Unbiased Derivative Estimation for Stationary Mean of Parameterized Markov chains

Jeffrey Wang, Chang-Han Rhee

November 11, 2024

Abstract

We propose a new approach to unbiased estimation of the gradients of the stationary means associated with parametrized families of Markov chains. Our estimators are particularly efficient when the Markov chains have slow mixing rate. Our approach does not require a specific parametrization except for an oracle to evaluate the transition density and its gradient at a given data point without any additional knowledge about the density function itself. It makes our estimator suitable for parametrizations associated with neural networks. The estimator can potentially achieve large improvement in terms of efficiency. Numerical experiments confirm the good performance predicted by the theory.

1 Introduction

Consider a Markov chain $X = \{X_i\}_{i\geq 0}$ on a general state space $\mathcal{X}$. Let $P(\theta) := (P(\theta, x, dy) : x, y \in \mathcal{X})$ denote its transition kernel parametrized by $\theta \in \Lambda$. We assume that there exists a unique invariant distribution $\pi(\theta)$ associated with $P(\theta)$. Now let $\gamma(\theta) := \mathbb{E}_{\pi(\theta)}[f(X)]$ denote the stationary mean for some integrable function $f : \mathcal{X} \to \mathbb{R}$ under $\pi(\theta)$. The goal of our work is to estimate the derivative of the stationary mean, denoted by

$$\gamma'(\theta) := \frac{d}{d\theta}\mathbb{E}_{\pi(\theta)}[f(X)]. \tag{1.1}$$

Such derivatives are essential for sensitivity analysis of the long-run average performance of a system or for optimization of the performance with respect to the parameter θ. However, the explicit evaluation of derivatives are rarely possible and one typically needs to estimate the derivatives computationally. When efficient and unbiased derivative estimators are available, numerical optimization algorithms such as stochastic gradient decent (SGD) converge more quickly. Throughout the paper, efficiency of an estimator is measured by the reciprocal of the product of the expected work and variance of the estimator; see [14].

Our approach builds on a few recent lines of research. The first line concerns the general theory [25] of *debiasing* simulation estimators. In [13], it was shown that this idea, together with a coupling technique, can be used to construct unbiased estimators for the stationary means of positive Harris recurrent chains. [17] further studied and exploits the coupling technique to bring it to the Markov chain Monte Carlo (MCMC) setting where they construct unbiased estimators for integrals with respect to a target distribution. Recently [6] also applied similar coupling idea to construct unbiased estimators for the solutions to the Poisson's equation associated with Markov chains, and from there they develop unbiased estimator for the asymptotic variance of the Markov chain. In this work, we address an estimation problem of yet another limiting quantity, the derivative of the stationary mean.

There have been a substantial body of works that aim to estimate such a derivative. Whenever an atom is available or can be easily constructed, we may utilize the regeneration structure of the Markov chain to construct a derivative estimator, see [10], for example. However, since the estimator is a non-linear function of sample averages, it is an asymptotically consistent estimator but biased for any finite realizations. More importantly, this approach can be hard to implement if regeneration states are hard to be identified or the constructed regeneration cycle is too long to be efficient. The Infinitesimal Perturbation Analysis (IPA) estimator for $\gamma'(\theta)$ has also been developed in [8]. It transforms the derivative estimation problem into a stationary mean estimation problem by constructing the *derivative process*. Although the estimator is unbiased and the efficiency typically triumphs other estimators when available, the method can be implemented on a very restricted class of parametrizations. [16],[15] provide another way to estimate $\gamma'(\theta)$ by applying the Hahn-Jordan decomposition to decompose the derivative of transition kernel $P'(\theta)$ into two positive kernels $P^{+}(\theta)$ and $P^{-}(\theta)$. They simulate two chains separately starting from those two kernels and produce an estimator, which they call Phantom estimator. Potentially, this could avoid requiring constructing atoms and be applied to a broader class of parametrizations but it still needs detailed knowledge about the transition kernel $P(\theta)$ to (i) perform the Hahn-Jordan decomposition, (ii) sample from the decomposed kernels and (iii) compute the normalizing constants for the decomposed parts. A more recent work [12] proposed two more likelihood ratio estimators for $\gamma'(\theta)$ that does not require constructing regenerative cycles. Moreover, this approach requires the knowledge of the the stationary mean $\gamma(\theta)$ which is often unavailable. Those likelihood ratio estimators are asymptotically consistent and so far have the fewest requirements to implement. However, as the bias gets smaller, the efficiency gets worse and a difficult trade-off has to be made here. Specific implementation details and numerical results of the aforementioned methods are deferred to Section 6.

Our contribution builds on the theoretical results from [24] and we propose an unbiased estimator for $\gamma'(\theta)$. The requirement to implement this estimator is modest. To be specific, it only requires an oracle which given a parameter θ and a state pair (x, y), outputs the one-step transition density $P(\theta, x, dy)$ and its derivative $P'(\theta, x, dy)$ with respect to θ. Our method does not require constructing regeneration time nor analytical knowledge of the transition kernel $P(\theta)$. This is desirable in modern optimization setting as often the kernel is parametrized by complicated function approximators such as neural networks (NNs), for which it is almost impossible to obtain analytic knowledge. Meanwhile, an oracle described above can be obtained if the transition kernel is parameterized by NNs. Specifically, given the current NN weights θ and a state pair (x, y), $P(\theta, x, dy)$ and $P'(\theta, x, dy)$ can be numerically computed using forward-feeding and back-propagation of the NNs respectively. Section 5 provides examples regarding Markov chains whose transition kernels are parametrized by NNs. In addition to the previously stated theoretical desirability, our estimator can be surprisingly efficient as well when tuned appropriately. Numerical experiments show that for slow mixing Markov chains which are the difficult scenarios, the efficiency of our derivative estimator achieves a similar level of performance with the IPA estimator and surpasses almost all previous estimators.

The rest of the paper is organized as follows: Section 2 contains the preliminary background for coupling technique and how it can be used to construct unbiased estimators for the stationary means and solutions to the Poisson's equation. Section 3 contains the development of our unbiased derivative estimators. Section 4 provides efficiency analysis of the previously developed estimator and gives a general guideline for choosing parameters based on the analysis results. Section 6 shows the empirical performance of our estimator comparing with the previously developed ones in a queueing example.

2 Preliminaries

2.1 Construction of coupled Markov chains

Assume we have a joint kernel $\bar{P}(\theta)$ that evolves on the joint space $\mathcal{X} \times \mathcal{X}$ such that the marginal transition probabilities follow $P(\theta)$, i.e. $\bar{P}(\theta, (x, y), \mathcal{A} \times \mathcal{X}) = P(\theta, x, \mathcal{A})$ and $\bar{P}(\theta, (x, y), \mathcal{X} \times \mathcal{A}) = P(\theta, y, \mathcal{A})$. Furthermore, $\bar{P}(\theta)$ satisfies the property that the marginal chains stay together after they meet. Let X_0, $Y_0 \sim \mu_0$ where μ_0 is some initial distribution on $\mathcal{X}$. For $L \geq 1$, we may construct a L-lagged coupled Markov chain $(X, Y) := \{X_i, Y_i\}_{i \geq 0}$. We first run only chain X for L steps where we generate $X_i \sim P(\theta, X_{i-1}, \cdot)$ for $i \leq L$. We then run the joint chain and generate future steps according to the joint kernel $\bar{P}(\theta)$ such that for $i \geq 1$, $(X_{i+L}, Y_i) \sim \bar{P}(\theta, (X_{i+L-1}, Y_{i-1}), \cdot)$. Let $\tau^L := \inf(i \geq L : X_i = Y_{i-L})$ and note that $f(X_i) = f(Y_{i-L})$ for all $i \geq \tau^L$. The coupling kernel $\bar{P}$ can be constructed using maximum coupling technique or using common random numbers with more details can be found in [18].

2.2 Notations and Assumptions

First we define some notations. For a function $h : \mathcal{X} \to \mathbb{R}$, a transition kernel P, a state $x \in \mathcal{X}$ and a probability measure μ, define the following notation:

$$Ph(x) := \int_{\mathcal{X}} h(x)P(x, dy); \tag{2.1}$$

$$\mu h := \int_{\mathcal{X}} h(x)\mu(dy). \tag{2.2}$$

Furthermore, for a function $V : \mathcal{X} \to \mathbb{R}$, we define the V-norm of function h as

$$\|h\|_V := \sup_{x \in X} \frac{h(x)}{V(x)}. \tag{2.3}$$

Assumption 1. *Assumption A4/A5 in [24] (in exact words): The family of one-step transition kernels $(P(\theta) : \theta \in \Lambda)$ is absolutely continuous with respect to $P(\theta_0)$, in the sense that there exists a density $(p(\theta, x, y) : \theta \in \Lambda, x, y \in S)$ for which*

$$P(\theta, x, dy) = p(\theta, x, y)P(\theta_0, x, dy)$$

for $x, y \in S$, and $\theta \in \Lambda$. Furthermore, there exists $\epsilon > 0$ for which $p(\cdot, x, y)$ is continuously differentiable on $[\theta_0 - \epsilon, \theta_0 + \epsilon]$ for each $x, y \in S$. Set $\omega_\epsilon(x, y) := \sup_{|\theta - \theta_0| < \epsilon} |p'(\theta, x, y)|$. Also assume that there exists a subset $A \subseteq S$, an integer $n \geq 1, \beta > 0$, and a probability φ for which

$$P^n(\theta, x, dy) \geq \lambda\varphi(dy)$$

for $x \in A, y \in S$, and $|\theta - \theta_0| < \epsilon$.

Assumption 2. *For a small set K, suppose there exists a measurable function $V : X \to [1, \infty)$ and constants $0 < \lambda < 1, b < \infty$ such that*

$$P(\theta)V(x) \leq \lambda V(x) + b\mathbb{1}(x \in K), \tag{2.4}$$

and the chains induced by $P(\theta)$ are irreducible and aperiodic for θ such that $|\theta - \theta_0| < \epsilon$.

Remark 1. *If we iteratively apply the $P(\theta_0)$ operator on both sides of (2.9), assumption 2 implies that for any $n \geq 1$:*

$$P^n(\theta)V(x) \leq V(x) + \frac{b}{1-\lambda} := \bar{V}(x). \tag{2.5}$$

Remark 2. *The Lyapunov condition assumed on function V implies that*

$$P(\theta)\sqrt{V(x)} \leq \sqrt{P(\theta)V(x)} \leq \sqrt{\lambda}\sqrt{V(x)} + \sqrt{b}\mathbb{1}(x \in K). \tag{2.6}$$

Since $0 < \sqrt{\lambda} < 1$ and $\sqrt{b} < \infty$, the function $\sqrt{V}$ is also a Lyapunov function.

Remark 3. *Assumption 2 also implies that for any $g : \mathcal{X} \to \mathbb{R}$ with $|g(x)| \leq |V(x)|$ for all $x \in \mathcal{X}$, there exist constants $R < \infty$ and $r < 1$ such that*

$$|\mathbb{E}_x[g(X_n)] - \pi(\theta)g| \leq RV(x)r^n, \tag{2.7}$$

and

$$\left|\mathbb{E}_x\left[\sqrt{g(X_n)}\right] - \pi(\theta)\sqrt{g}\right| \leq R\sqrt{V(x)}r^n, \tag{2.8}$$

for all $x \in \mathcal{X}$ and $n \geq 1$.

Assumption 3. *We also assume our starting distribution μ_0 satisfies that $\mu_0 V < \infty$.*

Assumption 4. *Let $\tau_{x,z}$ be the coupling time of two chains starting from state x and state z respectively. Assume the coupling time satisfies a geometrically decaying tail probability, i.e.*

$$P(\tau_{x,z} > t) \leq M \left(V(x) + V(z) \right) \rho^t, \tag{2.9}$$

for some $R < M < \infty$ and $r < \rho < 1$.

Assumption 1 is carried directly from [24]. Assumption 2 sometimes can be used to verify Assumption 4. See Section 3.2 in [17] for additional assumptions and details to do this. Assumption 4 sometimes can be extended to polynomially decaying tail probabilities, see [6][1].

2.3 Unbiased Estimation for the Stationary Mean

The intuitive arguments that lead to an unbiased estimator for the stationary mean starts from a telescoping sum technique [13] that for any $k \geq 0$:

$$\begin{aligned}
\gamma(\theta) &= \lim_{t \to \infty} \mathbb{E}\left[f(X_t) \right] & (2.10)\\
&= \mathbb{E}\left[f(X_k) \right] + \sum_{t=1}^{\infty} \mathbb{E}\left[f(X_{k+tL}) \right] - \mathbb{E}\left[f(X_{k+(t-1)L}) \right] & (2.11)\\
&= \mathbb{E}\left[f(X_k) \right] + \sum_{t=1}^{\infty} \mathbb{E}\left[f(X_{k+tL}) \right] - \mathbb{E}\left[f(Y_{k+(t-1)L}) \right] & (2.12)\\
&= \mathbb{E}\left[f(X_k) + \sum_{t=1}^{\left\lfloor \frac{\tau^L - k}{L} \right\rfloor} f(X_{k+tL}) - f(Y_{k+(t-1)L}) \right], & (2.13)
\end{aligned}$$

where the third equality follows from that both X and Y evolves according $P(\theta)$ and the last equality follows from that $f(X_{i+1}) = f(Y_i)$ for all $i \geq \tau^L$. Therefore,

$$H_k^L := f(X_k) + \sum_{t=1}^{\left\lfloor \frac{\tau^L - k}{L} \right\rfloor} f(X_{k+tL}) - f(Y_{k+(t-1)L}) \tag{2.14}$$

would be an unbiased estimator for $\gamma(\theta)$ for any $k \geq 0$ if (2.10) to (2.13) hold rigorously and so is an average of H_k^L for different k values $H_{k:m}^L := \frac{1}{m-k+1} \sum_{t=k}^{m} H_t^L$, where $m \geq k$. We may rewrite it as

$$H_{k:m}^L = \frac{1}{m-k+1} \sum_{t=k}^{m} f(X_t) + \frac{1}{m-k+1} \sum_{t=k+L}^{\tau^L - 1} c_t \left(f(X_t) - f(Y_{t-L}) \right), \tag{2.15}$$

where

$$c_t = \left\lfloor \frac{\min(m+L, t) - k - (t-k)\%L}{L} \right\rfloor. \tag{2.16}$$

If $L = 1$, then $H_{k:m}^L$ is equivalent to the stationary mean estimator developed in [17]. However, choosing a larger L is proposed and advocated in [27] as it generally improves the performance of the estimator, sometimes dramatically. The intuition behind choosing a larger L is that the coefficients c_t in can be greatly reduced and make $H_{k:m}^L$ behave more like a long-run average. [6], [1] and [26] contain more detailed analysis and examples of $H_{k:m}^L$.

2.4 Unbiased Estimation for the solution to the Poisson's Equation

Let g denote the solution to the Poisson's equation such that for any $x \in \mathcal{X}$, the following equality holds

$$g(x) - P(\theta)g(x) = f(x) - \pi(\theta)f, \tag{2.17}$$

The solution g has been shown to play an important role in analyzing the sum of the form $\sum_{t=0}^{n} f(X_t)$ for Markov chains [9]. Past works have analyzed the uniqueness of the solution and its connections to the Lyapunov conditions, see [2], [11] and [24]. The most basic form the solution function g could take is termed the **fundamental solution** to the Poisson's equation and is defined as:

$$g^{fu}(x) := \mathbb{E}_x^{\theta}\left[\sum_{t=0}^{\infty}(f(X_t) - \pi(\theta)f)\right]. \tag{2.18}$$

Note that (2.17) can admit various solutions that differ from each other by constants. One of the most exploited expression of the solution g depends on the regeneration structure of the analyzed Markov chain and attempt to estimate the following form:

$$g_{\alpha}^{re}(x) := \mathbb{E}_x^{\theta}\left[\sum_{t=0}^{T(\alpha)}(f(X_t) - \pi(\theta)f)\right] \tag{2.19}$$

where α is a regeneration state and $T(\alpha) := \inf\{t > 0 : X_t = \alpha\}$ is the regeneration time, e.g. in [4]. When the state space is uncountable, sometimes we are still able to construct regeneration times using a splitting technique [22]. However, the constructed regeneration time could be unnecessarily long causing a large variance for the estimator. Since we do not assume the existence or the ability to construct efficient regeneration times, we focuses on a different form that takes advantage of the existence of a coupling kernel, first introduced in [6]. For a pre-specified state $z \in \mathcal{X}$, define:

$$\begin{aligned} g_z(x) &:= g^{fu}(x) - g^{fu}(z) \\ &= \mathbb{E}_x^{\theta}\left[\sum_{t=0}^{\infty}(f(X_t) - \pi(\theta)f)\right] - \mathbb{E}_z^{\theta}\left[\sum_{t=0}^{\infty}(f(X_t) - \pi(\theta)f)\right] \\ &= \mathbb{E}_{x,z}^{\theta}\left[\sum_{t=0}^{\infty}(f(X_t) - f(Z_t))\right] \\ &= \mathbb{E}_{x,z}^{\theta}\left[\sum_{t=0}^{\tau_{x,z}}(f(X_t) - f(Z_t))\right] \end{aligned} \tag{2.20}$$

where $(X_t, Z_t)_{t=1,2,\ldots}$ is a joint Markov chain that starts from (x, z) and evolves according to $\bar{P}(\theta)$ and $\tau_{x,z} = \inf\{t \geq 0 : X_t = Z_t\}$. Thus naturally $\sum_{t=0}^{\tau_{x,z}}(f(X_t) - f(Z_t))$ would be an unbiased estimator for $g_z(x)$ which is a solution to (2.17).

3 Unbiased Derivative Estimators

3.1 Derivative estimator with stationary mean approach

Let θ_0 denote our current parameter at which we are trying to compute the derivative. In this section, we present differentiability criterion for the steady-state mean and a probabilistic representation of $\gamma'(\theta_0)$ based on which we develop an unbiased estimators.

Proposition 1. *(Theorem 4.1 in [24]) With Assumption 1, let $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$be a function for which $\kappa(x) \geq x$ and $\kappa(x)/x \to \infty$ as $x \to \infty$. Suppose that there exist positive constants ϵ, c_0, and c_1, and non-negative finite-valued functions q, v_0, and v_1 for which*

$$
\begin{aligned}
&(P(\theta)v_0)(x) \leq v_0(x) - (q(x) \vee 1) + c_0 \mathbb{I}(x \in A) \\
&(P(\theta)v_1)(x) \leq v_1(x) - \kappa\left(\int_S (1 \vee \omega_\epsilon(x,y))(v_0(y)+1) P(\theta, x, dy)\right) + c_1 \mathbb{I}(x \in A)
\end{aligned}
$$

for $x \in S, |\theta - \theta_0| < \epsilon$, and

$$\sup_{x \in A} v_0(x) < \infty.$$

Then

$$\gamma'(\theta_0) = \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[p'(\theta_0, X_0, X_1) g(X_1)\right]. \tag{3.1}$$

where g is a solution to (2.17).

Refer to Section 4 in [24] for more details and intuitions. Essentially (3.1) writes the gradient in (1.1) as a stationary mean of another function under $\pi(\theta_0)$. Note that Assumption 2-4 were not necessary in Proposition 1 but the geometric ergodicity entailed by those assumptions will be useful in efficiency analysis. Here we focus on estimating the derivative based on the probabilistic representation (3.1). First we consider g_z in place of g in (3.1) and define

$$G_z(x) := \sum_{t=0}^{\tau_{x,z}} (f(X_t) - f(Z_t)) \tag{3.2}$$

as an unbiased estimator for $g_z(x)$. Observe that by conditioning on X_0, we may rewrite (2.17) to

$$\gamma'(\theta_0) = \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[p'(\theta_0, X_0, X_1) g(X_1)\right] = \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[h(X_0)\right], \tag{3.3}$$

where

$$h(x) := \mathbb{E}^{\theta_0}_{x}\left[p'(\theta_0, x, X_1) g(X_1)\right]. \tag{3.4}$$

Now we define

$$H(x_0, x_1) = p'(\theta_0, x_0, x_1) G_z(x_1), \tag{3.5}$$

where $G_z(x_1)$ is independent given x_1 and we make the dependence on choice of state z implicit in $H(x_0, x_1)$ and whenever obvious later. Thus built upon the unbiased stationary mean estimator (2.14), we define our first naive derivative estimator:

$$H_1^{k,L}(X,Y) := H(X_k, X_{k+1}) + \sum_{t=1}^{\left\lfloor \frac{\tau^L - k}{L} \right\rfloor} H(X_{k+tL}, X_{k+tL+1}) - H(Y_{k+(t-1)L}, Y_{k+(t-1)L+1}). \tag{3.6}$$

$H_1^{k,L}(X,Y)$ is indeed an unbiased estimator for $\gamma'(\theta_0)$ with finite second moment and finite computation time with appropriate technical conditions in place. The formal proof will be delayed to later this section where we prove the result for a more generalized version of the derivative estimator. One of the key differences between $H_1^{k,L}(X,Y)$ and previously studied stationary mean estimators is that $H(x_0, x_1)$ is non-trivial to evaluate as it requires simulating a joint chain until they couple. Moreover, $H(x_0, x_1)$ is a random variable for fixed input (x_0, x_1) and could potentially introduces a significant amount of additional variance, especially when the coupling time tends to be long causing the variances of the G_z term to be large. Therefore, to make a more practical estimator, we need machinery to improve efficiency for each of the $H(x_0, x_1)$ terms in (3.6).

3.2 Derivative estimator with the L-skeleton chain approach

We note that $\gamma(\theta_0)$ stays the same when we consider the L-skeleton chain and so should its derivative $\gamma'(\theta_0)$. First we present an intuitive argument to develop the L-skeleton version of the derivative representation followed by a rigorously proved theorem. First see that for the L-step transition kernel $P^L(\theta) := (P(\theta, x, dy) : x, y \in \mathcal{X})$,

$$\begin{aligned}
P^L(\theta, x_0, dx_L) &= \int_{\mathcal{X}} P^{L-1}(\theta, x_0, dx_{L-1}) P(\theta, x_{L-1}, dx_L) dx_{L-1} \qquad (3.7)\\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} P^{L-2}(\theta, x_0, dx_{L-2}) P(\theta, x_{L-2}, dx_{L-1}) dx_{L-2} \right) P(\theta, x_{L-1}, dx_L) dx_{L-1} \qquad (3.8)\\
&= \int_{\mathcal{X}^{L-1}} \prod_{i=0}^{L-1} P(\theta, x_i, dx_{i+1}) dx_1 dx_2 \cdots dx_{L-1}. \qquad (3.9)
\end{aligned}$$

Then, assume the validity for interchanging derivatives and integrals, we may rewrite (3.1) as

$$\begin{aligned}
\gamma'(\theta_0) &= \mathbb{E}^{\theta_0}_{\pi(\theta_0)} \left[\frac{\frac{d}{d\theta} P^L(\theta, X_0, dX_L)|_{\theta_0}}{P^L(\theta_0, X_0, dX_L)} g^L(X_L) \right] \qquad (3.10)\\
&= \mathbb{E}_{X_0 \sim \pi(\theta_0)} \left[\mathbb{E} \left[\frac{\frac{d}{d\theta} P^L(\theta, X_0, dX_L)|_{\theta_0}}{P^L(\theta_0, X_0, dX_L)} g^L(X_L) | X_0 \right] \right] \qquad (3.11)\\
&= \mathbb{E}_{X_0 \sim \pi(\theta_0)} \left[\int \frac{d}{d\theta} P^L(\theta, X_0, dX_L)|_{\theta_0} g^L(X_L) \right] \qquad (3.12)\\
&= \mathbb{E}^{\theta_0}_{\pi(\theta_0)} \left[\left(\sum_{i=0}^{L-1} p'(\theta_0, X_i, X_{i+1}) \right) g^L(X_L) \right], \qquad (3.13)
\end{aligned}$$

where the first equality is (3.1) with the L-step kernel $P^L(\theta)$ and for now is assumed to hold without further assumptions; the last equality comes from applying chain rule and divide and multiply $\prod_{i=0}^{L-1} P(\theta_0, x_i, dx_{i+1})$, and g^L is a solution to the Poisson's equation associated with the L-skeleton Markov chain that satisfies the following equation

$$g^L - P^L(\theta) g^L = f - \pi(\theta) f. \tag{3.14}$$

Next we show that the L-skeleton version of the derivative representation holds rigorously without further assumptions other than the ones introduced in Proposition 1. This saves us the trouble to verify conditions for the L-step transition kernel, which is usually not straightforward.

Theorem 2. *Assume that* (3.1) *holds, then*

$$\gamma'(\theta_0) = \mathbb{E}^{\theta_0}_{\pi(\theta_0)} \left[\left(\sum_{i=0}^{L-1} p'(\theta_0, X_i, X_{i+1}) \right) g^L(X_L) \right], \tag{3.15}$$

for all $L \geq 1$.

The proof is deferred to Section 7.1. Naturally, here for a pre-specified state $z \in \mathcal{X}$, we develop the L-skeleton chain counterpart for (3.2) as:

$$G^L_z(x) := \sum_{i=0}^{\lfloor \frac{\tau_{x,z}}{L} \rfloor} f(X_{iL}) - f(Z_{iL}), \tag{3.16}$$

and for the same reason as before, this is an unbiased estimator for

$$g^L_z(x) := \mathbb{E} \left[\sum_{i=0}^{\infty} f(X_{iL}) - f(Z_{iL}) \right], \tag{3.17}$$

where g_z^L is a solution to (3.14). Thus similar to Section 3.1, first let

$$H(\{x_i\}_{i=0}^L) := \left(\sum_{i=0}^{L-1} p'(\theta_0, x_i, x_{i+1}) \right) G_z^L(x_L), \tag{3.18}$$

and then we can define our second derivative estimator to be

$$H_2^{k,L}(X,Y) := SE_k^L + BC_k^L \tag{3.19}$$

where

$$SE_k^L := H\left(\{X_i\}_{i=k}^{k+L}\right); \tag{3.20}$$

$$BC_k^L := \sum_{t=1}^{\left\lfloor \frac{\tau^L - k}{L} \right\rfloor} H\left(\{X_i\}_{i=k+tL}^{k+(t+1)L}\right) - H\left(\{Y_i\}_{i=k+(t-1)L}^{k+tL}\right). \tag{3.21}$$

represents the singleton estimator (SE) and and bias correction (BC) term to the singleton estimator respectively. The motivation to develop this estimator from the L-skeleton chain is the greatly reduced variance when considering G_z^L instead of G_z since G_z^L is a sum of much fewer random variables when L is large. The price we pay for the reduced variance in G_z^L is the increased variance in the $\sum_{i=0}^{L-1} p'(\theta_0, x_i, x_{i+1})$ term in (3.18) as L gets larger. However, under appropriate conditions, we can show that $\{M_n\}_{n=0,1,2}$ with $M_0 = 0$ and

$$M_n := \sum_{i=0}^{L-1} p'(\theta_0, x_i, x_{i+1}) \tag{3.22}$$

for $n \geq 1$ is a martingale. Thus the variance increase for the martingale term usually gets outweighed by the sometimes dramatic variance decrease in the estimation of the solution to the Poisson's equation associated with the L-skeleton chain and causing a great improvement in efficiency. Nevertheless, it is unwise to choose L to be arbitrarily large as the efficiency of our estimator does not monotonically improve as L gets larger. The analysis of the efficiency and a guideline for choosing L will be discussed in the next section. Here we first prove our result rigorously.

Assumption 5. *Suppose there exists a $p > 1$, $\kappa > 1$ and a $\delta > 0$ such that $\left|f^{2\kappa p+\delta}\right|_V, \left|\Omega^{\frac{2\kappa p}{p-1}}\right|_V < \infty$, where*

$$\Omega^{\frac{2\zeta p}{p-1}}(x) = \int_{\mathcal{X}} \omega_\epsilon(x,y)^{\frac{2\zeta p}{p-1}} P(\theta_0, x, dy), \tag{3.23}$$

$$f^{2\zeta p+\delta}(x) = f(x)^{2\zeta p+\delta}. \tag{3.24}$$

Lemma 3. *With assumptions 1-5, if we define for $1 \leq \zeta \leq \kappa$:*

$$\Gamma_\zeta^L(x) := \mathbb{E}_x^{\theta_0}\left[H\left(\{X_i\}_{i=0}^L\right)^{2\zeta}\right], \tag{3.25}$$

then we have $\Gamma_\zeta^L(x) \leq V(x) U_{\zeta,z}(L)$ where

$$U_{\zeta,z}(L) := \left(L^\zeta A_z + L^\zeta \left(\frac{\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}{1 - \left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L} \right)^{2\zeta} B_z \right), \tag{3.26}$$

and

$$A_z^\zeta := 2^{4\zeta-2} \left|f^{2\zeta p+\delta}\right|_V^{\frac{2\zeta}{2\zeta p+\delta}} \left(C_{\frac{2\zeta p}{p-1}} |\Omega^{\frac{2\zeta p}{p-1}}|_V \left(1 + \frac{b}{1-\lambda}\right)\right)^{\frac{p-1}{p}} \left(1 + \left(\frac{b}{1-\lambda}\right)^{\frac{2\zeta}{2\zeta p+\delta}} + \bar{V}(z)^{\frac{2\zeta}{2\zeta p+\delta}}\right), \tag{3.27}$$

$$B_z^\zeta := A_z^\zeta \cdot M^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}} \left(1 + \left(\frac{b}{1-\lambda}\right)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}} + V(z)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\right). \tag{3.28}$$

One direct implication is that $\left|\Gamma_\zeta^L\right|_V \leq U_{\zeta,z}(L) < \infty$*.*

The proof is in Section 7.2. Next we prove the formal results for our estimator $H_2^{k,L}(X,Y)$.

Theorem 4. *Under assumptions 1, 2, 3, 4 and 5, $H_2^{k,L}(X,Y)$ is an unbiased estimator for $\gamma'(\theta)$ with finite second moment and finite expected computation time. Furthermore,*

$$\mathbb{E}_{\mu_0}^{\theta_0}\left[\left(BC_k^L\right)^2\right] \to 0 \tag{3.29}$$

and

$$\left|\mathbb{E}_{\mu_0}^{\theta_0}\left[\left(SE_k^L\right)^2\right] - \pi(\theta_0)\Gamma_1^L\right| \to 0 \tag{3.30}$$

as $k \to \infty$.

The proof of the Theorem is in Section 7.3. Theorem 4 makes rigorous some obvious intuitions. The second moment (and thus the first moment) goes to 0 as the starting point k increases since X_k becomes closer and closer to being distributed as $\pi(\theta_0)$. Thus we may choose k to be large such that the SE_k^L term dominates and the variance of the estimator converges to the variance of the singleton estimator if we were to start the chain from stationary $\pi(\theta_0)$. However, even though choosing a large burn-in period k may reduce variance, it is not economic to just generate a one sample estimator and discard all previously simulated chains. Therefore, a running average version of the estimator may be of greater interest.

3.3 Averaging the L-skeleton chain estimator

Many previous works on estimating the stationary expectation calls for a running average version of their estimators, see [17], [6] [1]. However, in our scenario, evaluating the functional of which we are trying to take stationary expectation actually takes a significantly amount of time. To be specific, we need to independently run two chains until they couple every time we are computing G_z. Assume we define a running average version similar to (2.15), we will get for some $m \geq k$,

$$H_3^{k,m,L}(X,Y) := \frac{1}{m-k+1}\sum_{t=k}^{m} H_2^{t,L}(X,Y). \tag{3.31}$$

One major problem with this running average estimator is the correlation between consecutive estimators $H_2^{t,L}(X,Y)$ and $H_2^{t+1,L}(X,Y)$. Assume that we take k to be a large quantile of the coupling time $\tau^L - L$ so that the bias correction term plays little role [17], then notice that

$$H_2^{t,L}(X,Y) \approx H\left(\{X_i\}_{i=t}^{t+L}\right) = \left(\sum_{i=t}^{t+L-1} p'(\theta_0, X_i, X_{i+1})\right) G_z^L(X_{t+L}); \tag{3.32}$$

$$H_2^{t+1,L}(X,Y) \approx H\left(\{X_i\}_{i=t+1}^{t+1+L}\right) = \left(\sum_{i=t+1}^{t+L} p'(\theta_0, X_i, X_{i+1})\right) G_z^L(X_{t+L+1}), \tag{3.33}$$

where those two consecutive estimators can be strongly correlated even with moderate choice of L since they share a great number of terms in the summation, specifically $\sum_{i=t+1}^{t+L-1} p'(\theta_0, X_i, X_{i+1})$ appears in both of the consecutive estimators. Also, $G_z^L(X_{t+L})$ and $G_z^L(X_{t+L+1})$ can also be correlated since X_{t+L} and X_{t+L+1} are one step away. Meanwhile, two independent coupled chains needs to be simulated to compute $G_z^L(X_{t+L})$ and $G_z^L(X_{t+L+1})$, causing a significant computational burden. One may naturally have the doubt that is it worth it to spend this many computational resources to generate two highly correlated estimators? [23] studies exactly such a problem and argues that if the cost to evaluate the function once outweighs the cost to advance the Markov chain for one step, discarding a portion of the samples can significantly increase efficiency. Thus here we propose a more general form of the running average estimator that utilizes the thinning technique in [23]. For some $m \geq 1$, we define:

$$H_4^{k,m,L}(X,Y) := \frac{1}{m} \sum_{t=0}^{m-1} H_2^{k+tL,L}(X,Y). \tag{3.34}$$

Note that for each of the $t = 1, 2, \ldots, m-1$, Theorem 4 applies to $H_2^{k+tL,L}(X,Y)$. Since $H_4^{k,m,L}(X,Y)$ is an average of them, a simple application of Minkowski's inequality would yield that $H_4^{k,m,L}(X,Y)$ also has finite second moment and thus it is unbiased and has computation time. Similarly, if k is chosen to be a large quantile of $\tau^L - L$, then the average of the singleton estimators is going to dominate and

$$H_4^{k,m,L}(X,Y) \approx \frac{1}{m} \sum_{t=0}^{m-1} H\left(\{X_i\}_{i=k+tL}^{k+(t+1)L}\right), \tag{3.35}$$

where we usually take m to be large so that the burn-in cost k becomes less significant. The asymptotic variance of $H_4^{k,m,L}(X,Y)$ is then approximately $\sigma_L^2 + 2\sum_{j=1}^{\infty} \sigma_{j,L}$ where

$$\begin{aligned} \sigma_L^2 &:= Var_{\pi(\theta_0)}\left(H\left(\{X_i\}_{i=0}^{L}\right)\right), \\ \sigma_{j,L} &:= Cov_{\pi(\theta_0)}\left(H\left(\{X_i\}_{i=0}^{L}\right), H\left(\{X_i\}_{i=jL}^{(j+1)L}\right)\right). \end{aligned} \tag{3.36}$$

That is,

$$m \cdot Var(H_4^{k,m,L}(X,Y)) \approx \sigma_L^2 + 2\sum_{j=1}^{\infty} \sigma_{j,L}. \tag{3.37}$$

Notice that the consecutive estimators such as $H\left(\{X_i\}_{i=0}^{L}\right)$ and $H\left(\{X_i\}_{i=L}^{2L}\right)$ will not share any overlapping terms from chain X and thus the correlation should intuitively be small whenever L is not too small. We shall make more quantified analysis of the efficiency of our estimator in the next section.

4 Efficiency Analysis and Parameter Selections

In this section, we will focus on analyzing in detail the inefficiency of our estimator which is quantified by the work-variance product. From the analysis results, we aim to practical guidelines for choosing other important parameters such as the lagging parameter L and the choice of the fixed state z. To make the words concrete, if we choose k and m as suggested in the previous section such that approximations (3.35) and (3.37) hold, then the expected number of Markov transitions to produce one $H_4^{k,m,L}(X,Y)$ is approximately $k + m(L + 2\mathbb{E}[\tau_{\pi(\theta_0),z}])$ where $\tau_{\pi(\theta_0),z}$ is the coupling time of two chains starting from stationary distribution $\pi(\theta_0)$ and a pre-specified state z. For specific breakdowns, k is the burn-in time for chain X; $L + 2\mathbb{E}[\tau_{\pi(\theta_0),z}]$ is the cost to advance from X_{k+tL} to $X_{k+(t+1)L}$ plus the expected cost to generate independently a pair of coupled chain to estimate the solution to the Poisson's equation G_z^L. The multiplier m represents those steps will be repeated for m times. Thus we aim to analysis the following work-variance product:

$$\left(k+m\left(L+2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right)\right)Var(H_4^{k,m,L}(X,Y)) \approx \left(\frac{k}{m}+L+2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right)\left(\sigma_L^2+2\sum_{j=1}^{\infty}\sigma_{j,L}\right) \tag{4.1}$$

$$\approx \left(L+2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right)\left(\sigma_L^2+2\sum_{j=1}^{\infty}\sigma_{j,L}\right) \tag{4.2}$$

where the approximation follows from (3.37).

Lemma 5. *With Assumptions 1-5, the asymptotic variance is bounded by*

$$\left(\sigma_L^2+2\sum_{j=1}^{\infty}\sigma_{j,L}\right) \le U_{1,z}(L)\pi(\theta_0)V\left(3+\frac{2M\rho^L}{1-\rho^L}\right). \tag{4.3}$$

The proof is recorded in Section 7.4. With a quantified upper bound on the asymptotic variance, we then aim to develop general guidelines of parameter selection to reduce work-variance product of our estimator.

Choice of z: First we see that

$$\begin{aligned} P_{X_0\sim\pi(\theta_0)}\left(\tau_{X_0,z}>t\right) &= \int_{\mathcal{X}} P\left(\tau_{y,z}>t\right)\pi(\theta_0,dy) \\ &\le \int_{\mathcal{X}} M(V(y)+V(z))\rho^t\pi(\theta_0,dy) \\ &\le M\left(\pi(\theta_0)V+V(z)\right)\rho^t. \end{aligned} \tag{4.4}$$

Therefore $\mathbb{E}\left[\tau_{\pi(\theta),z}\right] \le \frac{M}{1-\rho}\left(\pi(\theta_0)V+V(z)\right)$. Furthermore, note that the upper bound $U_{1,z}(L)$ also decreases monotonically as $V(z)$ decreases. Thus the most obvious choice of z is to make it a state such that $V(z)$ is as small as possible. It makes sense in most cases as a smaller $V(z)$ usually entails that z is in a neighborhood that gets visited more often and more close to the typical behavior of the Markov chain.

Choice of L: On the other hand, the choice of L can be more sophisticated and can depend heavily on quantities that we do not have easy access to such as the smallest geometric convergence rate ρ that satisfies our assumptions. Nevertheless we provide a guideline that depends on available information and show that the performance can be at least x times better than not considering that L-skeleton chain, e.g. the case where $L=1$.

Theorem 6. *Suppose we define*

$$W(L) := \left(L+2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right)U_{1,z}(L)\pi(\theta_0)V\left(3+\frac{2M\rho^L}{1-\rho^L}\right), \tag{4.5}$$

then

$$\frac{W(1)}{W(\mathbb{E}[\tau_{\pi(\theta_0),z}]-1)} > \frac{\mathbb{E}[\tau_{\pi(\theta_0),z}]-1}{3}. \tag{4.6}$$

The proof is recorded in Section 7.5. Theorem 6 suggests that if we select $L=\mathbb{E}[\tau_{\pi(\theta_0),z}]-1$, then the work-variance product upper bound would be reduced to at most $\left(\mathbb{E}[\tau_{\pi(\theta_0),z}]-1\right)/3$ fraction of the original one with $L=1$. When the Markov chain mixes slowly, the expected coupling time can be quite long thus resulting a significant improvement. This is a useful guideline as it only requires estimation of the expected coupling time which can be done rather easily. The theorem suggests that the improvement is at least this much and usually the improvement can be much larger in actual practices. Also $L = \mathbb{E}[\tau_{\pi(\theta_0),z}] - 1$ is rarely the optimal choice to minimize inefficiency but rather one is encouraged to use it as a starting point and try bigger L values.

For the previously developed unbiased estimators for the regular stationary mean as described in Section 2.3, the choice of L is suggested to be a large quantile of the coupling time which is the same as the guideline for selecting k and such choice would incur very minor cost [1]. In our setting, the choice should be made with more care since the $L^2 A_z^1$ term in $W(L)$ increases with square rate. Nevertheless, for a slow mixing Markov chain, it is generally the case that choosing an L too large is still better than choosing and L too small because $\frac{\rho^L}{1-\rho^L}$ decreases much faster than L increases when ρ is close to 1, which corresponds to the slow mixing case.

5 Transition kernel represented by neural networks

One of the key advantages of our estimator is the minimum requirement to implement and the applicability to parametrizations with neural networks. Likelihood ratio based methods potentially can be used for NN parametrizations as well but require extra assumptions such as identifiable regenerative structure or knowledge of the exact value of the stationary mean which are often absent. We present two most fundamental examples of such parametrizations with no additional assumptions where our gradient estimator could be used.

5.1 Policies parametrized by Neural Networks

Consider the reinforcement learning (RL) setting where we can control the Markov Decision Process (MDP) by changing the parametrized policy. Given current state s, we abuse the notation a little in this section and denote for now that $\pi_\theta(a|s)$ as a policy represented by a neural network where θ is the weights and biases of the network. $\pi_\theta(\cdot|s)$ outputs a probability distribution on the action space $\mathcal{A}$. The system dynamics, represented by $P(\cdot|s,a)$ will give the probability distribution for the next state s'. The system dynamics here is assumed to be known or learned. In this case,

$$P(\theta, s, s') = \int_{\mathcal{A}} \pi_\theta(a|s) P(s'|s,a), \tag{5.1}$$

and if we may interchange the derivative and the integration,

$$\frac{d}{d\theta} P(\theta, s, s') = \int_{\mathcal{A}} \frac{d}{d\theta} \pi_\theta(a|s) P(s'|s,a). \tag{5.2}$$

If the action space $\mathcal{A}$ is finite (such as in a preemptive queuing control problem), then given (s, s'), both of the above expressions can be evaluated exactly as $\pi_\theta(a|s)$ can be obtained by forward feeding the neural network and $\frac{d}{d\theta}\pi_\theta(a|s)$ can be obtained by back-propagation. Thus the oracle to compute the transition density and its derivative is available. When the action space is uncountable or infinite, we may consider making the state-action pair (s, a) as a state of a markov chain on the joint state-action space $\mathcal{S} \times \mathcal{A}$. Then the transition kernel becomes

$$P(\theta, (s,a), (s',a)) = P(s'|s,a)\pi_\theta(a'|s'), \tag{5.3}$$

and assuming the interchange between the derivative and the integration,

$$\frac{d}{d\theta} P(\theta, (s,a), (s',a)) = P(s'|s,a) \frac{d}{d\theta} \pi_\theta(a'|s'). \tag{5.4}$$

Thus with the objective being optimizing the long-run average or the steady state mean of some performance/cost metric of the MDP, our gradient estimator can be easily implemented and be very useful when conducting gradient decent type of algorithms.

5.2 Dynamics Models parametrized by Neural Networks

A different setting is when the system transition dynamics are learned from data. This genre of RL methods are usually called model-based RL. With a learned dynamics model, we can generate samples from it and then conduct policy optimizations. Since the optimized is conducted upon the learned dynamics model instead of the true system dynamics, one natural question practitioners may ask is how sensitive the performance of the learned policy is when the learned dynamics changes. This is a sensitivity analysis type of question where our gradient estimator can be of use. Next we discuss some details on how system dynamics can be parametrized. The easiest way to parametrize the dynamics is through a NN that is a deterministic mapping from a state-action pair to another state. A more general and sophisticated parametrization is a probabilistic one where the NN maps a state-action pair to parameters of a parametrized distribution. The probabilistic model has been shown to exemplify great performance comparable model-free RL methods [3]. Let $\mathcal{D}$ denote a training dataset that stores the transitions that has been experienced, then one can minimize the negative log prediction density (NLPD) as our loss function:

$$loss(\phi) := -\sum_{(s_t,a_t,s_{t+1})\in\mathcal{D}} \log P_\phi\left(s_{t+1}|s_t,a_t\right). \tag{5.5}$$

where $P_\phi\left(s_{t+1}|s_t,a_t\right)$ is the transition density under NN parameter ϕ. Let $\phi^* := argmin_\phi loss(\phi)$ be the learned dynamics model parameter, then we can simulate the MDP with the learned transition dynamics $s_{t+1} \sim P_\phi\left(\cdot|s_t,a_t\right)$. In this case, let π^* be an optimized policy, then

$$P(\phi,(s,a),(s',a')) = P_\phi\left(s'|s,a\right)\pi^*(a'|s'), \tag{5.6}$$

and assuming the interchange between the derivative and the integration

$$\frac{d}{d\phi}P(\phi,(s,a),(s',a')) = \frac{d}{d\phi}P_\phi\left(s'|s,a\right)\pi^*(a'|s'). \tag{5.7}$$

A more specific example where the NN maps to a Gaussian distribution can be found in [3]. The transition densities and their derivatives can be computed from forward-feeding and back-propagations of NNs again. Thus, with no other significant requirements, for an optimized policy π^*, our estimator can be used to perform sensitivity analysis regarding to the learned dynamics model.

6 Numerical Experiments

In this section, we provide numerical experiment results on three different scenarios: a heavy-traffic single server queue waiting time sequence where the state space is continues; a multi-class queuing network control problem where the policy is parametrized by a neural network; and last an ising model control problem where the control is also parametrized by a neural network but with higher-dimensional state space.

6.1 $M/M/1$ Queue

We first illustrate our results in a simple $M/M/1$ customer waiting time sequence setting. Although our estimator has almost minimum requirement to implement, many of the previous gradient estimators that we want to compare with require analytical knowledge on the parametrized transition kernel or specific ways of parametrization and the $M/M/1$ example satisfies all the analytical requirements. Nevertheless, the simple queuing model can still pose complex behavior and make the estimation task hard under heavy traffic scenarios because of the long regeneration cycles and the slow mixing rate. For the remainder of the section, X_n represents the waiting time of the nth customer where $X_0 = 0$.

The inter-arrival rate is fixed at 5 and the service rate is θ. Thus our Markov chain evolves according to the Lindley recursion:

$$X_{n+1} = \max(0, X_n + S_n^\theta - T_n), \tag{6.1}$$

where S_n^θ's are independent exponential random variables with rate θ and T_n's are independent exponential random variables with rate 5. The mean waiting time under stationarity is of interest here so in this case $\gamma(\theta) = \mathbb{E}_{X\sim\pi(\theta)}[X]$. The goal is to estimate $\frac{d}{d\theta}\gamma(\theta)|_{\theta=\theta_0}$ where $\theta_0 = 5.2$ in our experiment. The true analytical derivative here is approximately -24.96. We then present several benchmark gradient estimators and the details on how they are implemented in the $M/M/1$ setting.

6.1.1 The IPA estimator [8]

Let $X(\theta) = \{X_n(\theta); n \geq 0\}$ be a Markov chain parametrized by θ. Then under certain conditions,

$$\gamma'(\theta) = \mathbb{E}[X_\infty(\theta)]' = \mathbb{E}[X'_\infty(\theta)] = \frac{\mathbb{E}\Big[\sum_{i=\tau_{k-1}}^{\tau_k} X'_i(\theta)\Big]}{\mathbb{E}[\tau_k - \tau_{k-1}]}, \tag{6.2}$$

where $X'(\theta) := \{X'_n(\theta), n \geq 0\}$ represents the derivative sequence of the process X and τ_k represents the k-th regeneration time of X'. This paper mainly considers the special case where

$$X_{n+1}(\theta) = \phi(X_n(\theta), U_n(\theta)), \quad n \geq 0, \tag{6.3}$$

for some recursive function ϕ and input sequence $U(\theta) = \{U_n(\theta); n \geq 0\}$. Our $M/M/1$ setting fits exactly here. Note that by inverse CDF method, we know $-\frac{1}{\theta}\ln(U)$ follows the same distribution as S_n^θ where $U \sim Unif(0,1)$. Thus in our case, the waiting time evolves according to

$$X_{n+1} = \max(0, X_n - \frac{1}{\theta}\ln(U_n) - T_n), \tag{6.4}$$

then

$$X'_{n+1} = X'_n + \frac{1}{\theta^2}\ln(U_n), \quad for \quad X_n(\theta) - \frac{1}{\theta}\ln(U_n) - T_n > 0, \tag{6.5}$$

$$X'_{n+1} = 0, \qquad\qquad otherwise. \tag{6.6}$$

Thus the problem is then transformed to estimating the stationary mean of the derivative process X' while X' possesses the same regenerative structure as X since both processes hit 0 at the same moment. Now define

$$H^N_{IPA} := \frac{\frac{1}{N}\sum_{i=1}^N H^{(i)}}{\frac{1}{N}\sum_{i=1}^N \tau^{(i)}}, \tag{6.7}$$

where $\tau^{(i)}$'s are i.i.d. copies of the length of one regeneration cycle, and $H^{(i)}$'s are i.i.d. copies (independent from the $\tau^{(i)}$'s already generated) of the sum over one regeneration cycle defined as

$$H^{(i)} := \sum_{j=\tau_{i-1}}^{\tau_i} X'_i. \tag{6.8}$$

Hence based on (6.2), H^N_{IPA} is the proposed estimator for $\gamma'(\theta_0)$. Table 1 records the performance of H^N_{IPA} with different values N. Note that the regeneration structure is utilized here and hence caused the initial biases for small N's but it is asymptotically unbiased.

6.1.2 The Phantom Estimator [16]

Assume that for any $x \in \mathcal{X}$, $P'(\theta)$ exists such that

$$\frac{d}{d\theta}\int_S P(\theta,x,dy)g(y) = \int_S P'(\theta,x,dy)g(y) \tag{6.9}$$

for any $x \in \mathcal{X}$. Then let $\left([P'(\theta)]^+(x;\cdot),[P'(\theta)]^-(x;\cdot)\right)$ denote the Hahn-Jordan decomposition of the signed measure $P'(\theta,x,\cdot)$ and define the normalizing constant to be

$$c_{P(\theta)}(x) := [P'(\theta)]^+(x;\mathcal{X}) = [P'(\theta)]^-(x;\mathcal{X}).$$

Also define $P^+(\theta)$ and P^- to be the normalized transition kernel corresponding to the Hahn-Jordan decomposition:

$$P^+(\theta,x,\cdot) = \frac{[P'(\theta)]^+(x;\cdot)}{c_{P(\theta)}(x)}, \quad P^-(\theta,x,\cdot) = \frac{[P'(\theta)]^-(x;\cdot)}{c_{P(\theta)}(x)}. \tag{6.10}$$

In the $M/M/1$ setting, an easy decomposition is to decompose the exponential service time $Exponential(\theta)$ to be a difference between an $Exponential(\theta)$ r.v. and a $Gamma(2,\theta)$ r.v. and the normalizing constant $c_{P(\theta)}(x) = 1/\theta$ for all x. For decompositions of more common distributions, refer to Table A.1 in [7]. Next, we introduce the so called phantom processes $X^+ := \{X_n^+, n \geq 0\}$ and $X^- := \{X_n^-, n \geq 0\}$ where $X_0^+ = X_0^- = x$ for some starting point x, the first transitions from X_0^+ to X_1^+ and from X_0^- to X_1^- are governed by kernels $P^+(\theta)$ and $P^-(\theta)$ respectively. After the first transition, for $n \geq 2$, X_n^+ and X_n^- evolve according to the regular coupled kernel $\bar{P}(\theta)$. Then under some ergodicity conditions, our desired quantity can be written as

$$\gamma'(\theta) = \mathbb{E}_{X_0\sim\pi(\theta)}\left[\mathbf{D}\left(P(\theta),g;X_0\right)\right], \tag{6.11}$$

where

$$\mathbf{D}\left(P(\theta),g;x\right) = \mathbb{E}_x\left[c_{P(\theta)}(x)\sum_{n=0}^{\tau_x^{\pm}}\left(g(X_n^+) - g\left(X_n^-\right)\right)\right], \tag{6.12}$$

and $\tau^{\pm} = \inf\{n \in \mathbb{N} : X_n^+ = X_n^-\}$ is the coupling time between X^+ and X^-. Thus (6.11) transforms the derivative estimation into a stationary mean estimation for a different function. In the single server queue setting, 0 is an atom for the waiting time sequence and with that available, [16] proposes to utilize the following expression to estimate the derivative:

$$\gamma'(\theta_0) = \frac{1}{\mathbb{E}\left[\tau_\theta^0\right]}\mathbb{E}_{X_0=0}\left[\sum_{i=0}^{\tau_\theta^0-1}\mathbf{D}\left(P(\theta),g;X_i\right)\right] \tag{6.13}$$

where $\tau_\theta^0 := \inf\{n \geq 1 : X_n = 0\}$ is the first time the chain hits 0. Now define

$$\hat{\mathbf{D}}\left(P(\theta),g;x\right) = \frac{1}{\theta}\sum_{n=0}^{\tau_x^{\pm}}\left(g(X_n^+) - g\left(X_n^-\right)\right), \tag{6.14}$$

to be a estimator for $\mathbf{D}\left(P(\theta),g;x\right)$. Then assume we generate N regeneration cycles. Then the phantom estimator is defined as

$$H_{Ph}^N := \frac{\frac{1}{N}\sum_{i=1}^N\sum_{j=0}^{\tau_{\theta,(i)}^0-1}\hat{\mathbf{D}}\left(P(\theta),g;X_j^{(i)}\right)}{\frac{1}{N}\sum_{i=1}^N\tau_{\theta,(i)}^0}, \tag{6.15}$$

where $\tau^0_{\theta,(i)}$ is the length of the i-th regeneration cycle, $X^{(i)}_j$ is the j-th state in the i-th regeneration cycle and each of the estimator $\hat{\mathbf{D}}$ is generated independently. Table 2 records the performance of H^N_{Ph} with different values of N. This estimator is also asymptotically unbiased. In [15], the authors propose another implementation of the phantom estimator in the single server queue setting with a more complex dependence on the regeneration cycle and thus the performance is worse in the heavy traffic scenarios comparing to the implementation of the phantom estimator experimented here.

6.1.3 The Regeneration Estimator [10]

Assume the Markov chain X possesses a regenerative structure and thus the stationary mean can be written as

$$\gamma(\theta) = \mathbb{E}[f(X_\infty)] = \frac{u(\theta)}{l(\theta)}, \tag{6.16}$$

where

$$u(\theta) := \mathbb{E}^\theta_s\Big[\sum_{k=1}^{\tau_\theta} f(X_k)\Big]; \tag{6.17}$$

$$l(\theta) := \mathbb{E}^\theta_s[\tau_\theta] \tag{6.18}$$

for some regenerative state s and regeneration time $\tau_\theta := \inf\{n \geq 1 : X_n = s\}$. This paper shows that under appropriate conditions,

$$\gamma'(\theta) = \frac{u'(\theta)l(\theta) - l'(\theta)u(\theta)}{l(\theta)^2} \tag{6.19}$$

where

$$u'(\theta) := \mathbb{E}^\theta_s\left[\left(\sum_{k=1}^{\tau_\theta} p'(\theta, X_{k-1}, X_k)\right)\left(\sum_{k=1}^{\tau_\theta} f(X_k)\right)\right]; \tag{6.20}$$

$$l'(\theta) := \mathbb{E}^\theta_s\left[\left(\sum_{k=1}^{\tau_\theta} p'(\theta, X_{k-1}, X_k)\right)\tau_\theta\right], \tag{6.21}$$

where p' is defined as in Assumption 1. Thus (6.19) is a nonlinear function of four expectations, each of which can be estimated by generating N regenerative cycles and averaging the corresponding quantities. Now we define the estimators for each of those terms:

$$\begin{aligned}
\hat{u}_N(\theta) &:= \frac{1}{N}\sum_{i=1}^N \left(\sum_{k=1}^{\tau^0_{\theta,(i)}} f(X^{(i)}_k)\right); \\
\hat{u'}_N(\theta) &:= \frac{1}{N}\sum_{i=1}^N \left(\sum_{k=1}^{\tau^0_{\theta,(i)}} p'(\theta, X^{(i)}_{k-1}, X^{(i)}_k)\right)\left(\sum_{k=1}^{\tau^0_{\theta,(i)}} f(X^{(i)}_k)\right); \\
\hat{l}_N(\theta) &:= \frac{1}{N}\sum_{i=1}^N \tau^0_{\theta,(i)}; \\
\hat{l'}_N(\theta) &:= \frac{1}{N}\sum_{i=1}^N \left(\sum_{k=1}^{\tau^0_{\theta,(i)}} p'(\theta, X^{(i)}_{k-1}, X^{(i)}_k)\right)\tau^0_{\theta,(i)},
\end{aligned} \tag{6.22}$$

where $\tau^0_{\theta,(i)}$ is the length of the i-th regeneration cycle with the regeneration state being 0, $X^{(i)}_k$ is the k-th state in the i-th regeneration cycle. Thus the final estimator is of the form

$$H^N_{Regen} := \frac{\hat{u'}_N(\theta)\hat{l}_N(\theta) - \hat{l'}_N(\theta)\hat{u}_N(\theta)}{\hat{l}_N(\theta)^2}. \tag{6.23}$$

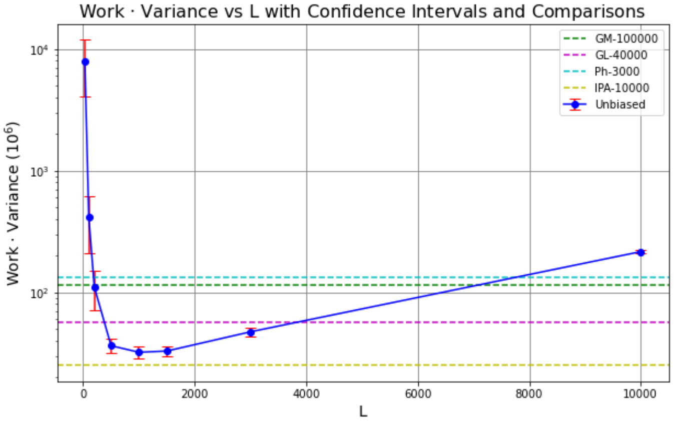


Figure 1: Comparison with previous methods.

The estimator is biased but consistent as $N \to \infty$. The performance is recorded in Table 3

6.1.4 The Likelihood Ratio Estimator [12]

This work proposes two likelihood ratio based estimators that do not depend on the availability of the regeneration structure. We implement the one with a better asymptotic variance [12]. Assume the stationary mean $\gamma(\theta)$ is known, then the proposed estimator takes the form

$$H_{LR}^N := \frac{1}{N} \sum_{k=1}^{N-1} \sum_{l=k}^{N-1} \left(f(X_l) - \gamma(\theta_0)\right) p'(\theta_0, X_{k-1}, X_k), \tag{6.24}$$

where p' is defined as in Assumption 1. The performance is is recorded in Table 4. Note that unlike the previous estimators, as N gets larger, the variance of H_{GO}^N does not go to 0. Thus the work-variance product grows without bound as we try make the bias go to 0.

6.1.5 Our estimator $H_4^{k,m,L}(X,Y)$ and comparison to others

The choice of of the parameters k and L are $k = L$ and $m = 100$. Table 5 records the performance of our estimator with different picks of L values. Notice that the inefficiency reduced very fast when L increases from small values and increases rather slowly when L is taken larger than the sweet spot. This also confirms the parameter choice guideline that we would rather choosing a larger L than necessary than choosing a smaller value. Figure 1 plots the work-variances of our unbiased derivative estimator with different choices of L and illustrate the performance comparison with previously discussed methods. The candidates from the previous methods are chosen such that the bias is negligible. In this figure, "IPA-10000" represents the performance of H_{IPA}^{10000} and the meaning of other indexes follow the same logic.

6.2 Multi-urn Ehrenfest model

For the second illustration, we introduce a multi-urn Ehrenfest model which is used to model the transitions of n gas particles in N interconnected urns. The model is described as the following. Let $X_t = \{x_{t,i}\}_{i=1}^N$ denote the state of the system at time t where $x_{t,i}$ describes the number of particles in urn i at time t. At each time step, one particle i is uniformly chosen from the population and

$$\begin{aligned} p_{ij} &= \frac{\theta_i}{n} \quad \text{for } i \neq j, \\ p_{ii} &= 1 - \frac{\theta_i(n-1)}{n} \end{aligned} \tag{6.25}$$

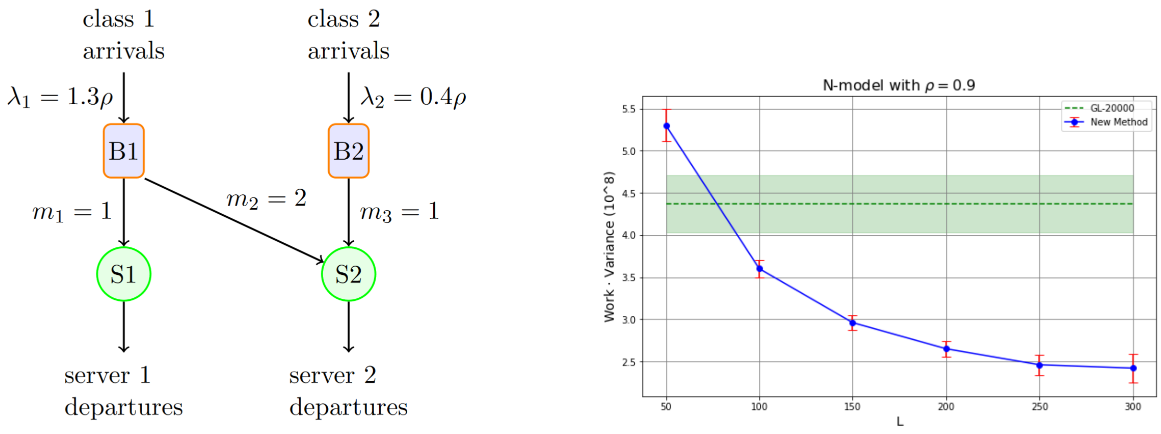
where $\theta = \{\theta_i\}_{i=1}^n$ dictates the rate at which a particle in urn i escapes and p_{ij} is the probability that the particle in urn i escapes to urn j. The diffusion rate θ_i associated with different urn is typically controlled by external forces such as extra heat sources or magnetic fields. The goal can vary depending on the application. Here we consider a function associated with state X_t as follows:

$$f(X_t) = \sum_{i=1}^{N} \mathbb{1}\{x_{t,i} \geq C_i\} \tag{6.26}$$

for some thresholds C_i associated with urn i. Function f gives the number of urns in which the number of particles reaches a desired threshold and the long run average $\gamma(\theta) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t)$ gives the expected number of urns that will reach the threshold under stationary distribution. For different threshold configurations, the optimal solutions will differ and it would be useful to estimate the derivative $\gamma'(\theta)$ to perform numerical optimization.

6.3 Two-class queuing network control

Next we present a parallel server queuing network control problem as a Markov decision process where the policy is parametrized by a neural network. Figure 2a shows the queuing network called the N-model with two classes of jobs whose arriving in the system with their inter-arrival time being exponentially distributed and with rates $\lambda_1 = 1.3\rho$ and $\lambda_2 = 0.4\rho$ where $\rho = 0.9$ is the traffic intensity parameter. Buffer 1 and Buffer 2 are buffers that hold class 1 and class 2 jobs respectively. Server 1 and Server 2 are two servers in the system. Server 1 can process only class 1 jobs with a exponential service rate $m_1 = 1$ and Server 2 can process class 1 jobs with a exponential service rate $m_2 = 2$ and can process class 2 jobs with a exponential service rate $m_3 = 1$. The state of the Markov decision process is a two dimensional vector (x_1, x_2) where x_1 is the number of jobs in Buffer 1 and x_2 is the number of jobs in Buffer 2. The holding cost of a state (x_1, x_2) is defined as $h((x_1, x_2)) := 10x_1 + x_2$. Let $\pi_{\boldsymbol{\theta}}$ denote the neural network parametrized policy that maps the state (x_1, x_2) to a probability distribution over a binary action space where $\boldsymbol{\theta}$ is the weights and biases in the neural network. An action $a \in \{1, 2\}$ is then sampled from the distribution $\pi_{\boldsymbol{\theta}}(a|x_1, x_2)$ whose gradient $\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(a|x_1, x_2)$ can be numerically computed by back-propagation. If $a = 1$, then Server 2 would preemptive class 1 jobs and if $a = 2$, Server 2 would preemptive class 2 jobs. Since actions are finite, we can compute the exact transition probabilities and their gradients by conditioning on the actions we take and summing them up. Check Section 5.3 in [4] for the full transition dynamics of the discrete time Markov chain model of this queuing network. In the reinforcement learning setting, we want to minimize the stationary mean holding cost of the network denoted as $\gamma(\boldsymbol{\theta})$. Thus the goal here is to estimate the gradient of the weights and biases of the policy neural network with respect to the stationary mean holding cost incurred by the system, i.e. $\nabla_{\boldsymbol{\theta}}\gamma(\boldsymbol{\theta})$. Here we implement our estimator and the Glynn-L'ecuyer regeneration estimator as they are the only eligible options. To the best of our knowledge, we do not know how to derive an IPA estimator or perform a Jordan decomposition on non-trivial neural network parametrized transition kernels. Furthermore, we do not have exact knowledge of the stationary mean so the likelihood ratio based method in [12] is also not applicable here. The entire gradient can be estimated but for the purpose of comparison and visualization, we only report the estimation of the partial derivative with respect to the last bias term of the neural network. Figure 2b presents the performance comparison between our method and the regeneration method and it shows that at optimal choice of L, the work-variance product decreases almost 50% and for a wide range of choices for L, the performance of our estimator triumphs the regeneration estimator.



(a) N-model network.

(b) Comparison with previous methods.

Figure 2: N-model and its comparison with previous methods.

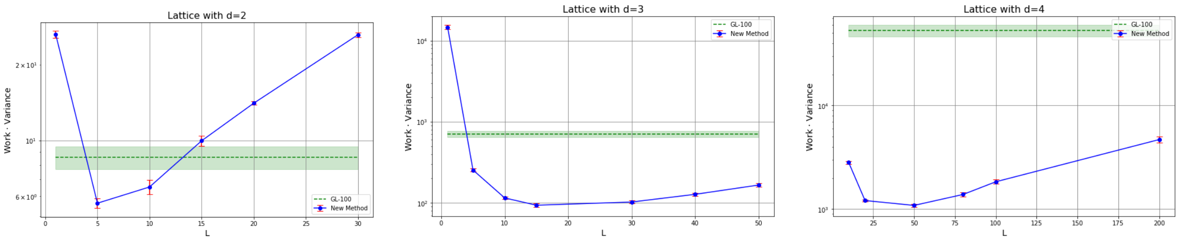
6.4 Ising model Control

In the $M/M/1$ example, we show that our estimator outperforms the almost all previous methods except for the IPA estimator for a wide range choice of the skipping parameter L. In the 2-class queuing network control problem, our estimator also exhibit the similar trait comparing to the Glynn-L'ecuyer regeneration based likelihood ratio estimator when the parametrization is by neural networks. In this section, we consider a non-queuing setting where the regeneration state becomes less accessible as dimension of the problem scales.

There has been some recent research that aims to maximize influence in social networks that are modeled by ising models [21][19][20]. Ising models can represents general social networks with different interaction frameworks but here for the illustrative purposes we adapt the simple 2-dimensional $d \times d$ square lattice as our example where each node only connects to its adjacent neighbors. There is in total d^2 nodes and node on the i-th row and j-th column is indexed as (i, j). Each node (i, j) is assigned a binary value $\sigma_{(i,j)} \in \{-1, 1\}$. In the social network setting, σ_i is the current opinion of individual i such as being prone to republican or democratic party and each node holds some influence to its neighbors. The control element of this model is introduced by a external magnetic field. We represent the external field by a neural network $h^{\boldsymbol{\theta}}$ that maps the node index (i, j) to a real number where $\boldsymbol{\theta}$ denotes the weights and biases of the neural network network. This external magnetic field is usually used to model the effect of a campaign or advertisement over the entire population's opinion formation. To introduce the time element, we let $\sigma_{(i,j)}(t)$ to be the opinion of individual (i, j) at time step t and let the vector $\boldsymbol{\sigma}(t) = \{\sigma_{(i,j)}(t)\}_{i,j\in\{1,2,\ldots,d\}^2}$ denote the current opinions of the entire network at time step t. The goal in this case is to maximize the total opinion in the long run or the stationary expected opinion, $\gamma(\boldsymbol{\theta}) := \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} M(\boldsymbol{\sigma}(t))$ where $M(\boldsymbol{\sigma}(t)) = \sum_{i,j\in\{1,2,\ldots,d\}^2} \sigma_{(i,j)}(t)$ is the total opinion or the magnetization of the lattice at time t. The transition of the system follows the Glauber dynamics [21]: at each time t, we sample an individual (i, j) uniformly from the entire population and update its state according to

$$P\left(\sigma_{(i,j)}(t+1) = 1 \mid \sigma(t)\right) = \frac{e^{\beta\left(\sum_{(k,l)\in\mathcal{N}_{(i,j)}} \sigma_{(k,l)}(t) + h^{\boldsymbol{\theta}}(i,j)\right)}}{e^{-\beta\left(\sum_{(k,l)\in\mathcal{N}_{(i,j)}} \sigma_{(k,l)}(t) + h^{\boldsymbol{\theta}}(i,j)\right)} + e^{\beta\left(\sum_{(k,l)\in\mathcal{N}_{(i,j)}} \sigma_{(k,l)}(t) + h^{\boldsymbol{\theta}}(i,j)\right)}}, \tag{6.27}$$

where $\mathcal{N}_{(i,j)}$ is the set of nodes adjacent to node (i, j) and β is the inverse temperature which reflects the overall interaction strength. The most intuitive way to construct a coupling kernel is by sampling the individuals with the same indices in both lattices and sampling their opinions following the maximum coupling distribution. Our gradient estimator can be of help when one is trying to estimate $\nabla_{\boldsymbol{\theta}}\gamma(\boldsymbol{\theta})$ during numerical optimization of the control neural network to target correct individual to drop advertisement to maximize influence under certain budget constraints. Again, as in the queuing network control example, the entire gradient can be estimated but we only report the estimation of the partial derivative with respect to the last bias term of the neural network. We test our estimator and the Glynn-L'ecuyer regeneration estimator on three cases where the dimension of the lattice $d = 2, 3, 4$ respectively. We report the performance of the Glynn-L'ecuyer estimator where the regeneration state is set to be the state where $sigma_{(i,j)} = 1$ for all indexes (i, j). Other choices of regeneration states (since any state can be a regeneration state in this setting) have been explored and non performs better. Figure 3 shows the work-variance products of our new estimator under different skipping parameter L comparing to the Glynn-L'ecuyer regeneration method for $d = 2, 3, 4$ respectively. It can be seen and is expected that as d increases, our method triumph the regeneration method by a larger and larger margin. The main reason is that regeneration becomes very rare to happen and thus causes the work-variance to increase significantly. For larger social networks, the difference between the two estimators becomes qualitative.



(a) 2-D Lattice

(b) 3-D Lattice

(c) 4-D Lattice

Figure 3: Work-Variance analysis for 2-D, 3-D, and 4-D Ising lattices. Each subfigure represents the corresponding lattice dimension's Work-Variance vs. L plot with confidence intervals.

7 Proofs

7.1 Proof of Theorem 2

Proof. First see that for any $i \geq 0$,

$$\gamma'(\theta_0) = \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[p'(\theta_0, X_i, X_{i+1})g(X_{i+1})\right], \tag{7.1}$$

since the chain starts from stationarity. Next we define some notations. Define the fundamental solution to the L-skeleton chain Poisson's equation to be

$$g^{fu,L}(x) := \mathbb{E}^{\theta_0}_{x}\left[\sum_{j=0}^{\infty}\bar{f}(X_{jL})\right], \tag{7.2}$$

where $\bar{f}(x) = f(x) - \pi(\theta_0)f$. Then we utilize (7.1) to obtain the following equality,

$$\begin{aligned} L\gamma'(\theta_0) &= \sum_{i=0}^{L-1}\mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_i, X_{i+1})g^{fu,1}(X_{i+1})\right] \\ &= \mathbb{E}_{\pi(\theta_0)}\left[\sum_{i=0}^{L-1}\left(p'(\theta_0, X_i, X_{i+1})Lg^{fu,L}(X_L)\right)\right] \\ &\quad + \sum_{i=0}^{L-1}\mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_i, X_{i+1})\left(g^{fu,1}(X_{i+1}) - Lg^{fu,L}(X_L)\right)\right]. \end{aligned} \tag{7.3}$$

The only thing remains to show is that the second term equals to 0. Again since the chain starts from stationarity,

$$\mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_i, X_{i+1})\left(g^{fu,1}(X_{i+1}) - Lg^{fu,L}(X_L)\right)\right] = \mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_0, X_1)\left(g^{fu,1}(X_1) - Lg^{fu,L}(X_{L-i})\right)\right]. \tag{7.4}$$

Therefore,

$$\begin{aligned}
&\sum_{i=0}^{L-1} \mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_i, X_{i+1})\left(g^{fu,1}(X_{i+1}) - Lg^{fu,L}(X_L)\right)\right] \\
&= \mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_0, X_1)\left(Lg^{fu,1}(X_1) - \sum_{i=0}^{L-1} Lg^{fu,L}(X_{L-i})\right)\right] \\
&= \mathbb{E}_{\pi(\theta_0)}\left[p'(\theta_0, X_0, X_1) L \mathbb{E}^{\theta_0}_{X_1 = X_1}\left[\left(g^{fu,1}(X_1) - \sum_{i=0}^{L-1} g^{fu,L}(X_{L-i})\right)\right]\right] \\
&= 0
\end{aligned} \tag{7.5}$$

where the first equality is due to (7.4), the second equality is from the tower property by conditioning on X_1 and the final equality holds by realizing the following equality:

$$\mathbb{E}^{\theta_0}_x\left[\sum_{i=0}^{L-1} g^{fu,L}(X_i)\right] = \mathbb{E}^{\theta_0}_x\left[\sum_{i=0}^{L-1}\sum_{j=0}^{\infty} \bar{f}(X_{i+jL})\right] = \mathbb{E}^{\theta_0}_x\left[\sum_{j=0}^{\infty} \bar{f}(X_j)\right] = g^{fu,1}(x). \tag{7.6}$$

Hence it completes the proof. □

7.2 Proof of Theorem 3

Proof. Applying Hölder's inequality and get

$$\Gamma^L_\zeta(x) \le \mathbb{E}^{\theta_0}_x\left[G^L_z(X_L)^{2\zeta p}\right]^{\frac{1}{p}} \mathbb{E}^{\theta_0}_x\left[\left(\sum_{i=0}^{L-1} p'(\theta_0, X_i, X_{i+1})\right)^{\frac{2\zeta p}{p-1}}\right]^{\frac{p-1}{p}}. \tag{7.7}$$

First we obtain an upper bound for the second term of the right-hand side. Recall that we defined

$$M_n := \sum_{i=0}^{L-1} p'(\theta_0, x_i, x_{i+1}) \tag{7.8}$$

with $M_0 = 0$. We now show that $\{M_n\}_{n=1,2,\ldots}$ is a martingale. For fixed states x, y and for all $h \le \epsilon$,

$$\frac{|p(\theta_0 + h, x, y) - p(\theta_0, x, y)|}{h} \le \omega_\epsilon(x, y) \tag{7.9}$$

and

$$\int_{\mathcal{X}} \omega_\epsilon(x, y) P(\theta_0, x, dy) \le V(x)\,|\Omega(x)|_V < \infty, \tag{7.10}$$

where

$$\Omega(x) = \int_{\mathcal{X}} \omega_\epsilon(x, y) P(\theta_0, x, dy), \tag{7.11}$$

and the second inequality holds since $|\Omega|_V < \infty$ is implied by $\left|\Omega^{2q}\right|_V < \infty$. Thus

$$\int_{\mathcal{X}} p'(\theta_0, x, y) P(\theta_0, x, dy) = \int_{\mathcal{X}} \lim_{h\to 0} \frac{p(\theta_0 + h, x, y) - p(\theta_0, x, y)}{h} P(\theta_0, x, dy) \tag{7.12}$$

$$= \lim_{h\to 0} \int_{\mathcal{X}} \frac{p(\theta_0 + h, x, y) - p(\theta_0, x, y)}{h} P(\theta_0, x, dy) \tag{7.13}$$

$$= \lim_{h\to 0} \frac{1}{h} \left(\int_{\mathcal{X}} p(\theta_0 + h, x, y) P(\theta_0, x, dy) - \int_{\mathcal{X}} p(\theta_0, x, y) P(\theta_0, x, dy) \right) \tag{7.14}$$

$$= 0, \tag{7.15}$$

where the second equality follows from the dominated convergence theorem and the last equality follows from

$$\int_{\mathcal{X}} p(\theta, x, y) P(\theta_0, x, dy) = \int_{\mathcal{X}} P(\theta, x, dy) = 1, \tag{7.16}$$

for all θ. Then

$$\mathbb{E}_x^{\theta_0}\left[p'(\theta_0, X_i, X_{i+1})\right] = \mathbb{E}_x^{\theta_0}\left[\mathbb{E}^{\theta_0}\left[p'(\theta_0, X_i, X_{i+1}) | X_i\right]\right] = 0, \tag{7.17}$$

and thus M_n is a martingale. Furthermore, define

$$\gamma_{\frac{2\zeta p}{p-1}, n} := \mathbb{E}_x^{\theta_0}\left[|p'(\theta_0, X_n, X_{n+1})|^{\frac{2\zeta p}{p-1}}\right], \tag{7.18}$$

and observe that

$$\gamma_{\frac{2\zeta p}{p-1}, n} = \mathbb{E}_x^{\theta_0}\left[\mathbb{E}^{\theta_0}\left[|p'(\theta_0, X_n, X_{n+1})|^{\frac{2\zeta p}{p-1}} | X_n\right]\right] \tag{7.19}$$

$$\leq \mathbb{E}_x^{\theta_0}\left[\mathbb{E}^{\theta_0}\left[\omega_\epsilon(X_n, X_{n+1})^{\frac{2\zeta p}{p-1}} | X_n\right]\right] \tag{7.20}$$

$$\leq |\Omega^{\frac{2\zeta p}{p-1}}|_V \mathbb{E}_x^{\theta_0}\left[V(X_n)\right] \tag{7.21}$$

$$\leq |\Omega^{\frac{2\zeta p}{p-1}}|_V \left(V(x) + \frac{b}{1-\lambda}\right) \tag{7.22}$$

$$= |\Omega^{\frac{2\zeta p}{p-1}}|_V \left(1 + \frac{b}{1-\lambda}\right) V(x), \tag{7.23}$$

where the first inequality holds because (7.9); the second inequality holds due to Assumption 5 with $\zeta \leq \kappa$; and the third inequality holds since (2.5). Thus by Theorem 1 in [5], we obtain an upper bound on the second term of (7.7):

$$\mathbb{E}\left[|M_L|^{\frac{2\zeta p}{p-1}}\right]^{\frac{p-1}{p}} \leq L^\zeta \left(C_{\frac{2\zeta p}{p-1}} |\Omega^{\frac{2\zeta p}{p-1}}|_V \left(1 + \frac{b}{1-\lambda}\right) V(x)\right)^{\frac{p-1}{p}} \tag{7.24}$$

where $C_l := \left[8(l-1)\max(1, 2^{l-3})\right]^l$. Now we look at the first term in (7.7). We define $(X_{L+t}, Z_t)_{t\in\mathbb{N}}$ to be a coupled Markov chains that evolves according to $\bar{P}(\theta_0)$ with $X_L \sim P^L(\theta_0, x, \cdot)$ and $Z_0 = z$.

Then see that

$$\begin{aligned}
\mathbb{E}_x\left[G_z^L(X_L)^{2\zeta p}\right]^{\frac{1}{2\zeta p}} &= \mathbb{E}_{x,z}^{\theta_0}\left[\left(\sum_{t=0}^{\infty}\left(f(X_{(t+1)L}) - f(Z_{tL})\mathbb{1}\left(X_{(t+1)L} \neq Z_{tL}\right)\right)^{2\zeta p}\right)\right]^{\frac{1}{2\zeta p}} \\
&\leq \sum_{t=0}^{\infty}\Big(\mathbb{E}_{x,z}^{\theta_0}\left[f(X_{(t+1)L})^{2\zeta p}\mathbb{1}\left(X_{(t+1)L} \neq Z_{tL}\right)\right]^{\frac{1}{2\zeta p}} \\
&\quad + \mathbb{E}_{x,z}^{\theta_0}\left[f(Z_{tL})^{2\zeta p}\mathbb{1}\left(X_{(t+1)L} \neq Z_{tL}\right)\right]^{\frac{1}{2\zeta p}}\Big) \\
&\leq \sum_{t=0}^{\infty}\Big(\mathbb{E}_{x,z}^{\theta_0}\left[f(X_{(t+1)L})^{2\zeta p+\delta}\right]^{\frac{1}{2\zeta p+\delta}} \\
&\quad + \mathbb{E}_{x,z}^{\theta_0}\left[f(Z_{tL})^{2\zeta p+\delta}\right]^{\frac{1}{2\zeta p+\delta}}\Big) P_x\left(\tau_{X_L,z} > tL\right)^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}} \\
&\leq \left|f^{2\zeta p+\delta}\right|_V^{\frac{1}{2\zeta p+\delta}}\left(\bar{V}(x)^{\frac{1}{2\zeta p+\delta}} + \bar{V}(z)^{\frac{1}{2\zeta p+\delta}}\right)\sum_{t=0}^{\infty}P_x\left(\tau_{X_L,z} > tL\right)^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}} \\
&\leq \left|f^{2\zeta p+\delta}\right|_V^{\frac{1}{2\zeta p+\delta}}\left(\bar{V}(x)^{\frac{1}{2\zeta p+\delta}} + \bar{V}(z)^{\frac{1}{2\zeta p+\delta}}\right) \\
&\quad \cdot\left(1 + \left(M\bar{V}(x) + MV(z)\right)^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\frac{\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}{1-\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}\right)
\end{aligned} \tag{7.25}$$

where the first inequality is due to Minkowski's inequality; the second inequality is from Hölder's inequality; the third inequality holds because Assumption 5 and 2, specifically

$$\mathbb{E}_{x,z}^{\theta_0}\left[f(X_{tL})^{2\zeta p+\delta}\right] \leq \left|f^{2\zeta p+\delta}\right|_V^{\frac{1}{2\zeta p+\delta}}\mathbb{E}_{x,z}^{\theta_0}\left[V(X_{tL})\right] \leq \left|f^{2\zeta p+\delta}\right|_V^{\frac{1}{2\zeta p+\delta}}\bar{V}(x); \tag{7.26}$$

and the forth inequality holds if we apply Assumption 4 and obtain

$$P_x\left(\tau_{X_L,z} > tL\right) = \int_{\mathcal{X}} P\left(\tau_{y,z} > tL\right) P^L(\theta_0, x, dy) \tag{7.27}$$

$$\leq \int_{\mathcal{X}} M\left(V(y) + V(z)\right)\rho^{tL}P^L(\theta_0, x, dy) \tag{7.28}$$

$$\leq M\left(\bar{V}(x) + V(z)\right)\rho^{tL}. \tag{7.29}$$

Therefore by raising both sides to the power 2ζ, we obtain

$$\begin{aligned}
\mathbb{E}_x\left[G_z^L(X_L)^{2\zeta p}\right]^{\frac{1}{p}} &\leq 2^{4\zeta-2}\left|f^{2\zeta p+\delta}\right|_V^{\frac{2\zeta}{2\zeta p+\delta}}\left(\bar{V}(x)^{\frac{2\zeta}{2\zeta p+\delta}} + \bar{V}(z)^{\frac{2\zeta}{2\zeta p+\delta}}\right) \\
&\quad \cdot\left(1 + M^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\left(\bar{V}(x)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}} + V(z)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\right)\left(\frac{\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}{1-\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}\right)^{2\zeta}\right).
\end{aligned} \tag{7.30}$$

where we use the inequality $(x+y)^{2\zeta} \leq 2^{2\zeta-1}(x^{2\zeta}+y^{2\zeta})$. To emphasize the dependence of the starting state x, we rewrite the upper bound to be

$$\begin{aligned}\mathbb{E}_x\left[G_z^L(X_L)^{2\zeta p}\right]^{\frac{1}{p}} \leq V(x)^{\frac{1}{p}} 2^{4\zeta-2}\left|f^{2\zeta p+\delta}\right|_V^{\frac{2\zeta}{2\zeta p+\delta}}\left(1+\left(\frac{b}{1-\lambda}\right)^{\frac{2\zeta}{2\zeta p+\delta}}+\bar{V}(z)^{\frac{2\zeta}{2\zeta p+\delta}}\right)\\ \cdot\left(1+M^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\left(1+\left(\frac{b}{1-\lambda}\right)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}+V(z)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\right)\left(\frac{\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}{1-\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}\right)^{2\zeta}\right).\end{aligned} \tag{7.31}$$

Thus combining it with the bound on the martingale term (7.24), we obtain that

$$\Gamma_\zeta^L(x) \leq V(x)\left(L^\zeta A_z + L^\zeta\left(\frac{\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}{1-\left(\rho^{\frac{\delta}{2\zeta p(2\zeta p+\delta)}}\right)^L}\right)^{2\zeta} B_z\right) := V(x)U_{\zeta,z}(L) \tag{7.32}$$

where

$$A_z^\zeta := 2^{4\zeta-2}\left|f^{2\zeta p+\delta}\right|_V^{\frac{2\zeta}{2\zeta p+\delta}}\left(C_{\frac{2\zeta p}{p-1}}|\Omega^{\frac{2\zeta p}{p-1}}|_V\left(1+\frac{b}{1-\lambda}\right)\right)^{\frac{p-1}{p}}\left(1+\left(\frac{b}{1-\lambda}\right)^{\frac{2\zeta}{2\zeta p+\delta}}+\bar{V}(z)^{\frac{2\zeta}{2\zeta p+\delta}}\right); \tag{7.33}$$

$$B_z^\zeta := A_z^\zeta\cdot M^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\left(1+\left(\frac{b}{1-\lambda}\right)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}+V(z)^{\frac{2\zeta\delta}{2\zeta p(2\zeta p+\delta)}}\right) \tag{7.34}$$

are independent of the starting state x or L. Therefore,

$$\frac{\Gamma_\zeta^L(x)}{V(x)} \leq U_{\zeta,z}(L) \tag{7.35}$$

for all $x \in \mathcal{X}$ and hence $\left|\Gamma_\zeta^L\right|_V \leq U_{\zeta,z}(L) < \infty$. □

7.3 Proof of Theorem 4

Proof. First note that

$$\begin{aligned}\mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{X_i\}_{i=k+tL}^{k+(t+1)L}\right)^2\mathbb{1}\left(X_{k+tL}\neq Y_{k+(t-1)L}\right)\right]^{\frac{1}{2}} &\leq \mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{X_i\}_{i=k+tL}^{k+(t+1)L}\right)^{2\kappa}\right]^{\frac{1}{2\kappa}} P\left(\tau^L > k+tL\right)^{\frac{\kappa-1}{2\kappa}}\\ &= \mathbb{E}_{\mu_0}^{\theta_0}\left[\Gamma_\kappa^L(X_{k+tL})\right]^{\frac{1}{2\kappa}} P\left(\tau^L > k+tL\right)^{\frac{\kappa-1}{2\kappa}}\\ &\leq \left|\Gamma_\kappa^L\right|_V^{\frac{1}{2\kappa}}\mathbb{E}_{\mu_0}^{\theta_0}\left[V(X_{k+tL})\right]^{\frac{1}{2\kappa}} P\left(\tau^L > k+tL\right)^{\frac{\kappa-1}{2\kappa}}\\ &\leq \left|\Gamma_\kappa^L\right|_V^{\frac{1}{2\kappa}}\mu_0\bar{V}^{\frac{1}{2\kappa}} P\left(\tau^L > k+tL\right)^{\frac{\kappa-1}{2\kappa}},\end{aligned} \tag{7.36}$$

where the first inequality is due to Hölder's inequality; the second equality is by the tower property; the second inequality follows from Lemma 3 and the last inequality follows from (2.5). The argument holds similarly for chain Y. Next we bound the tail probability of the coupling time.

$$\begin{aligned}P_x\left(\tau^L > k+tL\right) &= \int_{\mathcal{X}}\int_{\mathcal{X}} P\left(\tau_{x,y} > k+(t-1)L\right)P^L(\theta_0,x,dy)\mu_0(dx)\\ &\leq \int_{\mathcal{X}}\int_{\mathcal{X}} M\left(V(x)+V(y)\right)\rho^{k+(t-1)L}P^L(\theta_0,x,dy)\mu_0(dx)\\ &\leq M\left(\mu_0 V+\mu_0\bar{V}\right)\rho^{k+(t-1)L},\end{aligned} \tag{7.37}$$

where the first inequality follows from (2.9). Now we are equipped to show that $H_2^{k,L}(X,Y)$ has finite second moment.

$$\begin{aligned}\mathbb{E}_{\mu_0}^{\theta_0}\left[H_2^{k,L}(X,Y)^2\right]^{\frac{1}{2}} &\leq \mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{X_i\}_{i=k}^{k+L}\right)^2\right]^{\frac{1}{2}} + \sum_{t=1}^{\infty}\Bigg(\mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{X_i\}_{i=k+tL}^{k+(t+1)L}\right)^2 \mathbb{1}\left(X_{k+tL} \neq Y_{k+(t-1)L}\right)\right]^{\frac{1}{2}} \\ &\qquad + \mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{Y_i\}_{i=k+(t-1)L}^{k+tL}\right)^2 \mathbb{1}\left(X_{k+tL} \neq Y_{k+(t-1)L}\right)\right]^{\frac{1}{2}}\Bigg) \\ &\leq \left|\Gamma_\kappa^L\right|_V^{\frac{1}{2\kappa}} \mu_0 \bar{V}^{\frac{1}{2\kappa}} + 2\left|\Gamma_\kappa^L\right|_V^{\frac{1}{2\kappa}} \mu_0 \bar{V}^{\frac{1}{2\kappa}} \sum_{t=1}^{\infty} P\left(\tau^L > k+tL\right)^{\frac{\kappa-1}{2\kappa}} \\ &\leq \left|\Gamma_\kappa^L\right|_V^{\frac{1}{2\kappa}} \mu_0 \bar{V}^{\frac{1}{2\kappa}} + 2\left|\Gamma_\kappa^L\right|_V^{\frac{1}{2\kappa}} \mu_0 \bar{V}^{\frac{1}{2\kappa}} M^{\frac{\kappa-1}{2\kappa}} \left(\mu_0 V + \mu_0 \bar{V}\right)^{\frac{\kappa-1}{2\kappa}} \left(\rho^{\frac{\kappa-1}{2\kappa}}\right)^k \sum_{t=0}^{\infty}\left(\rho^{\frac{\kappa-1}{2\kappa}}\right)^{tL},\end{aligned} \tag{7.38}$$

where the first inequality follows from Minkowski's inequality; the second inequality follows from (7.36) and the third inequality follows from (7.37). Also observe that the second term of the upper bound is actually an upper bound for $\mathbb{E}_{\mu_0}^{\theta_0}\left[\left(BC_k^L\right)^2\right]^{1/2}$. Since $\left(\rho^{\frac{\kappa-1}{2\kappa}}\right)^k$ goes to 0 as k increases to ∞ and the other terms are constants, we may conclude (3.29):

$$\mathbb{E}_{\mu_0}^{\theta_0}\left[\left(BC_k^L\right)^2\right] \to 0 \tag{7.39}$$

as $k \to \infty$. Furthermore, see that

$$\mathbb{E}_{\mu_0}^{\theta_0}\left[\left(SE_k^L\right)^2\right] = \mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{X_i\}_{i=k}^{k+L}\right)^2\right] = \mathbb{E}_{\mu_0}^{\theta_0}\left[\Gamma_{\frac{1}{2}}^L(X_k)\right] \tag{7.40}$$

and $\left|\Gamma_1^L\right|_V < \infty$ is implied by Lemma 3. Therefore,

$$\left|\mathbb{E}_{\mu_0}^{\theta_0}\left[\Gamma_1^L(X_k)\right] - \pi(\theta_0)\Gamma_1^L\right| \leq \left|\Gamma_1^L\right|_V M(\mu_0 V)\rho^k, \tag{7.41}$$

where the upper bound goes to 0 as k increases to infinity. Hence we showed (3.30). To prove $H_2^{k,L}(X,Y)$ is unbiased, first realize that

$$\mathbb{E}_{\mu_0}^{\theta_0}\left[H\left(\{X_i\}_{i=k}^{k+L}\right)\right] = \mathbb{E}_{\mu_0}^{\theta_0}\left[h(X_k)\right]. \tag{7.42}$$

Then using the fact that $H_2^{k,L}(X,Y)$ has finite second moment, we can guarantee that it has finite first moment and thus we may invoke the Fubini's theorem to guarantee the validity of the interchange of the summation and expectation from (2.10) to (2.13) and conclude that $H_2^{k,L}(X,Y)$ is indeed an unbiased estimator. The finite computation time is implied by the geometrically decaying tail probability of the coupling time made in Assumption 4. □

7.4 Proof of Lemma 5

Proof. First we bound the variance σ_L^2 by its second moment:

$$\sigma_L^2 \leq \pi(\theta_0)\Gamma_1^L \leq \pi(\theta_0)V|\Gamma_1^L|_V \tag{7.43}$$

Now we turn to obtain an upper bound for the covariance term. To ease the notations, we let

$$H^L(j) := H\left(\{X_i\}_{i=j}^{j+L}\right). \tag{7.44}$$

Observe that

$$\begin{aligned}
\mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)H^L(jL)\right] &= \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)\mathbb{E}^{\theta_0}_{X_L}\left[H^L(jL)\right]\right] \\
&= \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)\left(\gamma'(\theta_0)+\mathbb{E}^{\theta_0}_{X_L}\left[H^L(jL)\right]-\gamma'(\theta_0)\right)\right] \\
&= \gamma'(\theta_0)^2+\mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)\left(\mathbb{E}^{\theta_0}_{X_L}\left[H^L(jL)\right]-\gamma'(\theta_0)\right)\right] \\
&\le \gamma'(\theta_0)^2+\mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)^2\right]^{\frac{1}{2}}\mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[\left(\mathbb{E}^{\theta_0}_{X_L}\left[H^L(jL)\right]-\gamma'(\theta_0)\right)^2\right]^{\frac{1}{2}}.
\end{aligned} \tag{7.45}$$

Now note that since $j \ge 1$,

$$\mathbb{E}^{\theta_0}_x\left[H^L(jL)\right] = \mathbb{E}^{\theta_0}_x\left[\Gamma^L_{\frac{1}{2}}\left(X_{(j-1)L}\right)\right], \tag{7.46}$$

$$\gamma'(\theta_0) = \pi(\theta_0)\Gamma^L_{\frac{1}{2}}, \tag{7.47}$$

and by Jensen's inequality,

$$|\Gamma^L_{\frac{1}{2}}(x)|^2 \le |\Gamma^L_1(x)| \le \left|\Gamma^L_1\right| V(x). \tag{7.48}$$

Thus

$$\left|\mathbb{E}^{\theta_0}_x\left[H^L(jL)\right]-\gamma'(\theta_0)\right| \le \left|\Gamma^L_1\right|_V^{\frac{1}{2}} M\sqrt{V(x)}\rho^{(j-1)L} \tag{7.49}$$

Therefore,

$$\sigma_{j,L} = \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)H^L(jL)\right]-\gamma'(\theta_0)^2 \tag{7.50}$$

$$\le \mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)^2\right]^{\frac{1}{2}}\mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[\left(\mathbb{E}^{\theta_0}_{X_L}\left[H^L(jL)\right]-\gamma'(\theta_0)\right)^2\right]^{\frac{1}{2}} \tag{7.51}$$

$$\le \left(\pi(\theta_0)\Gamma^L_1\right)^{\frac{1}{2}}\left(\pi(\theta_0)V\right)^{\frac{1}{2}}\left|\Gamma^L_1\right|_V^{\frac{1}{2}} M\rho^{(j-1)L} \tag{7.52}$$

$$\le \left(\pi(\theta_0)V\right)\left|\Gamma^L_1\right|_V M\rho^{(j-1)L} \tag{7.53}$$

where the first inequality follows from (7.45) and the second inequality follows from realizing that $\mathbb{E}^{\theta_0}_{\pi(\theta_0)}\left[H^L(0)^2\right] = \pi(\theta_0)\Gamma^L_1$ and inequality (7.49). Thus we can derive an upper bound for the sum of the covariance terms

$$\sum_{j=1}^{\infty}\sigma_{j,L} \le \sigma_{1,L}+\sum_{j=2}^{\infty}\sigma_{j,L} \le \sigma^2_L+\left(\pi(\theta_0)V\right)\left|\Gamma^L_1\right|_V\frac{M\rho^L}{1-\rho^L} \tag{7.54}$$

Finally we combine the upper bounds for variance (7.43) and covariance terms (7.54) and invoke Lemma 3 to get the upper bound for the asymptotic variance:

$$\sigma^2_L+2\sum_{j=1}^{\infty}\sigma_{j,L} \le U_{1,z}(L)\pi(\theta_0)V\left(3+\frac{2M\rho^L}{1-\rho^L}\right) \tag{7.55}$$

□

7.5 Proof of Theorem 6

Proof. Note that

$$\frac{\frac{2M\rho^L}{1-\rho^L}}{\frac{2M\rho}{1-\rho}} \ge 1, \tag{7.56}$$

for all $L \geq 1$. Thus

$$\frac{W(1)}{W(L)} \geq \frac{\left(1 + 2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right) U_{1,z}(1)}{\left(L + 2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right) U_{1,z}(L)}, \tag{7.57}$$

and our goal is to develop a lower bound for the right hand side. Before doing so, we make obvious some useful inequalities. First see that

$$\begin{aligned} \mathbb{E}\left[\tau_{\pi(\theta_0),z}\right] - 1 &\leq \sum_{t=1}^{\infty} P\left(\tau_{\pi(\theta_0),z} > t\right) \\ &\leq \sum_{t=1}^{\infty} \left(P\left(\tau_{\pi(\theta_0),z} > t\right)\right)^{\frac{\delta}{2p(2p+\delta)}} \\ &\leq M^{\frac{\delta}{2p(2p+\delta)}} \left(\pi(\theta_0)V + V(z)\right)^{\frac{\delta}{2p(2p+\delta)}} \frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1 - \rho^{\frac{\delta}{2p(2p+\delta)}}}, \end{aligned} \tag{7.58}$$

where the last inequality follows from (4.4). Rearrange the inequality to get

$$\begin{aligned} \left(\frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1 - \rho^{\frac{\delta}{2p(2p+\delta)}}}\right)^{-2} &\leq \left(\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right] - 1\right)^{-2} M^{\frac{2\delta}{2p(2p+\delta)}} \left(\pi(\theta_0)V + V(z)\right)^{\frac{2\delta}{2p(2p+\delta)}} \\ &\leq \left(\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right] - 1\right)^{-2} M^{\frac{2\delta}{2p(2p+\delta)}} \left(\pi(\theta_0)V^{\frac{2\delta}{2p(2p+\delta)}} + V(z)^{\frac{2\delta}{2p(2p+\delta)}}\right) \\ &\leq \left(\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right] - 1\right)^{-2} M^{\frac{2\delta}{2p(2p+\delta)}} \left(\left(\frac{b}{1-\lambda}\right)^{\frac{2\delta}{2p(2p+\delta)}} + V(z)^{\frac{2\delta}{2p(2p+\delta)}}\right), \end{aligned} \tag{7.59}$$

where the last inequality follows from that for any $x \in \mathcal{X}$:

$$\pi(\theta_0)V = \lim_{n\to\infty} \mathbb{E}_x^{\theta_0}\left[V(X_n)\right] \leq \lim_{n\to\infty} \lambda^n V(x) + \frac{b}{1-\lambda} = \frac{b}{1-\lambda}. \tag{7.60}$$

Now recall that

$$\frac{A_z^1}{B_z^1} = M^{-\frac{2\delta}{2p(2p+\delta)}} \left(1 + \left(\frac{b}{1-\lambda}\right)^{\frac{2\delta}{2p(2p+\delta)}} + V(z)^{\frac{2\delta}{2p(2p+\delta)}}\right)^{-1}, \tag{7.61}$$

and we now have the following inequality:

$$\frac{A_z^1}{B_z^1} \left(\frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1 - \rho^{\frac{\delta}{2p(2p+\delta)}}}\right)^{-2} \leq \left(\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right] - 1\right)^{-2}. \tag{7.62}$$

Now we go back to developing the lower bound for the performance ratio:

$$\begin{aligned}
\frac{\left(1+2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right)U_{1,z}(1)}{\left(L+2\mathbb{E}[\tau_{\pi(\theta_0),z}]\right)U_{1,z}(L)} &= \frac{\left(2\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]+1\right)\left(A_z^1+\left(\frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1-\rho^{\frac{\delta}{2p(2p+\delta)}}}\right)^2 B_z^1\right)}{\left(2\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]+L\right)\left(LA_z^1+L\left(\frac{\left(\rho^{\frac{\delta}{2p(2p+\delta)}}\right)^L}{1-\left(\rho^{\frac{\delta}{2p(2p+\delta)}}\right)^L}\right)^2 B_z^1\right)} \\
&\geq \frac{2\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]\left(\frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1-\rho^{\frac{\delta}{2p(2p+\delta)}}}\right)^2 B_z^1}{\left(2\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]+L\right)\left(LA_z^1+\frac{1}{L}\left(\frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1-\rho^{\frac{\delta}{2p(2p+\delta)}}}\right)^2 B_z^1\right)} \qquad (7.63) \\
&= \frac{1}{\left(1+\frac{L}{2\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]}\right)\left(L\left(\frac{\rho^{\frac{\delta}{2p(2p+\delta)}}}{1-\rho^{\frac{\delta}{2p(2p+\delta)}}}\right)^{-2}\frac{A_z^1}{B_z^1}+\frac{1}{L}\right)} \\
&\geq \frac{1}{\left(1+\frac{L}{2\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]}\right)\left(L\frac{1}{\left(\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]-1\right)^2}+\frac{1}{L}\right)},
\end{aligned}$$

where the first inequality follows from the fact that $\frac{h^x}{1-h^x} <= \frac{1}{x}\frac{h}{1-h}$ for any $0 < h < 1$ and $x \geq 1$ and the last inequality follows from (7.62). Finally if we let $L = \mathbb{E}\left[\tau_{\pi(\theta_0),z}\right] - 1$, then we observe

$$\frac{W(1)}{W(\mathbb{E}\left[\tau_{\pi(\theta_0),z}\right]-1)} > \frac{1}{3}\mathbb{E}\left[\tau_{\pi(\theta_0),z}-1\right]. \qquad (7.64)$$

which completes the proof. □

8 Tables

N	Work	Work·Variance (10^6)	Estimate (True Value:-24.963)
100	$5,235 \pm 24$	66.90 ± 9.88	-37.14 ± 1.07
200	$10,404 \pm 28$	47.74 ± 9.43	-30.90 ± 0.53
500	$26,151 \pm 160$	33.60 ± 7.46	-27.28 ± 0.99
1000	$51,943 \pm 77$	30.15 ± 1.58	-26.18 ± 0.23
2000	$104,019 \pm 150$	27.54 ± 1.80	-25.47 ± 0.21
5000	$259,609 \pm 395$	25.98 ± 1.42	-25.00 ± 0.21
10000	$520,093 \pm 457$	25.27 ± 0.92	-24.97 ± 0.07
40000	$2,080,745 \pm 1147$	26.20 ± 0.83	-25.10 ± 0.09

Table 1: Performance stats of H_{IPA}^N with different N

References

[1] Yves F Atchadé and Pierre E Jacob. Unbiased markov chain monte carlo: what, why, and how. *arXiv preprint arXiv:2406.06851*, 2024. 2.2, 2.3, 3.3, 4

N	Work	Work·Variance (10^6)	Estimate (True Value:-24.963)
100	186071 ± 4646	38.17 ± 2.37	-18.16 ± 0.30
200	382259 ± 6929	63.03 ± 4.85	-21.15 ± 0.29
500	943558 ± 10339	88.74 ± 5.23	-22.92 ± 0.20
1000	1896955 ± 13497	112.83 ± 5.90	-23.97 ± 0.15
2000	3755492 ± 27625	115.11 ± 6.87	-24.38 ± 0.16
3000	5643306 ± 28646	132.32 ± 8.47	-24.56 ± 0.15
5000	$\pm$	$\pm$	$\pm$

Table 2: Performance stats of H_{Ph}^N with different N

N	Work (10^4)	Work ·Variance (10^6)	Estimate (True Value:-24.963)
1000	2.59 ± 0.001	15.33 ± 0.36	-18.40 ± 0.07
2000	5.19 ± 0.006	28.60 ± 2.25	-21.28 ± 0.18
3000	7.80 ± 0.005	34.86 ± 1.50	-22.38 ± 0.11
5000	12.97 ± 0.03	39.42 ± 2.01	-22.75 ± 0.42
10000	25.99 ± 0.02	45.23 ± 1.55	-24.02 ± 0.08
20000	52.02 ± 0.03	51.37 ± 2.88	-24.54 ± 0.14
40000	104.00 ± 0.05	57.30 ± 3.66	-24.86 ± 0.04

Table 3: Performance stats of H_{GL}^N with different N

[2] Sandjai Bhulai and Flora M Spieksma. On the uniqueness of solutions to the poisson equations for average cost markov chains with unbounded cost functions. *Mathematical Methods of Operations Research*, 58:221–236, 2003. 2.4

[3] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018. 5.2, 5.2

[4] Jim G Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022. 2.4, 6.3

[5] SW Dharmadhikari, V Fabian, and K Jogdeo. Bounds on the moments of martingales. *The Annals of Mathematical Statistics*, pages 1719–1723, 1968. 7.2

[6] Randal Douc, Pierre E Jacob, Anthony Lee, and Dootika Vats. Solving the poisson equation using coupled markov chains. *arXiv preprint arXiv:2206.05691*, 2022. 1, 2.2, 2.3, 2.4, 3.3

[7] Taoying Farenhorst-Yuan. *Efficient simulation algorithms for optimization of discrete event systems based on measure-valued differentation.* Number 467. Rozenberg Publishers, 2010. 6.1.2

[8] Paul Glasserman. Regenerative derivatives of regenerative sequences. *Advances in applied probability*, 25(1):116–139, 1993. 1, 6.1.1

[9] Peter W Glynn and Alex Infanger. Solution representations for poisson's equation, martingale structure, and the markov chain central limit theorem. *Stochastic Systems*, 14(1):47–68, 2024. 2.4

[10] Peter W Glynn and Pierre L'ecuyer. Likelihood ratio gradient estimation for stochastic recursions. *Advances in applied probability*, 27(4):1019–1053, 1995. 1, 6.1.3

[11] Peter W Glynn and Sean P Meyn. A liapounov bound for solutions of the poisson equation. *The Annals of Probability*, pages 916–931, 1996. 2.4

N	Work·Variance (10^6)	Estimate (True Value:-24.963)
2000	0.60 ± 0.009	-10.19 ± 0.03
5000	4.64 ± 0.07	-16.33 ± 0.05
10000	14.41 ± 0.32	-20.17 ± 0.08
20000	31.61 ± 1.28	-22.57 ± 0.12
50000	64.02 ± 3.32	-23.98 ± 0.21
100000	114.23 ± 3.90	-24.62 ± 0.18
500000	479.88 ± 28.82	-25.26 ± 0.38

Table 4: Performance stats of H_{GO}^N with different N

L	Work (10^4)	Work ·Variance (10^6)	Estimate (True Value:-24.963)
30	5.09 ± 0.01	7957.25 ± 3905.91	-25.11 ± 1.57
100	6.74 ± 0.01	412.25 ± 201.43	-24.77 ± 0.31
200	8.07 ± 0.02	110.36 ± 39.80	-25.06 ± 0.22
500	11.27 ± 0.05	36.40 ± 4.98	-24.85 ± 0.37
1000	16.46 ± 0.06	32.15 ± 3.59	-25.12 ± 0.45
1500	21.46 ± 0.05	32.91 ± 2.98	-24.65 ± 0.38
3000	36.65 ± 0.05	47.34 ± 3.66	-24.89 ± 0.38
10000	107.50 ± 0.03	215.92 ± 7.80	-24.85 ± 0.35

Table 5: Performance stats of H_{GL}^N with different N

[12] Peter W Glynn and Mariana Olvera-Cravioto. Likelihood ratio gradient estimation for steady-state parameters. *Stochastic Systems*, 9(2):83–100, 2019. 1, 6.1.4, 6.3

[13] Peter W Glynn and Chang-han Rhee. Exact estimation for markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014. 1, 2.3

[14] Peter W Glynn and Ward Whitt. The asymptotic efficiency of simulation estimators. *Operations research*, 40(3):505–520, 1992. 1

[15] Bernd Heidergott, Taoying Farenhorst-Yuan, and Felisa Vázquez-Abad. A perturbation analysis approach to phantom estimators for waiting times in the g/g/1 queue. *Discrete Event Dynamic Systems*, 20:249–273, 2010. 1, 6.1.2

[16] Bernd Heidergott, Arie Hordijk, and Heinz Weisshaupt. Measure-valued differentiation for stationary markov chains. *Mathematics of Operations Research*, 31(1):154–172, 2006. 1, 6.1.2, 6.1.2

[17] Pierre E Jacob, John O'Leary, and Yves 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, 2.2, 2.3, 3.3, 3.3

[18] Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002. 2.1

[19] Shihuan Liu, Lei Ying, and Srinivas Shakkottai. Influence maximization in social networks: An ising-model-based approach. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 570–576. IEEE, 2010. 6.4

[20] Christopher Lynn and Daniel Lee. Maximizing activity in ising networks via the tap approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 6.4

[21] Christopher Lynn and Daniel D Lee. Maximizing influence in an ising network: A mean-field optimal solution. *Advances in neural information processing systems*, 29, 2016. 6.4

[22] Esa Nummelin. A splitting technique for harris recurrent markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(4):309–318, 1978. 2.4

[23] Art B Owen. Statistically efficient thinning of a markov chain sampler. *Journal of Computational and Graphical Statistics*, 26(3):738–744, 2017. 3.3

[24] Chang-Han Rhee and Peter Glynn. Lyapunov conditions for differentiability of markov chain expectations: the absolutely continuous case. *arXiv preprint arXiv:1707.03870*, 2017. 1, 1, 2.2, 2.4, 1, 3.1

[25] Chang-han Rhee and Peter W Glynn. Unbiased estimation with square root convergence for sde models. *Operations Research*, 63(5):1026–1043, 2015. 1

[26] Francisco JR Ruiz, Michalis K Titsias, Taylan Cemgil, and Arnaud Doucet. Unbiased gradient estimation for variational auto-encoders using coupled markov chains. In *Uncertainty in Artificial Intelligence*, pages 707–717. PMLR, 2021. 2.3

[27] Paul Vanetti and Arnaud Doucet. Discussion of "unbiased markov chain monte carlo with couplings" by jacob et al. *Journal of the Royal Statistical Society Series B*, 82(3):592–593, 2020. 2.3