in JOA estimator, we generate 10^5 unbiased estimates of $g_K(\cdot)$ using our algorithm with different geometric success parameter p. As is shown in Figure 2, the estimates vary little for different success parameters. For values in the range of $p \in (0.6, 0.8)$, results tand to be very similar. Thus, in order to ensure both small mean square error and efficient computation, we set $p = 0.7$ in the following examples. Then, for different K, the results are illustrated in Figure 3, where estimates and their standard deviation [GW: just to make sure, standard deviation is for our final estimator, or JOA estimator?] present a very similar trend. The reason behind this phenomenon is that when all parameters are fixed, the standard deviation is only effected by magnitude of the estimator. [GW: This I do not totally understand]

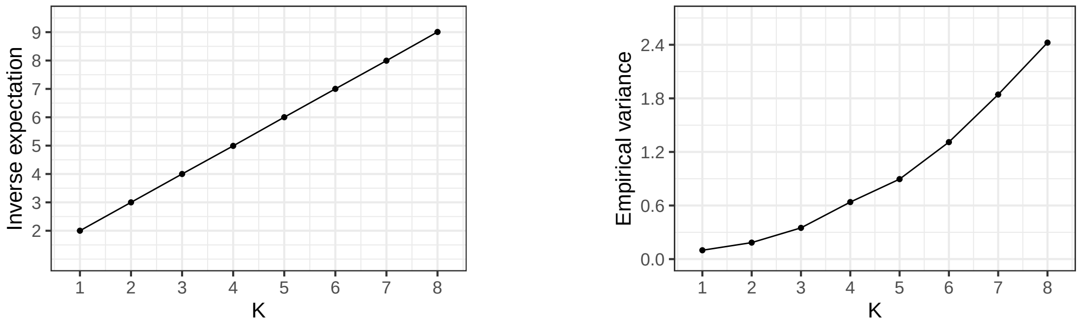


(a) Inverse expectation

(b) Empirical variance

Figure 2: Results for $K = 1$ with different geometric success parameters.

[GW: For the left plot, can we add a black dashed line to present the ground truch (though it seems our estimator looks very accurate, so the two lines should be nearly identical.)]



(a) Product of inverse expectation

(b) Empirical variance

Figure 3: $g_K(\cdot)$ in multivariate distribution.

4.2 Ising model

Here we apply the method to a $n \times n$ square lattice Ising model. Let $\sigma \in \{-1, 1\}^{n^2}$ be the spin configuration such that each individual spin $\sigma_i \in \{-1, +1\}$. Then the Hamiltonian of the Ising model is $H(\sigma) = -\sum_{i,j} \sigma_i \sigma_j$, [GW: If we sum over neighbors, then we should try to avoid using $\sum_{i,j}$, people seem to use $\sum_{\langle i,j \rangle}$, but we can check on Wiki/other notes]where the sum is over all pairs of neighboring lattice sites. Then given $\theta \geq 0$ the inverse temperature, the configuration probability is

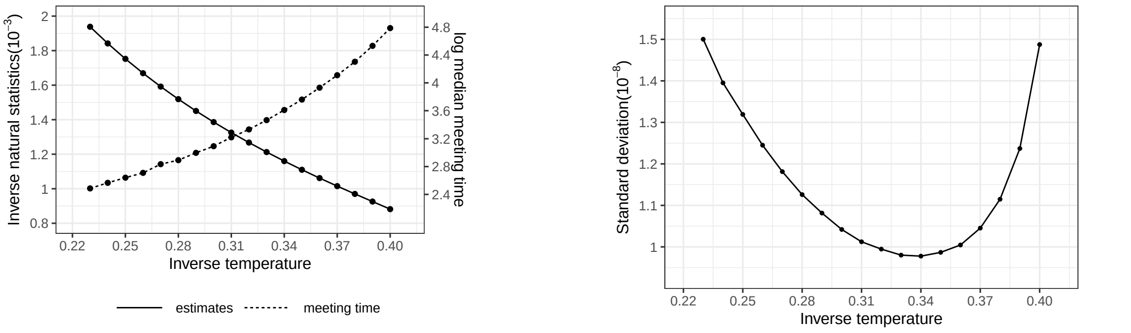
$$p_\theta(\sigma) = \frac{\exp(-\theta H(\sigma))}{Z(\theta)},$$

where $Z(\theta) = \sum_\sigma \exp(-\theta H(\sigma))$ is the normalization constant.

4.2.1 Inverse of natural statistics

In the Ising model, denote $h(\sigma) = -H(\sigma)$, then $\sum_\sigma p_\theta(\sigma) h(\sigma)$ the denominator is defined as "natural statistics" in Jacob et al. [2020]. Here we consider the inverse expected natural statistics $g(h(\sigma)) = 1/\sum_\sigma p_\theta(\sigma) h(\sigma)$, which should exhibit degeneracy as θ increases. Set $n = 32$ and $p = 0.7$ [TW: need another graph put above to show when p=0.7, the estimator has low variance and computing is efficient], we illustrated the reason for variance change in Figure 4. [GW: Can we change the label of the left plot to the unit of 10^{-3}? It looks slightly better than a lot of zeros. Can we change the right one to 10^{-8} instead of e^{-8}? Ideally we mention in the y-label that we use (10^{-3} or 10^{-8}), instead of repeating it several times.]

[GW: What p have we used?]



(a) $1/\left(\sum_\sigma p_\theta(x) h(x)\right)$ and log meeting time (b) Standard deviation of $1/\left(\sum_\sigma p_\theta(x) h(x)\right)$

Figure 4: estimates, meeting time and standard deviations of $1/\left(\sum_\sigma p_\theta(x) h(x)\right)$ at difference θ

4.2.2 Ratio of normalizing constant

Use the same notation as previous section, now we consider the ratio of normlization constant $Z(\theta_1)/Z(\theta_2)$. JOA estimator enabled us to sample

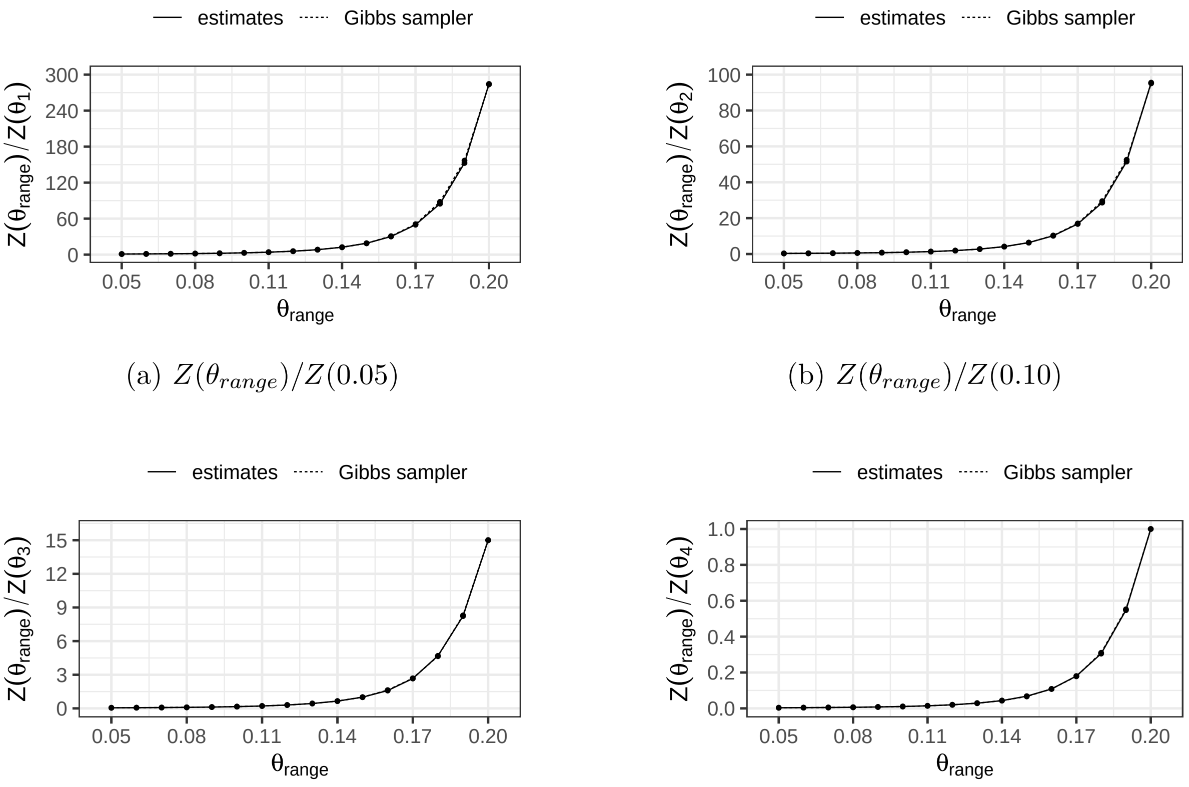
$$\sum_\sigma p_\theta(\sigma) \cdot \frac{1}{\exp(\theta h(\sigma))} = \sum_\sigma \frac{\exp(\theta h(\sigma))}{Z(\theta)} \cdot \frac{1}{\exp(\theta h(\sigma))} = \sum_\sigma \frac{1}{Z(\theta)} = \frac{2^{n^2}}{Z(\theta)}$$

Thus, with same size of the square lattice, the ratio can be generated by

$$\begin{aligned} \frac{Z(\theta_1)}{Z(\theta_2)} &= \sum_\sigma \frac{p_{\theta_2}(\sigma)}{\exp(\theta_2 h(\sigma))} / \sum_\sigma \frac{p_{\theta_1}(\sigma)}{\exp(\theta_1 h(\sigma))} \\ &= g\left(\sum_\sigma \frac{p_{\theta_1}(\sigma)}{\exp(\theta_1 h(\sigma))}, \sum_\sigma \frac{p_{\theta_2}(\sigma)}{\exp(\theta_2 h(\sigma))}\right) \end{aligned}$$

In practice, due to the preformance of coupling method, it would be expensive to generate unbiased estimators for either size n greater than 16 or inverse temperature θ larger than 0.30. Here for $n = 12$, we present results of the normalizing constant ratio in Figure 5. The solid lines represent our estimates for $Z(\theta_{range})/Z(\theta_i)$ such that $\theta_{range} = \{0.05, 0.07, \dots, 0.20\}$ and θ_i denotes the i^{th} value in $\{0.05, 0.10, 0.15, 0.20\}$. To check the results, we run Gibbs sampling 500 times to sample each $Z(\theta_{range})$ with 10^5 iterations, calculated $Z(\theta_{range})/Z(\theta_i)$. Then we get the averages and illustrated them by dash lines.

5 Conclusion



(a) $Z(\theta_{range})/Z(0.05)$

(b) $Z(\theta_{range})/Z(0.10)$

(c) $Z(\theta_{range})/Z(0.15)$

(d) $Z(\theta_{range})/Z(0.20)$

Figure 5: Ratio of normalizing constant $Z(\theta_{range})/Z(\theta_i)$ for $n = 12$. Solid lines represents our estimates and dash lines are estimates generated by Gibbs sampling

6 Appendix

6.1 Auxiliary Lemmas

In this section we prove some auxiliary results that will be used throughout the technical proofs. We start (without proof) the well-known Marcinkiewicz-Zygmund inequality, and then prove two useful corollaries based on this inequality.

Lemma 2 (Marcinkiewicz-Zygmund inequality [Marcinkiewicz and Zygmund, 1937]). *If $X_1, \cdots, X_n$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^p] < \infty$ for some $p > 2$. Then,*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^p\right] \leq C_p \mathbb{E}\left[\left(\sum_{i=1}^{n} |X_i|^2\right)^{p/2}\right],$$

where C_p is a constant that only depends on p.

One corollary of the Marcinkiewicz-Zygmund inequality is:

Corollary 1. *With all the assumptions as above, if we further assume that $X_1, \cdots, X_n$ are i.i.d. . Then,*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|^p\right] \leq C_p \frac{\mathbb{E}|X_1|^p}{n^{p/2}}$$

for every $p \geq 2$.

Proof of Corollary 1. Applying the Marcinkiewicz-Zygmund inequality on $(X_1/n, X_2/n, \ldots, X_n/n)$, we have:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|^p\right] \leq C_p \mathbb{E}\left[\left(\sum_{i=1}^{n}\left|\frac{X_i}{n}\right|^2\right)^{p/2}\right].$$

Since $x^{p/2}$ is convex, we have

$$\left(\sum_{i=1}^{n}\left|\frac{X_i}{n}\right|^2\right)^{p/2} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{|X_i|^2}{n}\right)^{p/2} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{|X_i|^p}{n^{p/2}}.$$

Taking expectation on both sides of the above inequality yields

$$\mathbb{E}\left[\left(\sum_{i=1}^{n}\left|\frac{X_i}{n}\right|^2\right)^{p/2}\right] \leq \frac{\mathbb{E}|X_1|^p}{n^{p/2}},$$

and our desired inequality follows. □

The Marcinkiewicz-Zygmund inequality naturally generalizes to random vectors.

Corollary 2 (Multivariate Marcinkiewicz-Zygmund inequality)**.** *Let $X_1, \cdots, X_n$ be i.i.d. random vectors in $\mathbb{R}^m$, with $\mathbb{E}[X_i] = \mathbf{0}$ and $\mathbb{E}[\|X_i\|_p^p] = \mathbb{E}[\sum_{j=1}^m |X_{i,j}|^p] < \infty$. Then*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_p^p\right] \leq C_p \frac{\mathbb{E}\left[\|X_1\|_p^p\right]}{n^{p/2}}$$

for every $p \geq 2$.

Proof of Corollary 2. We know

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_p^p\right] = \sum_{j=1}^{m}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right|^p\right].$$

Applying Corollary 1 on each component of each X_i yields

$$\sum_{j=1}^{m}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right|^p\right] \leq C_p \sum_{j=1}^{m}\frac{\mathbb{E}|X_{1,j}|^p}{n^{p/2}} = C_p \frac{\mathbb{E}\left[\|X_1\|_p^p\right]}{n^{p/2}},$$

as desired. □

We also need the following inequality to compare $\|x\|_p$ and $\|x\|_q$ for $p \neq q$ and $x \in \mathbb{R}^m$. The proof follows directly from the Hölder's inequality.

Lemma 3. *For any $x \in \mathbb{R}^m$ and $p < q$, we have:*

$$\|x\|_p \leq m^{1/p-1/q}\|x\|_q.$$

Proof.

$$\|x\|_p^p = \sum_{i=1}^{m}|x_i|^p \cdot 1 \leq \left(\sum_{i=1}^{m}|x_i|^q\right)^{p/q} m^{1-p/q}$$

where the last inequality follows from the Hölder's inequality. Our result follows by taking the $(1/p)$-th power on both sides. □

6.2 Bounding $\mathbb{E}[|\Delta_n|^2]$

Recall that $\Delta_n = g\left(S_H(2^n)/2^n\right) - \frac{1}{2}\left(g\left(S_H^O(2^{n-1})/2^{n-1}\right) + g\left(S_H^E(2^{n-1})/2^{n-1}\right)\right)$, and the final estimator takes the form $\Delta_N/p_N + g(H_1)$. Therefore, understanding the theoretical properties of Δ_n is crucial for studying our estimator.

Proof of Lemma 1. For simplicity, we denote $m(\pi)$ by μ. By Assumption 3.3, there exists $\epsilon > 0$ such that g is α-Hölder continuous on $(\mu - \epsilon, \mu + \epsilon)$, we can then write Δ_n as:

$$|\Delta_n| = |\Delta_n|\mathbf{1}(A_1) + |\Delta_n|\mathbf{1}(A_2), \tag{6}$$

where A_1 is the event

$$\left\{\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\| < \epsilon\right\} \cap \left\{\left\|\frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}} - \mu\right\| < \epsilon\right\},$$

and A_2 is the event

$$\left\{\max\left(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\|, \left\|\frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}} - \mu\right\|\right) \geq \epsilon\right\}.$$

Under the event A_1, we have $\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\| < \epsilon$ and $\left\|\frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}} - \mu\right\| < \epsilon$. This further implies

$$\left\|\frac{S_H(2^n)}{2^n} - \mu\right\| < \epsilon$$

by the triangle inequality and the fact $\frac{S_H(2^n)}{2^n} = \frac{1}{2}\left(\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} + \frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right)$.

Then we can write Δ_n as:

$$\begin{aligned}
\Delta_n &= g\left(\frac{S_H(2^n)}{2^n}\right) - \frac{1}{2}\left(\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} + \frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right) \\
&= \frac{1}{2}\left(g\left(\frac{S_H(2^n)}{2^n}\right) - g\left(\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}\right)\right) + \frac{1}{2}\left(g\left(\frac{S_H(2^n)}{2^n}\right) - g\left(\frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right)\right) \\
&= \frac{1}{2}Dg(\xi_n^{\mathsf{O}})\left(\frac{S_H(2^n)}{2^n} - \frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}\right) + \frac{1}{2}Dg(\xi_n^{\mathsf{E}})\left(\frac{S_H(2^n)}{2^n} - \frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right) \\
&= \frac{1}{4}\left(Dg(\xi_n^{\mathsf{O}}) - Dg(\xi_n^{\mathsf{E}})\right)\frac{S_H^{\mathsf{E}}(2^{n-1}) - S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}},
\end{aligned}$$

where ξ_n^{O} is a convex combination of $\frac{S_H(2^n)}{2^n}$ and $\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}$, ξ_n^{E} is a convex combination of $\frac{S_H(2^n)}{2^n}$ and $\frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}$ by the Multivariate Mean value Theorem. Under A_1, both ξ_n^{O} and ξ_n^{E} are within the ϵ-neighbor of μ, applying the α-Hölder continuous assumption yields

$$|\Delta_n| \leq c_1(\epsilon)\left\|\xi_n^{\mathsf{O}} - \xi_n^{\mathsf{E}}\right\|^{\alpha} \cdot \left\|\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right\| \leq c_2(\epsilon)\left\|\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right\|^{1+\alpha}.$$

Then,

$$\mathbb{E}\left[|\Delta_n|^2\mathbf{1}(A_1)\right] \le c_2(\epsilon)\mathbb{E}\left[\left\|\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right\|^{2(1+\alpha)}\right]. \tag{7}$$

Since $S_H^{\mathsf{O}}(2^{n-1})$ and $S_H^{\mathsf{E}}(2^{n-1})$ are vectors in $\mathbb{R}^m$, applying Lemma 3 on $p = 2, q = 2(1+\alpha)$ gives:

$$\left\|\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right\|^{2(1+\alpha)} \le m^\alpha \left\|\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right\|_{2(1+\alpha)}^{2(1+\alpha)} \tag{8}$$

Since $S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})$ is the sum of 2^{n-1} i.i.d. random variables, each with the same distribution as $H_2 - H_1$, applying the Multivariate Marcinkiewicz-Zygmund inequality (Corollary 2) gives us:

$$\mathbb{E}\left[\left\|\frac{S_H^{\mathsf{O}}(2^{n-1}) - S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}}\right\|_{2(1+\alpha)}^{2(1+\alpha)}\right] \le C_{2(1+\alpha)} \cdot \frac{\mathbb{E}\left[\|H_2 - H_1\|_{2(1+\alpha)}^{2(1+\alpha)}\right]}{2^{(1+\alpha)(n-1)}} \tag{9}$$

$$\le C_{2(1+\alpha)} \cdot 2^{3(1+\alpha)} \cdot \frac{\mathbb{E}\left[\|H_1\|_{2(1+\alpha)}^{2(1+\alpha)}\right]}{2^{(1+\alpha)n}}. \tag{10}$$

where the last step uses the inequality $(a+b)^p \le 2^{p-1}(|a|^p + |b|^p)$ for $p \ge 2$. It is worth mentioning that the right hand side of (10) is finite as Assumption 3.4 guarantees H_1 has finite l-th moment with $l > 2 + \alpha$. Combining (7), (8), and (10), we have:

$$\mathbb{E}\left[\|\Delta_n\|^2\mathbf{1}(A_1)\right] \le C_1(m, \alpha, \epsilon)2^{-n(1+\alpha)}, \tag{11}$$

where $C_1(m, \alpha, \epsilon) = c_2(\epsilon) \cdot C_{2(1+\alpha)} \cdot 2^{3(1+\alpha)} \cdot \mathbb{E}\left[\|H_1\|_{2(1+\alpha)}^{2(1+\alpha)}\right]$ is a constant when Assumption 3.1 – 3.4 are satisfied.

Under A_2, we have:

$$|\Delta_n|^2\mathbf{1}(A_2) \le |\Delta_n|^2\mathbf{1}\left(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\| > \epsilon\right) + |\Delta_n|^2\mathbf{1}\left(\left\|\frac{S_H^{\mathsf{E}}(2^{n-1})}{2^{n-1}} - \mu\right\| > \epsilon\right) \tag{12}$$

Now we upper bound the first term's expectation,

$$\mathbb{E}\left[|\Delta_n|^2\mathbf{1}\Big(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\| > \epsilon\Big)\right] \le \mathbb{E}[|\Delta_n|^{2s}]^{1/s}\mathbb{P}\left(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\| > \epsilon\right)^{(s-1)/s} \tag{13}$$

$$\le \mathcal{C}_s^{1/s}2^{-\alpha_s n/s}\mathbb{P}\left(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\| > \epsilon\right)^{(s-1)/s} \tag{14}$$

$$\le \mathcal{C}_s^{1/s} \cdot (\epsilon^{-l(s-1)/s}) \cdot 2^{-\alpha_s n/s} \cdot \mathbb{E}\left[\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}} - \mu\right\|^l\right]^{(s-1)/s}. \tag{15}$$

Here (13) follows from the Hölder's inequality, (14) uses Assumption 3.5, and (15) follows from the Markov's inequality. Again, using Lemma 3 and Corollary 2, the term

$\mathbb{E}\left[\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}-\mu\right\|^l\right]$ can be upper bounded by:

$$\mathbb{E}\left[\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}-\mu\right\|^l\right] \leq m^{l/2-1}\mathbb{E}\left[\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}-\mu\right\|_l^l\right] \leq 2^{l/2}\cdot m^{l/2-1}\cdot C_l\cdot\frac{\mathbb{E}\left[\|H_1\|_l^l\right]}{2^{nl/2}}. \tag{16}$$

Combining (15) and (16), we have

$$\begin{aligned}\mathbb{E}\left[|\Delta_n|^2\mathbf{1}\Big(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}-\mu\right\|>\epsilon\Big)\right] &\leq C_2(m,l,\epsilon,s)2^{-\alpha_s n/s}2^{-nl(s-1)/(2s)}\\ &= C_2(m,l,\epsilon,s)2^{-n\left(\frac{\alpha_s}{s}+\frac{(s-1)l}{2s}\right)},\end{aligned}$$

where $C_2(m,l,\epsilon,s) = \mathcal{C}_s^{1/s}\cdot\left(\epsilon^{-l}2^{l/2}\cdot m^{l/2-1}\cdot C_l\cdot\mathbb{E}\left[\|H_1\|_l^l\right]\right)^{(s-1)/s}$ is a constant when Assumption 3.1 – 3.5 are satisfied. Furthermore, by Assumption 3.5, $2\alpha_s+(s-1)l>2s$. It is clear that $\frac{\alpha_s}{s}+\frac{(s-1)l}{2s}>1$, and therefore

$$\mathbb{E}\left[|\Delta_n|^2\mathbf{1}\Big(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}-\mu\right\|>\epsilon\Big)\right]\leq C_2(m,l,\epsilon,s)2^{-(1+\tilde{\alpha})n}, \tag{17}$$

where $\tilde{\alpha}=\frac{\alpha_s}{s}+\frac{(s-1)l}{2s}-1>0$. The same argument also shows

$$\mathbb{E}\left[|\Delta_n|^2\mathbf{1}\Big(\left\|\frac{S_H^{\mathsf{O}}(2^{n-1})}{2^{n-1}}-\mu\right\|>\epsilon\Big)\right]\leq C_2(m,l,\epsilon,s)2^{-(1+\tilde{\alpha})n}. \tag{18}$$

Combining (17), (18), and (12), we have

$$\mathbb{E}\left[|\Delta_n|^2\mathbf{1}(A_2)\right]\leq 2C_2(m,l,\epsilon,s)2^{-(1+\tilde{\alpha})n}. \tag{19}$$

Finally, taking $\gamma=\min\{\alpha,\tilde{\alpha}\}$, $C=C_1+2C_2$, and using (6), (11), and (19), we conclude:

$$\mathbb{E}[|\Delta_n|^2]\leq C2^{-n(1+\gamma)}. \tag{20}$$

□

6.3 The Moment Assumption 3.4 and Markov chain mixing

In this subsection we discuss the relation between the Moment Assumption 3.4 and the mixing time of the underlying Markov chain. Throughout this subsection, the unbiased estimator H of $m(\pi)$ is assumed to be the JOA estimator $H_k(Y,Z)$ defined in (2), which also extends to $H_{k:m}(Y,Z)=(m-k+1)^{-1}\sum_{l=k}^m H_l(Y,Z)$ naturally.

Before giving a formal statement of Proposition 2, we first recall some definitions in Markov chain theory. We say a π-invariant, ϕ-irreducible and aperiodic Markov transition kernel P satisfies a geometric drift condition if there exists a measurable function $V:\Omega\to[1,\infty)$, $\lambda\in(0,1)$, and a measurable set $\mathcal{S}$ such that for all $x\in\Omega$:

$$\int P(x,\mathrm{d}y)V(y)\leq\lambda V(x)+b\mathbf{1}(x\in\mathcal{S}). \tag{21}$$

Moreover, the set $\mathcal{S}$ is called a small set if there exists a positive integer m, $\epsilon > 0$, and a probability measure ν on such that for every $x \in \mathcal{S}$:

$$P^m(x, \cdot) \geq \epsilon\mu(\cdot). \tag{22}$$

The technical definitions for irreducibility, aperiodicity and small sets can be found in Chapter 5 of Meyn and Tweedie [2012]. The geometric drift condition is a key tool guaranteeing the geometric ergodicity of a Markov chain, meaning the Markov chain P converges to its stationary distribution π at a geometric rate (see Theorem 9 of Roberts and Rosenthal [2004], Jones and Hobert [2001]). It is known that the geometric drift condition is satisfied for a wide family of Metropolis-Hastings algorithms, we refer the readers to Mengersen and Tweedie [1996], Roberts and Tweedie [1996a,b], Jarner and Hansen [2000], and Roberts and Rosenthal [1997] for existing results.

Now we give a formal statement of Proposition 2.

Proposition 3 (Verifying Assumption 3.4, formal version of Proposition 2)**.** *Suppose the Markov transition kernel described in Section 3.1.1 satisfies a geometric drift condition with a small set $\mathcal{S}$ of the form $\mathcal{S} = \{x : V(x) \leq L\}$ for $\lambda + b/(1+L) < 1$. Suppose there exists $\tilde{\epsilon} \in (0,1)$ such that*

$$\inf_{(x,y)\in\mathcal{S}\times\mathcal{S}} \bar{P}((x,y),\mathcal{D}) \geq \tilde{\epsilon},$$

where $\mathcal{D} := \{(x,x) : x \in \Omega\}$ is the diagonal of $\Omega \times \Omega$. Suppose also there exists $p > l$ and $D_p > 0$ such that $\mathbb{E}[\|f(Y_t)\|_p^p] < D_p$ for every t. Then $\mathbb{E}[\|H_k(Y,Z)\|_l^l] < \infty$ for every k.

The main ingredient in the proof of Proposition 3 is to control the tail probability of the meeting time τ. We say τ has a β-polynomial tail if there exists a constant $K_\beta > 0$ such that

$$\mathbb{P}(\tau > n) \leq K_\beta n^{-\beta}. \tag{23}$$

We say τ has an exponential tail if there exists a constant $K > 0$ and $\gamma \in (0,1)$ such that

$$\mathbb{P}(\tau > n) \leq K\gamma^n. \tag{24}$$

Our next result gives sufficient conditions to ensure Assumption 3.4.

Lemma 4. *Suppose one of the following holds:*

- *There exists $p > l$, $\beta > 0$, and $D_p > 0$ such that $\frac{1}{p} + \beta > \frac{1}{l}$; $\mathbb{E}[\|f(Y_t)\|_p^p] < D_p$ for every t, and τ has a β-polynomial tail;*
- *There exists $p > l$ and $D_p > 0$ such that $\mathbb{E}[\|f(Y_t)\|_p^p] < D_p$ for every t, and τ has an exponential tail.*

Then $\mathbb{E}[\|H_k(Y,Z)\|_l^l] < \infty$ for every k.

Proof of Lemma 4. We start with the first case. Without loss of generality, we assume $k = 0$ and the estimator $H_0(Y,Z) := f(Y_0) + \sum_{i=1}^{\tau-1}(f(Y_i) - f(Z_{i-1}))$ takes scalar value. Let $D_k := f(Y_k) - f(Z_{k-1})$ for $k \geq 1$, and $D_0 = f(Y_0)$, the estimator can be written as:

$$H_0(Y,Z) = \sum_{k=0}^{\infty} D_k \mathbf{1}(\tau > k).$$

The meeting time τ is almost surely (a.s.) finite by the β-polynomial assumption, therefore $H_0(Y,Z)$ is the limit of $H_0^n(Y,Z) := \sum_{k=0}^n D_k \mathbf{1}(\tau > k)$ in the a.s. sense. We will now prove $H_0^n(Y,Z) \to H_0(Y,Z)$ in L^l , which further implies $\mathbb{E}[|H_0(Y,Z)|^l] < \infty$.

By the Minkowski's inequality on the probability space $L^l(\Omega)$, we have

$$\Big(\mathbb{E}[|H_0^n(Y,Z) - H_0(Y,Z)|^l]\Big)^{1/l} = \Big(\mathbb{E}\big[\Big|\sum_{k=n+1}^{\infty} D_k \mathbf{1}(\tau > k)\Big|^l\big]\Big)^{1/l} \tag{25}$$

$$\leq \sum_{k=n+1}^{\infty} \Big(\mathbb{E}[|D_k \mathbf{1}(\tau > k)|^l]\Big)^{1/l}. \tag{26}$$

Every term in (26) can be upper bounded by the Hölder's inequality

$$\Big(\mathbb{E}[|D_k \mathbf{1}(\tau > k)|^l]\Big)^{1/l} \leq \Big(\mathbb{E}[|D_k|^p]\Big)^{1/p}\Big(\mathbb{P}(\tau > k)\Big)^{1/q} \qquad \text{here } 1/q = 1/l - 1/p \tag{27}$$

$$\leq (2D_p)^{1/p} K_\beta^{1/q} k^{-\beta/q} = (2D_p)^{1/p} K_\beta^{1/q} k^{-\frac{\beta}{\frac{1}{l}-\frac{1}{p}}}. \tag{28}$$

Since $\beta > \frac{1}{l} - \frac{1}{p} > 0$, the right hand side of (28) is summable. Therefore we conclude $\sum_{k=n+1}^{\infty} \Big(\mathbb{E}[|D_k \mathbf{1}(\tau > k)|^l]\Big)^{1/l} \to 0$ as $n \to \infty$, so $H_0^n(Y,Z) \to H_0(Y,Z)$ in L^l.

In the second case, exponential light tail implies β-polynomial tail for every $\beta > 0$, our result immediately follows from the first case.

□

The assumption $\mathbb{E}[\|f(Y_t)\|^p] < D_p$ in Lemma 4 is generally satisfied as long as f has p-th moment under the stationary distribution π. It remains to verify the tail conditions of τ, i.e., formula (23) or (24). The exponential tail (24) and polynomial tail (23) are closely related to the geometric ergodicity and polynomial ergodicity of the underlying marginal Markov chain P, respectively. For simplicity, we only give conditions for the exponential tail here, which is provided in Jacob et al. [2020]. The sufficient conditions of polynomial tail of τ can be founded in Theorem 2 of Middleton et al. [2020].

Proposition 4 (Proposition 3.4 in Jacob et al. [2020])**.** *Suppose the Markov transition kernel described in Section 3.1.1 satisfies a geometric drift condition with a small set $\mathcal{S}$ of the form $\mathcal{S} = \{x : V(x) \leq L\}$ for $\lambda + b/(1+L) < 1$. Suppose there exists $\tilde{\epsilon} \in (0,1)$ such that*

$$\inf_{(x,y)\in\mathcal{S}\times\mathcal{S}} \bar{P}((x,y),\mathcal{D}) \geq \tilde{\epsilon},$$

where $\mathcal{D} := \{(x,x) : x \in \Omega\}$ is the diagonal of $\Omega \times \Omega$. Then the meeting time τ has a exponential light tail.

Combining Lemma 4 and Proposition 4, the proof of Proposition 3 is immediate.

Proof of Proposition 3. By Proposition 4, we know τ has an exponential tail. Using the second case of Lemma 4, our result follows. □

It is still possible to further strengthen Proposition 3 given extra assumptions on τ or f. For example, when τ has an exponential tail and $\mathbb{E}_\pi[e^{\theta f}] < \infty$ for a univerate f and some $\theta > 0$, one can then prove the JOA estimator also has an exponential moment, and thus has every finite-order moment. The existence of an exponential moment may be helpful for analyzing the concentration properties of the JOA estimator.

References

N. Biswas, P. E. Jacob, and P. Vanetti. Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 2, 5

J. H. Blanchet and P. W. Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. *2015 Winter Simulation Conference (WSC)*, pages 3656–3667, 2015. 2, 3, 6, 10, 11

J. H. Blanchet, N. Chen, and P. W. Glynn. Unbiased monte carlo computation of smooth functions of expectations via taylor expansions. In *2015 Winter Simulation Conference (WSC)*, pages 360–367. IEEE, 2015. 3, 7

J. H. Blanchet, P. W. Glynn, and Y. Pei. Unbiased Multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*, 2019. 2, 11

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. 12

P. Fearnhead, O. Papaspiliopoulos, G. O. Roberts, and A. Stuart. Random-weight particle filtering of continuous time processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):497–512, 2010. 7

M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008. 2

M. B. Giles. Multilevel Monte Carlo methods. *Acta Numer.*, 24:259–328, 2015. 2

P. W. Glynn and C.-h. Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014. 6

S. Heinrich. Multilevel Monte Carlo methods. In *International Conference on Large-Scale Scientific Computing*, pages 58–67. Springer, 2001. 2

J. Heng and P. E. Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302, 2019. 2

P. E. Jacob and A. H. Thiery. On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784, 2015. 3, 8

P. E. Jacob, J. O'Leary, and Y. F. Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020. 1, 5, 6, 10, 13, 15, 23

S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000. 22

G. L. Jones and J. P. Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001. 22

M. Keane and G. L. O'Brien. A bernoulli factory. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 4(2):213–219, 1994. 3

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-Scale Methods for Distributionally Robust Optimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020. 2

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015. 7

J. Marcinkiewicz and A. Zygmund. Quelques théoremes sur les fonctions indépendantes. *Fund. Math*, 29:60–90, 1937. 17

D. McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17(4):301–315, 2011. 2

X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996. 7

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996. 22

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012. 22

L. Middleton, G. Deligiannidis, A. Doucet, and P. E. Jacob. Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14 (2):2842–2891, 2020. 2, 5, 23

I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006. 7

Ş. Nacu and Y. Peres. Fast simulation of new coins from old. *The Annals of Applied Probability*, 15(1A):93–115, 2005. 3

O. Papaspiliopoulos. A methodological framework for Monte Carlo probabilistic inference for diffusion processes. 2009. 7

C.-h. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015. 2, 11

G. Roberts and J. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997. 22

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability surveys*, 1:20–71, 2004. 22

G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996a. 22

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996b. 22

J. S. Rosenthal. Faithful couplings of Markov chains: now equals forever. *Advances in Applied Mathematics*, 18(3):372–381, 1997. 4

F. J. Ruiz, M. K. Titsias, T. Cemgil, and A. Doucet. Unbiased gradient estimation for variational auto-encoders using coupled markov chains. *arXiv preprint arXiv:2010.01845*, 2020. 2

Y. Shi and R. Cornish. On Multilevel Monte Carlo Unbiased Gradient Estimation for Deep Latent Variable Models. In *International Conference on Artificial Intelligence and Statistics*, pages 3925–3933. PMLR, 2021. 2

M. Vihola. Unbiased estimators and multilevel Monte Carlo. *Operations Research*, 66(2): 448–462, 2018. 3, 10, 11

W. Wagner. Unbiased Monte Carlo evaluation of certain functional integrals. *Journal of Computational Physics*, 71(1):21–33, 1987. 7

G. Wang, J. O'Leary, and P. Jacob. Maximal Couplings of the Metropolis-Hastings Algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 1225–1233. PMLR, 2021. 2

Z. Zhou, G. Wang, J. Blanchet, and P. W. Glynn. Unbiased Optimal Stopping via the MUSE. *arXiv preprint arXiv:2106.02263*, 2021. 2, 10