Applied Probability Trust (15 September 2023)

CONVERGENCE SPEED AND APPROXIMATION ACCURACY OF NUMERICAL MCMC

TIANGANG CUI,* *University of Sydney*

JING DONG,** *Columbia University*

AJAY JASRA,*** *King Abdullah University of Science and Technology*

XIN T. TONG,**** *National University of Singapore*

* Postal address: School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
* Email address: tiangang.cui@sydney.edu.au
** Postal address: Graduate School of Business, Columbia University, U.S.A.
** Email address: jing.dong@gsb.columbia.edu
*** Postal address: Applied Mathematics and Computational Science Program, Computer, Electrical, Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, KSA
*** Email address: ajay.jasra@kaust.edu.sa
**** Postal address: Department of Mathematics, National University of Singapore, Singapore
**** Email address: mattxin@nus.edu.sg

Abstract

When implementing Markov Chain Monte Carlo (MCMC) algorithms, perturbation caused by numerical errors is sometimes inevitable. This paper studies how the perturbation of MCMC affects the convergence speed and approximation accuracy. Our results show that when the original Markov chain converges to stationarity fast enough and the perturbed transition kernel is a good approximation to the original transition kernel, the corresponding perturbed sampler has fast convergence speed and high approximation accuracy as well. Our convergence analysis is conducted under either the Wasserstein metric or the χ^2 metric, both are widely used in the literature. The results can be extended to obtain non-asymptotic error bounds for MCMC estimators. We demonstrate how to apply our convergence and approximation results to the analysis of specific sampling algorithms, including Random walk Metropolis, Metropolis adjusted Langevin algorithm with perturbed target densities, and parallel tempering Monte Carlo with perturbed densities. Finally, we present some simple numerical examples to verify our theoretical claims.

Keywords: Bayesian inverse problems, Markov Chain Monte Carlo, Convergence Speed, Perturbation Analysis

2020 Mathematics Subject Classification: Primary 35R30
Secondary 65C40, 37A25

1. Introduction

Markov Chain Monte Carlo (MCMC) is one of the main sampling methods in Bayesian statistics. Given a target density π with respect to Lebsegue measure on $\mathbb{R}^d$, an MCMC algorithm often simulates a Markov chain $(X_n)_{n\geq 0}$ with transition kernel P, such that π is its corresponding invariant measure. Under some generic conditions, the distribution of X_n converges to π geometrically fast. This indicates the existence of some mixing time n_0, such that the distribution of X_n is close to π when $n > n_0$. In other words, given a test function $f : \mathbb{R}^d \to \mathbb{R}$, we can use the following approximation

$$\mathbb{E}f(X_n) \approx \mathbb{E}^\pi f(X) := \int f(x)\pi(x)dx. \tag{1}$$

In practice, this allows us to approximate the average of a test function $\mathbb{E}^\pi f(X)$ using the temporal average of the Markov chain:

$$F_n := \frac{1}{n}\sum_{i=1}^{n} f(X_{n_0+i}). \tag{2}$$

The efficiency of the approximation scheme (2) is largely determined by the convergence speed of the Markov Chain $(X_n)_{n\geq 0}$ to π or the mixing time n_0. In particular, it will take $O(n_0)$ iterations to produce an approximately independent sample. In this context, convergence analysis has been a key component in the MCMC literature (see, for example, Section 4.1 of [1] and [27]).

When implementing MCMC on complicated target densities, it is often the case that we can only simulate a perturbed Markov chain $(\widehat{X}_n)_{n\geq 0}$ with transition kernel $\widehat{P}$. This is mainly due to two reasons:

1. The transition kernel P cannot be simulated directly. For example, if $(X_n)_{n\geq 0}$ is described by a stochastic differential equation (SDE) evaluated at a countable set of time points, using numerical schemes like the Euler-Maruyama method will induce discretization errors.

2. We do not have direct access to $\pi(x)$ or its derivatives. This is quite common in Bayesian inverse problems [40], where the target density can be written as

$$\pi(x) \propto p_0(x)\exp\left(-\tfrac{1}{2}\|G(x)-y\|^2\right). \tag{3}$$

   In (3), p_0 is the prior density of the unknown parameter x, G describes the data generating process, and y is the observed data. In many cases, G is formulated through an involved partial differential equation, and we can only compute an approximation of it, $\widehat{G}$ [7, 21, 5]. The corresponding "numerical" density becomes

$$\widehat{\pi}(x) \propto p_0(x)\exp\left(-\tfrac{1}{2}\|\widehat{G}(x)-y\|^2\right). \tag{4}$$

   In other settings, we may have access to $\pi(x)$, but accurate evaluation of its gradient $\nabla\pi(x)$ may not be accessible since it often involves high dimensional adjoint models. If we want to use gradient based MCMC, e.g. Metropolis Adjusted Langevin Algorithm (MALA), we can only use an approximately correct proposal. However, the Metropolis-Hastings step can guarantee that the target density remains the same, i.e. $\widehat{\pi} = \pi$.

In the above-mentioned scenarios, we run an MCMC $(\widehat{X}_n)_{n\geq 0}$ with transition kernel $\widehat{P}$ and target density $\widehat{\pi}$, which is the invariant measure of $(\widehat{X}_n)_{n\geq 0}$. In both scenarios (1) and (2) listed above, we would like to approximate $\mathbb{E}^{\pi} f(X)$ using

$$\mathbb{E} f(\widehat{X}_{n_0}) \quad \text{or} \quad \widehat{F}_n = \frac{1}{n}\sum_{i=1}^{n} f(\widehat{X}_{n_0+i}). \tag{5}$$

There are two key questions to address when using estimators of the form (5). The first question is about the convergence speed of $(\widehat{X}_n)_{n\geq 0}$ towards its invariant measure $\widehat{\pi}$, which determines the efficiency of the estimators in (5). In particular, if we use D to denote some metric between two distributions and use ν to denote the distribution of $\widehat{X}_0$, we are interested in how fast $D(\nu\widehat{P}^n, \widehat{\pi})$ converges to zero. The second question is about approximation accuracy, which can be measured by either the distance between the two invariant measures, $D(\widehat{\pi}, \pi)$, or the distance between the distribution of $\widehat{X}_n$ and π, $D(\nu\widehat{P}^n, \pi)$.

For MCMC based on $(\widehat{X}_n)_{n\geq 0}$ to achieve fast convergence and high approximation accuracy, we need to impose the following two high-level conditions (these conditions will be made more precise in our subsequent development):

1. $\widehat{P}$ is a good approximation of P.

2. $(X_n)_{n\geq 0}$ converges to its invariant measure π fast enough.

Condition 1 is necessary, because if $\widehat{P}$ is not a good approximation of P, $\widehat{\pi}$ is unlikely to be close to π, and the convergence property of $(X_n)_{n\geq 0}$ will not be useful in inferring the convergence property of $(\widehat{X}_n)_{n\geq 0}$. Condition 2 is also necessary. Otherwise, the approximation error may grow with the number of iterations. For example, one can think of an unstable autoregressive sequence, where numerical errors often grow exponentially with the number of iterations. Since Condition 1 involves only the one-step transition kernels, it is easier to fulfill. Hence, it is reasonable to study Condition 2 first and then formulate a version of Condition 1 that is compatible with the corresponding Condition 2.

In the literature on Markov chains, convergence to stationarity is often studied using one of two frameworks. The main differences between the two frameworks are the metrics and analytical tools involved. The first type of metric is the Wasserstein metrics [31]. The associated convergence results are often termed "geometric

ergodicity". Establishing convergence under the Wasserstein metrics often involves finding an appropriate Lyapunov function V and constructing an appropriate coupling [28]. For simplicity, we will call this framework **"Wasserstein convergence"**. The second framework uses the χ^2 distance [3] (or KL-divergence as in [45]). Establishing convergence under the χ^2 distance often involves functional analysis or other partial differential equation (PDE) tools such as Poincaré inequality and log Sobolev inequality. For simplicity, we will refer to this framework as the **"χ^2 convergence"**.

Establishing Wasserstein convergence is often viewed as being more intuitive, as it involves standard Lyapunov function and coupling construction. Establishing χ^2 convergence can be more delicate, but it often provides tighter quantification, especially in high-dimensional settings. For example, for the unadjusted Langevin algorithm, [9] uses Wasserstein convergence and the analysis works only for a fixed dimension; [45, 11] use χ^2 convergence and the analysis works in high-dimensional settings. We also note that these two frameworks are related. In particular, on one hand, under suitable regularity conditions, geometric ergodicity leads to the existence of a spectral gap, and hence χ^2 convergence (see, e.g., Proposition 2.8 in [21], and [19] for a more complete discussion of the connection). On the other hand, under proper regularity conditions, convergence under the χ^2 distance leads to convergence under the total variation distance.

We discuss both frameworks in this paper, because for some Markov chains, we may only have knowledge of one form of convergence. For example, to the best of our knowledge, the parallel tempering methods are only studied under the χ^2 distance [46, 12]. Preconditioned Crank–Nicolson algorithms are only studied under the Wasserstein distance [21]. The unadjusted Langevin algorithm was first studied under the Wasserstein distance [14, 16], and later under the KL divergence [45]. While there might be theoretical value to establishing convergence in both metrics, this is practically unnecessary. In this paper, we assume the convergence of $(X_n)_{n\geq 0}$ under either the Wasserstein distance or χ^2 distance, and study the convergence of $(\widehat{X}_n)_{n\geq 0}$ under one of the two metrics accordingly. We not only address the question qualitatively, but also quantitatively establish bounds for the convergence speed and the approximation accuracy of $(\widehat{X}_n)_{n\geq 0}$.

1.1. Related literature

The approximation and convergence questions we study here are fundamental for MCMC and have been studied in various settings before. Most existing works focus on specific approximation schemes. For example, [35, 6] study ergodicity of $\widehat{P}$ if $\widehat{P}(x, A) = P(x, h^{-1}(A))$ for some round-off function h. [22] studies the ergodicity property of finite-rank non-negative sub-Markov kernels in relation to the ergodicity property of the original Markov kernel. [4] studies the convergence and approximation problems of an adaptive subsampling approach under the assumption of uniform ergodicity. [30] studies the approximation problem for Monte Carlo within Metropolis algorithms. In general, there is a lack of a unified framework.

To provide a comprehensive overview of how perturbation of MCMC affects the approximation accuracy and convergence speed, we put together four sets of results. First, we study the approximation problem under the Wasserstein distance, i.e., the ergodicity framework. The corresponding result (Theorem 2.1) is taken directly from [38]. Approximation accuracy, i.e., bounds for the difference between the nth step distributions of the perturbed chain and the original chain, under the Wasserstein distance has also been studied in [39, 34, 25] under similar but arguably stronger assumptions. For example, [39] requires the perturbed chain to remain close to the original Markov chain uniformly over a bounded number of iterations, while we only require controlling the errors of one-step transition kernels. [34] focuses on MCMC algorithms with the subsampling type of errors. It requires the existence of a subset $\hat{G}$ in which both the unperturbed and perturbed chains remain with a high probability and there is a uniform bound on the errors of one-step transition kernels on $\hat{G}$. [25] requires the Markov chains to be uniformly ergodic, which limits the applicability of the results to non-compact state spaces. Second, we study the convergence problem under the Wasserstein distance. The corresponding result (Theorem 2.2) is new but follows from similar lines of analysis as [38]. The papers [34, 24] also analyze the convergence of the perturbed chain, only under the total variation distance though (and thus requires uniform ergodicity rather than geometric ergodicity). The convergence problem has also been studied in [18], but it does not quantify how the convergence rate depends on the perturbation size. Third and fourth, we study the convergence and approximation

problems under the χ^2-distance. A recent work [32] studies the approximation accuracy and convergence rate of $\widehat{P}$ under the χ^2 distance. Compared to our work, [32] requires stronger assumptions. For example, [32] requires $\widehat{P}$ to be $L_2(\pi) \to L_2(\pi)$, which can be difficult to verify in practice.

As reviewed above, many existing works focus on studying the approximation problem. The convergence problem is less studied. It is worth pointing out that, most of these approximation results only show

$$D(\nu\widehat{P}^n, \pi) = O(\rho^n + \epsilon),$$

where ρ is the convergence rate of P and ϵ is the difference between P and $\widehat{P}$. This does not directly imply the convergence of $D(\nu\widehat{P}^n, \widehat{\pi})$ to zero, since ϵ is nonzero. Moreover, if one has a convergence result, e.g. $D(\nu\widehat{P}^n, \widehat{\pi}) = O(\hat{\rho}^n)$ for some $\hat{\rho} \in (0, 1)$, and $\widehat{\pi}$ has an explicit smaller approximation error $D(\pi, \widehat{\pi}) = o(\epsilon)$, using triangular inequality, we can establish a tighter upper bound for $D(\nu\widehat{P}^n, \pi)$.

Lastly, while the connection between the Wasserstein convergence and MCMC sampling error is well known, most results are asymptotic, i.e., in the form of central limit theorem [26]. Non-asymptotic error bounds are more useful in practice [27]. Our work provides a comprehensive list of finite-sample performance quantifications for numerical MCMC samplers. We demonstrate that our results can be easily applied to the analysis of various algorithms in Sections 4 and 5.

1.2. Notations

For a probability measure μ on $\mathbb{R}^d$, we define

$$\mu f = \int_{\mathbb{R}^d} f(x)\mu(dx), \quad \operatorname{var}_\mu f = \int_{\mathbb{R}^d} (f(x) - \mu f)^2 \mu(dx).$$

We also use $\mu(dx)$ to denote the corresponding density function. For μ-squared integrable functions $f, g : \Omega \to \mathbb{R}$, we define the inner product with respect to μ as

$$\langle f, g \rangle_\mu = \int_{\mathbb{R}^d} f(x)g(x)\mu(dx).$$

Then, $\|f\|_\mu^2 = \langle f, f \rangle_\mu = \int_{\mathbb{R}^d} f(x)^2 \mu(dx)$. In what follows, we omit $\mathbb{R}^d$ from the integral notation when it is clear from the context. For a transition kernel P, define

$$\mu P(A) = \int P(x, A)\mu(dx).$$

For a measurable function f, we define $\delta_x Pf = Pf(x) = \int f(y)P(x,dy)$. We say P is symmetric with respect to π, if for any measurable functions f, g,

$$\langle Pf, g\rangle_\pi = \langle f, Pg\rangle_\pi.$$

Lastly, we denote C as a generic constant whose value can change from line to line.

1.3. Organization

We start by developing general results for the Wasserstein convergence in Section 2, and the χ^2 convergence in Section 3. We demonstrate how to apply these frameworks on two popular Metropolis–Hastings-MCMC algorithms in Section 4, and on the more involved parallel tempering algorithm in Section 5. Finally in Section 6, we verify our claims numerically on a Bayesian inverse problem, which tries to infer the initial condition and model parameters in the predator-prey system.

2. Wasserstein Convergence

We start our discussion with the Wasserstein convergence. Following [38], we first introduce the metric we use and the notion of ergodicity. For a lower semi-continuous function $V : \mathbb{R}^d \to [1, \infty]$, define

$$d_V(x, y) = (V(x) + V(y))1_{x\neq y}.$$

For two probability measures μ and ν on $\mathbb{R}^d$, define

$$\|\mu - \nu\|_V = \sup_{|f|\leq V} \left| \int f(x)(\mu(dx) - \nu(dx)) \right|.$$

It can be shown that $\|\mu-\nu\|_V = W_{d_V}(\mu, \nu)$ where W denotes the Wasserstein distance (Lemma 3.1 in [38]). If we use the constant function $V(x) = 1$, this gives the well-known total variation distance, i.e.,

$$\|\mu - \nu\|_{TV} = \sup_{|f|\leq 1} \left| \int f(x)(\mu(dx) - \nu(dx)) \right|.$$

Note that using $V(x) = 1$ neglects the location information of x. This location information can be crucial for problems with unbounded domains.

In general, for problems with unbounded domains, one often chooses V to be a Lyapunov function. Given a Markov chain X_n with transition kernel P, which we

will write them together as (X_n, P) for short, we say $V : \mathbb{R}^d \to [1, \infty)$ is a Lyapunov function if there exist $\lambda \in (0, 1)$ and $L > 0$, such that

$$PV(x) = \int P(x, dy)V(y) \leq \lambda V(x) + L, \tag{6}$$

and the sublevel sets of V are compact. The choice of V depends on the Markov chain and the target density. It can often be chosen as $\|x\|^2$ or $-\log \pi(x)$ or functions of certain moments, see, for example, [20, 21, 38].

If π is the invariant measure of X_n, we say X_n is geometrically ergodic under $\|\cdot\|_V$ (see Theorem 16.1 in [31]) if there are constants $\rho \in (0, 1)$ and $C_0 \in (0, \infty)$, such that for any $n \in \mathbb{Z}^+$,

$$\|\delta_x P^n - \pi\|_V \leq C_0 \rho^n V(x). \tag{7}$$

We refer to ρ as the ergodicity coefficient. Note that the smaller the value of ρ, the faster the convergence to stationarity. From (7), using the triangle inequality we obtain an equivalent definition of geometric ergodicity, which requires that for any x and y,

$$\|\delta_x P^n - \delta_y P^n\|_V \leq C_0' \rho^n d_V(x, y). \tag{8}$$

The equivalence can be seen from

$$\|\delta_x P^n - \pi\|_V \leq \int \pi(dy)\|\delta_x P^n - \delta_y P^n\|_V \leq C_0' \rho^n (V(x) + \pi V) \leq C_0 \rho^n V(x). \tag{9}$$

where $C_0 = C_0'(1 + \pi V)$.

The approximation problem under Wasserstein distance has been studied in [38]. We present one of their main results here which is related to our subsequent development. Interested readers can find more general discussion in the original work.

Theorem 2.1. (Corollary 3.3 in [38].) *Suppose (X_n, P) is geometrically ergodic, i.e., as in* (8)*. Suppose $\widehat{V}$ is a Lyapunov function for $(\widehat{X}_n, \widehat{P})$ in the sense of* (6)*, and*

$$\|\delta_x P - \delta_x \widehat{P}\|_V \leq \epsilon \widehat{V}(x). \tag{10}$$

Then, for some constant C, we have

$$\|\delta_x P^n - \delta_x \widehat{P}^n\|_V \leq C\epsilon \frac{1 - \rho^n}{1 - \rho} \left(\widehat{V}(x) + \frac{L}{1 - \lambda} \right). \tag{11}$$

The bound in (11) and the triangular inequality give us an approximation error bound

$$\|\delta_x \widehat{P}^n - \pi\|_V \leq \|\delta_x P^n - \delta_x \widehat{P}^n\|_V + \|\delta_x P^n - \pi\|_V \leq C'(\epsilon + \rho^n)(\widehat{V}(x) + V(x))$$

for some constant C'.

Remark 1. In this paper, we focus on a specific form of Wasserstein distance due to its connection to geometric ergodicity. The work [38] also studies the approximation problem under a more general form of Wasserstein ergodicity (see Theorem 3.1 in [38]).

Note that the right-hand side of (11) is not converging to zero as $n \to 0$. Thus, it cannot help us learn the ergodicity of $\widehat{X}_n$ or whether $\widehat{X}_n$ has a unique invariant measure. The next result shows that ergodicity can be obtained with essentially the same conditions as Theorem 2.1 (note that condition (7) leads to (12) through (9)).

Theorem 2.2. *Suppose V is a Lyapunov function for P in the sense of* (6)*. In addition, assume there exist $N \in \mathbb{Z}^+$ and $\rho \in (0,1)$, such that for any $n \geq N$,*

$$\|\delta_x P^n - \delta_y P^n\|_V \leq \rho^n d_V(x,y). \tag{12}$$

Lastly, there is an $\epsilon_0 > 0$ and the following holds for some $\epsilon \in (0, \epsilon_0]$,

$$\|\delta_x P - \delta_x \widehat{P}\|_V \leq \epsilon V(x). \tag{13}$$

Then, V is a Lyapunov function for $\widehat{P}$ as well with

$$\widehat{P}V(x) \leq (\lambda + \epsilon)V(x) + L.$$

Moreover, $\widehat{X}_n$ has a unique invariant measure $\widehat{\pi}$ and there exist $C_1, D_1 \in (0, \infty)$ independent of ϵ, such that

$$\|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V \leq C_1(\rho + D_1\epsilon)^n d_V(x,y).$$

In Theorem 2.2, ϵ_0 is chosen such that $\lambda + \epsilon_0 < 1$ and $\rho + D_1\epsilon_0 < 1$. Theorem 2.2 indicates that if P is geometrically ergodic with ergodicity coefficient ρ and $\widehat{P}$ is ϵ-close to P as characterized by (13), $\widehat{P}$ is also geometrically ergodic. Moreover, the ergodicity coefficient of $\widehat{P}$ is bounded above by $\rho + D_1\epsilon$.

In statistical applications, we are more interested in turning convergence results into error bounds for the Monte Carlo estimators. The central limit theorem of ergodic

Markov chains was studied in [44, 26], which provides asymptotic error quantifications. In practice, non-asymptotic bounds for finite values of n may be more desirable. The following proposition appeared in [37, 27]. We provide an explicit statement here to show the variance bound along with a simple proof for self-completeness. For simplicity, we assume the Markov chain is initialized with the invariant measure, i.e., $\widehat{X}_0 \sim \widehat{\pi}$, so a burn-in period is not necessary.

Proposition 2.1. *Suppose* $\|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V \leq \widehat{\rho}^n d_V(x,y)$ *for some* $\widehat{\rho} \in (0,1)$. *Then, for any* f *that is 1-Lipschitz under* $\|\cdot\|_V$,

$$\left|\delta_x \widehat{P}^n f - \widehat{\pi} f\right| \leq \widehat{\rho}^n (V(x) + \widehat{\pi} V).$$

In addition, if we use $\hat{f}_M = \frac{1}{M}\sum_{k=1}^M f(\widehat{X}_k)$ *as an estimator of* $\widehat{\pi} f$ *starting from* $\widehat{X}_0 \sim \widehat{\pi}$,

$$\mathbb{E}_{\widehat{\pi}}\left[(\hat{f}_M - \widehat{\pi} f)^2\right] \leq \frac{2}{(1-\widehat{\rho})M} \mathbb{E}_{\widehat{\pi}}\Big[|f(\widehat{X}_0)|(V(\widehat{X}_0) + \widehat{\pi} V)\Big].$$

In many applications, we are interested in the properties of $\widehat{\pi}$ on compact regions. In these scenarios, the associated test functions will be bounded, and 1-Lipschitz under $\|\cdot\|_V$. To learn tail properties of $\widehat{\pi}$ such as intermittency, the test functions often need to grow with $\|x\|$. In order to apply Proposition 2.1, we would need the Lyapunov function V to grow at a similar scale.

3. χ^2 convergence

In this section, we discuss convergence under the χ^2 divergence. We first introduce some notations. For a transition kernel P and density μ, define

$$\|P\|_\mu = \max_{f: 0 < \|f\|_\mu < \infty} \frac{\|Pf\|_\mu}{\|f\|_\mu},$$

where $\|f\|_\mu^2 = \langle f, f\rangle_\mu$. For two probability measures μ and ν on $\mathbb{R}^d$, where ν is absolutely continuous with respect to μ, define the χ^2 divergence of ν from μ as:

$$D_{\chi^2}(\nu\|\mu) = \int \left(\frac{\nu(x)}{\mu(x)} - 1\right)^2 \mu(x)dx = \int \frac{\nu(x)^2}{\mu(x)}dx - 1.$$

For a transition kernel P that is reversible with respect to π, the spectral gap of P is defined as [21]

$$\kappa(P) = 1 - \sup\left\{\frac{\|Pf - \pi f\|_\pi^2}{\|f - \pi f\|_\pi^2} : f \in L^2(\pi), \operatorname{var}_\pi f \neq 0\right\}. \tag{14}$$

Note that by repeatedly applying (14), we have for any $f \in L^2(\pi)$,

$$\|P^n f - \pi f\|_\pi^2 \leq (1-\kappa(P))^n \|f - \pi f\|_\pi^2.$$

Thus, the larger the spectral gap, the faster X_n converges to its invariant measure.

Remark 2. An alternative definition of the spectral gap takes the form [3, 2]

$$\kappa_a(P) = \inf\left\{ \frac{\langle f, (I-P)f\rangle_\pi}{\operatorname{var}_\pi f} : f \in L^2(\pi), \operatorname{var}_\pi f \neq 0 \right\}.$$

Note that these two spectral gaps are related through $1-\kappa(P) = (1-\kappa_a(P))^2$.

3.1. General χ^2 approximation and convergence

Our first result assumes that (X_n, P) has a spectral gap and $\widehat{P}$ is a close approximation of P:

Theorem 3.1. *Suppose P is a reversible transition kernel with invariant measure π and a spectral gap $\kappa(P) > 0$ in the sense of (14). There is an $\epsilon_0 > 0$ such that for any transition kernel $\widehat{P}$ satisfying $\|P - \widehat{P}\|_\pi \leq \epsilon < \epsilon_0$ and any $a \in (0,1)$, there exists a constant C that may depend on $\kappa(P)$ such that the following holds with $\widehat{\kappa} = (1-a)\kappa(P) - \frac{C\epsilon^2}{a} \in (0,1)$:*

1. *For any $f \in L^2(\pi)$,*

$$\|\widehat{P}^n f - \pi\widehat{P}^n f\|_\pi^2 \leq (1-\widehat{\kappa})^n \operatorname{var}_\pi f$$

2. *$\widehat{P}$ has an invariant measure $\widehat{\pi}$, which satisfies*

$$|(\widehat{\pi} - \pi\widehat{P}^n)f|^2 \leq C\epsilon^2(1-\widehat{\kappa})^n \operatorname{var}_\pi f.$$

Moreover, $D_{\chi^2}(\widehat{\pi}\|\pi) \leq C\epsilon^2$.

In Theorem 3.1, ϵ_0 is chosen such that $\hat{\kappa} \in (0,1)$. Theorem 3.1 indicates that if $\widehat{P}$ and P are ϵ-close to each other as quantified by $\|\widehat{P} - P\|_\pi \leq \epsilon$, $\widehat{P}$ has a stationary distribution $\widehat{\pi}$. Moreover, $\widehat{\pi}$ and π are ϵ-close to each other as quantified by $D_{\chi^2}(\widehat{\pi}\|\pi) \leq C\epsilon^2$. We also note that showing that $\|\widehat{P}^n f - \pi\widehat{P}^n f\|_\pi^2 \leq (1-\widehat{\kappa})^n \operatorname{var}_\pi f$ is different from finding the spectral gap of $\widehat{P}$, since the latter would need a similar inequality but with π replaced by $\widehat{\pi}$. In other words, Theorem 3.1 does not provide a spectral gap for $\widehat{P}$. On the other hand, we can obtain error bounds for Monte Carlo estimators using the bounds established in Theorem 3.1:

Proposition 3.1. *Under the same conditions as those in Theorem 3.1, for any $f \in L^2(\pi)$ and any initial distribution $\widehat{X}_0 \sim \nu \ll \pi$, there exists a constant C such that*

$$\left|\nu\widehat{P}^n f - \widehat{\pi} f\right|^2 \leq (1-\widehat{\kappa})^n \operatorname{var}_\pi(f) \left(\sqrt{D_{\chi^2}(\nu\|\pi)+1} + C\epsilon\right)^2.$$

In addition, if f is bounded, there exists a constant C such that

$$\mathbb{E}_{\widehat{\pi}}[(\hat{f}_M - \widehat{\pi} f)^2] \leq \frac{C}{M(1-(1-\widehat{\kappa})^{1/4})} \sqrt{\operatorname{var}_{\widehat{\pi}}(f) \operatorname{var}_\pi(f)},$$

where $\hat{f}_M = \frac{1}{M}\sum_{k=1}^M f(\widehat{X}_k)$.

Remark 3. [32] provides a result similar to the first claim in Theorem 3.1 (see Lemma A.6 in [32]). But it requires stronger assumptions on $\widehat{P}$, namely it requires that $\widehat{P}$ is ergodic and aperiodic, and is a mapping from $L_2(\pi)$ to $L_2(\pi)$. Our result does not require these assumptions.

3.2. Spectral gap with density ratio bounds

In this section, we show that stronger results can be established if we can bound the ratio between the invariant densities π and $\widehat{\pi}$. Such a bound is assessable if we have an explicit characterization of $\widehat{\pi}$. For example, in Bayesian inverse problems, $\pi(x) \propto p_0(x)\exp(-\frac{1}{2}\|G(x)-y\|^2)$ while $\widehat{\pi}(x) \propto p_0(x)\exp(-\frac{1}{2}\|\widehat{G}(x)-y\|^2)$. In this case, a density ratio bound can be obtained if $\|G(x)-\widehat{G}(x)\|$ is bounded. This is practically feasible by using an accurate numerical approximation of G, and the approximation error can be estimated by the grid size or Galerkin truncation used in the numerical scheme (see, for example, [23]). Similar assumptions have also been imposed in existing Bayesian computation literature, see, for example, [5].

Theorem 3.2. *Suppose P and $\widehat{P}$ are two reversible transition kernels with invariant densities π and $\widehat{\pi}$ respectively. We further assume $\pi(x)/\widehat{\pi}(x) \in [(1+\epsilon)^{-1}, 1+\epsilon]$ and $\|P-\widehat{P}\|_\pi \leq \epsilon$. Then, there exists a universal constant C such that*

$$\kappa(\widehat{P}) \geq \kappa(P) - C\epsilon.$$

Note that for ϵ small enough, $\kappa(P) - C\epsilon > 0$. Based on the spectral gap, we have the following non-asymptotic Monte Carlo error bound.

Proposition 3.2. *Suppose $(\widehat{X}_n, \widehat{P})$ has a spectral gap $\widehat{\kappa}$. Suppose the initial distribution is ν, i.e., $\widehat{X}_0 \sim \nu$. Then,*

$$\left|\mathbb{E}f(\widehat{X}_n) - \widehat{\pi}f\right|^2 \leq (1-\widehat{\kappa})^n \operatorname{var}_{\widehat{\pi}} f(D_{\chi^2}(\nu\|\widehat{\pi}) + 1).$$

In addition, if $\nu = \widehat{\pi}$, then for $\hat{f}_M = \frac{1}{M}\sum_{k=1}^{M}$, we have

$$\mathbb{E}_{\widehat{\pi}}[(\hat{f}_M - \widehat{\pi}f)^2] \leq \frac{2}{M(1-(1-\widehat{\kappa})^{1/2})} \operatorname{var}_{\widehat{\pi}}[f].$$

Note that Proposition 3.2 is not a new result. A more delicate central-limit-theorem version of it can be found as Theorem 4.4 of [21]. We provide a short proof of the proposition in the appendix for self-completeness.

Before we conclude our discussion of the χ^2 convergence, we remark that even though the condition $\|P-\widehat{P}\|_\pi \leq \epsilon$ is reasonable for the spectral gap analysis, it can be hard to verify directly in some applications. To remedy this issue, the next proposition shows that we can bound $\|P-\widehat{P}\|_\pi$ through a bound for $\|\delta_x P - \delta_x \widehat{P}\|_{TV}$, which can be easier to obtain using coupling techniques.

Proposition 3.3. *Suppose there exists a π-measurable function $V : \mathbb{R}^d \to [1,\infty)$ such that $\|\delta_x P - \delta_x \widehat{P}\|_{TV} \leq \epsilon V(x)$. In addition, suppose $\frac{1}{a} \leq \pi(x)/\hat{\pi}(x) \leq a$ for some constant $a > 0$. Then,*

$$\|P - \widehat{P}\|_\pi \leq \sqrt{2(1+a^2)}\sqrt{\epsilon}\|V\|_\pi^{1/2}.$$

4. Application: Metropolis-Hastings MCMC on perturbed densities

Random walk Metropolis (RWM) and Metropolis adjusted Langevin algorithm (MALA) are two popular MCMC samplers when it comes to sampling a generic density π. Many existing works have already studied their spectral gap under suitable conditions on π [36, 21, 15]. When implementing these samplers, it is often the case that we only have access to an approximation of π, which we denote as $\widehat{\pi}$. In this section, we will demonstrate how to apply our analysis framework to establish proper bounds for the spectral gap of the "numerical" RWM and MALA.

In fact, we can develop some general results for the Metropolis-Hastings (MH) type of Monte Carlo algorithm. Assume the proposals are given by some smooth transition density $R(x,x')$. Due to the possibility of rejection, MH Monte Carlo transition

densities w.r.t. Lebesgue measure can be written as $P(x,x') = \gamma(x)\delta_x(x') + \beta(x,x')$ with

$$\beta(x,x') = \min\left\{\frac{\pi(x')R(x',x)}{\pi(x)}, R(x,x')\right\}, \quad \gamma(x) = 1 - \int \beta(x,x')dx'. \tag{15}$$

The perturbed transition density can be written as $\widehat{P}(x,x') = \hat{\gamma}(x)\delta_x(x') + \widehat{\beta}(x,x')$. We provide some sufficient conditions under which the difference between P and $\widehat{P}$ is of order ϵ.

Lemma 4.1. *If the transition density is of the form $P(x,x') = \gamma(x)\delta_x(x') + \beta(x,x')$ with $\nu(x)P(x,x') = \nu(x')P(x',x)$, suppose $\widehat{P}(x,x') = \hat{\gamma}(x)\delta_x(x') + \hat{\beta}(x,x')$ with*

$$|\hat{\gamma}(x) - \gamma(x)| \le C\epsilon \quad \textit{and} \quad (1 - C\epsilon)\beta(x,x') \le \hat{\beta}(x,x') \le (1 + C\epsilon)\beta(x,x').$$

for some constant $C \in (0,\infty)$. Then, there exists a constant $C_1 \in (0,\infty)$ such that $\|P - \widehat{P}\|_\nu \le C_1\epsilon$.

4.1. Random walk Metropolis

RWM considers implementing the MH procedure on random walk proposals. That is, we use

$$R(x,x') = \frac{1}{(2\pi h)^{d/2}} \exp\left(-\frac{1}{4h}\|x' - x\|^2\right)$$

in (15). It is worth noting that using a perturbed density $\widehat{\pi}$ does not affect this proposal.

Proposition 4.1. *For RWM, there is an $\epsilon_0 > 0$ so that for any $\epsilon < \epsilon_0$ and $\sup_x |\log \pi(x) - \log \hat{\pi}(x)| \le C\epsilon$, there is a constant C_1 so that*

$$\|P_{RWM} - \widehat{P}_{RWM}\|_\pi \le C_1\epsilon.$$

If the original RWM has a spectral gap and $\sup_x |\log \pi(x) - \log \hat{\pi}(x)| \le C\epsilon$, then Proposition 4.1 together with Theorem 3.2 implies that the perturbed RWM has a proper spectral gap as well. In practice, an estimate of ϵ can be obtained by analyzing the numerical scheme used, see Section 3.2 for more details.

4.2. Metropolis adjusted Langevin algorithm

MALA considers implementing the MH procedure on proposals following the Langevin diffusion. That is, we use

$$R(x, x') = \frac{1}{(4\pi h)^{d/2}} \exp\left(-\frac{1}{4h}\|x' - x - h\nabla \log \pi(x)\|^2\right)$$

in (15). Using a perturbed density $\widehat{\pi}$ does change this proposal. We discuss the perturbation in two separate cases. In particular, we shall verify that the condition $\|P_{MALA} - \widehat{P}_{MALA}\|_\pi \leq \epsilon$ holds under appropriate assumptions on $\widehat{\pi}$ in the two cases. Then, if P_{MALA} has a spectral gap, the numerical sampler $\widehat{P}_{MALA}$ has a proper spectral gap as well.

4.2.1. *Bounded domain* When the support of π and $\widehat{\pi}$ are bounded, the analysis is quite straightforward with Lemma 4.1.

Proposition 4.2. *For MALA, there is an $\epsilon_0 > 0$ so that for any $\epsilon < \epsilon_0$, if $\sup_x |\log \pi(x) - \log \hat{\pi}(x)| \leq C\epsilon$, $\sup_x \|\nabla \log \pi(x) - \nabla \log \hat{\pi}(x)\| \leq C\epsilon$, and the support of π and $\widehat{\pi}$ are bounded, then*

$$\|P_{MALA} - \widehat{P}_{MALA}\|_\pi = O(\epsilon).$$

4.2.2. *Unbounded support* When the support of the density is unbounded, directly bounding $\|P_{MALA} - \widehat{P}_{MALA}\|_\pi$ becomes difficult. Instead, we consider establishing $\|\delta_x P - \delta_x \widehat{P}\|_{TV} = O(\epsilon)$.

Proposition 4.3. *For MALA, if $\log \pi$ is Lipschitz, $\sup_x |\log \pi(x) - \log \hat{\pi}(x)| \leq L_\pi \epsilon$, and moreover $\sup_x \|\nabla \log \pi(x) - \nabla \log \widehat{\pi}(x)\| \leq L_\pi \epsilon$, for any $\delta > 0$, there exists $C_\delta \in (0, \infty)$, such that for $h < (\frac{5L_\pi}{\delta} + 20L_\pi)^{-1}$,*

$$\|\delta_x P - \delta_x \widehat{P}\|_{TV} \leq C_\delta \epsilon \exp(\delta \|x\|^2).$$

When $\pi(x)$ is sub-Gaussian, we can find a $\delta > 0$ such that $V(x) = \exp(\delta\|x\|^2)$ is L_2-integrable under π. Then Proposition 3.3 indicates that $\|P - \widehat{P}\|_\pi = O(\sqrt{\epsilon})$.

5. Application: Parallel Tempering with Perturbed Densities

In this section, we demonstrate how to apply our framework to parallel tempering (PT) algorithms [17, 42, 43]. These algorithms are also referred to as the replica

exchange methods [41, 13, 11]. Compared with regular MCMC samplers like RWM and MALA, PT tries to sample a multiple-tempered version of the target density. Such a design can improve the convergence rate on densities with multiple isolated modes.

To implement PT, a sequence of distributions $\pi_0, \ldots, \pi_K$ are considered where the last one is the target density $\pi_K = \pi$. The first density π_0 is usually a distribution that is easy to draw samples from. The intermediate distributions, π_k's $1 \leq k \leq K-1$, are set up so that the two neighboring densities are similar to each other. A common choice for the intermediate distributions is to consider interpolations between π_K and π_0:

$$\pi_k(x) \propto \pi^{\beta_k}(x)\pi_0^{1-\beta_k}(x),$$

where $0 = \beta_0 < \beta_1 < \ldots < \beta_K = 1$ is a sequence of parameters. PT intends to generate samples from the product density

$$\Pi = \pi_0 \times \pi_1 \times \cdots \times \pi_K \text{ on } \mathbb{R}^{d(K+1)}.$$

To do so, its iterations consist of $K+1$ parts, i.e., $X_n = (X_n^0, \ldots, X_n^K)$, and the updating rule is given by the following two steps.

1. Updating each X_n^k to X_{n+1}^k according to a transition kernel M_k, whose stationary distribution is π_k. In practice, M_k is often taken as the transition kernel obtained by repeating RWM or MALA update for t_k steps. That is $M_k = P_{RWM}^{t_k}$ or $M_k = P_{MALA}^{t_k}$.

2. Pick an index $k \in \{0, \ldots, K-1\}$ uniformly at random and swap the values of X_{n+1}^k and X_{n+1}^{k+1} with probability $\alpha_k(X_{n+1}^k, X_{n+1}^{k+1})$, where

$$\alpha_k(x, x') = \min\left\{1, \frac{\pi_k(x')\pi_{k+1}(x)}{\pi_k(x)\pi_{k+1}(x')}\right\}.$$

The pseudo-code of PT is given in Algorithm 1. The exchange procedure can be described by the transition probability on $R^{(K+1)d} \times R^{(K+1)d}$:

$$Q_k(x, x) = 1 - \alpha_k(x^k, x^{k+1}), \quad Q_k(x, S_k(x)) = \alpha_k(x^k, x^{k+1}),$$

where $S_k(x) = (x^0, \ldots, x^{k-1}, x^{k+1}, x^k, x^{k+2}, \ldots, x^K)$. With a little abuse of notation, we write the transition kernel as Q_k as well, i.e., $Q_k f(x) = Q_k(x, x)f(x) +$

Algorithm 1 Parallel Tempering

Input: Replica counts K, target densities π_k for $k = 0, \ldots, K$, transition kernels M_k targeting π_k.

Output: $(x_t^k)_{t=0,\ldots,T_k,k=0,\ldots,K}$ as samples from π_k

Initialize x_0^k for all k

for $t = 0$ to T **do**

  for $k = 0$ to K **do** %Run MCMC at each level

    Generate $x_{t+1}^k \sim M_k(x_t^k, \cdot)$.

  end for

  %Consider swapping at a random level

  Let k be a random index in $\{0, \ldots, K-1\}$

  Let U be a random sample from Unif$[0, 1]$

  if $U < a_k(x_{t+1}^k, x_{t+1}^{k+1})$ **then**

    $(x_{t+1}^k, x_{t+1}^{k+1}) = (x_{t+1}^{k+1}, x_{t+1}^k)$

  end if

end for

$Q_k(x, S_k(x))f(S_k(x))$. The transition kernel of PT can then be written as

$$P = (M_0 \otimes \cdots \otimes M_K) \left(\frac{1}{K} \sum_{\mathbf{k} \in \{0,\ldots,K-1\}} Q_k \right), \tag{16}$$

where the direct product of two transition kernels is given by

$$M_0 \otimes M_1 f(x^0, x^1) = \int \int M_0(x^0, y^0) M_1(x^1, y^1) f(y^0, y^1) dy^0 dy^1.$$

The spectral gap of P in (16) has been studied in [46]. Assume the state space can be partition into $\mathbb{R}^d = \cup_{i=1}^J A_j$, it is shown that $\kappa(P)$ can be seen as the product of three elements: 1) the maximal spectral gap when sampling π_k, $k \geq 1$, constrained on one piece A_j; 2) the spectral gap when sampling π_0 using M_0; and 3) the density ratio: $\pi_k(A_j)/\pi_{k+1}(A_j)$. In particular, if π_0 is easy to sample, π_k is not so different from π_{k+1}, and the sampling of π_k constrained on A_j is efficient, then PT can be highly efficient.

When implementing PT numerically, we may not have access to the exact values of π_k, but only an ϵ-approximation, which we denote as $\widehat{\pi}_k$. Then, the corresponding

PT uses a sampler $\widehat{M}_k$ with invariant measure $\widehat{\pi}_k$ at each replica, while the exchange probability is given by

$$\widehat{\alpha}_k(x, x') = \min\left\{1, \frac{\widehat{\pi}_k(x')\widehat{\pi}_{k+1}(x)}{\widehat{\pi}_k(x)\widehat{\pi}_{k+1}(x')}\right\}.$$

The corresponding transition kernel can be written as

$$\widehat{P} = \left(\widehat{M}_0 \otimes \cdots \otimes \widehat{M}_K\right)\left(\frac{1}{K}\sum_{\mathbf{k}\in\{0,\ldots,K-1\}} \widehat{Q}_k\right).$$

It is natural to ask whether this numerical PT will inherit the spectral gap of P. The next result together with Theorem 3.2 indicates that under appropriate regularity conditions on $\widehat{\pi}_k$'s, the numerical PT also has a proper spectral gap.

Proposition 5.1. *Suppose for each replica the target distribution satisfies* $\sup_x |\log\widehat{\pi}_k(x) - \log\pi_k(x)| \le \epsilon$ *and the transition kernel satisfies* $\|P_k - \widehat{P}_k\|_{\pi_k} \le \epsilon$*, then the transition kernel of PT satisfies the following for some constant* C*:*

$$\|P - \widehat{P}\|_\Pi \le C\epsilon.$$

Before we prove Proposition 5.1, we first prove two auxiliary lemmas. The first lemma shows that different compositions of approximated transition kernels yield approximation kernels of similar accuracy. In particular, it helps us establish the condition $\|P_k - \widehat{P}_k\|_{\pi_k} \le \epsilon$ in Proposition 5.1 if we use $M_k = P_{RWM}^{t_k}$ or $M_k = P_{MALA}^{t_k}$.

Lemma 5.1. *1) For two transition kernels* R *and* S*, both with invariant measure* ν*, if* $\|R - \widehat{R}\|_\nu \le C\epsilon$ *and* $\|S - \widehat{S}\|_\nu \le C\epsilon$*, then there is a constant* C' *so that*

$$\|RS - \widehat{R}\widehat{S}\|_\nu \le C'\epsilon.$$

2) For two transition kernels R_1 *and* R_2 *with invariant measure* ν_1 *and* ν_2 *respectively, if* $\|R_1 - \hat{R}_1\|_{\nu_1} \le C\epsilon$ *and* $\|R_2 - \hat{R}_2\|_{\nu_2} \le C\epsilon$*, then there is a constant* C' *so that*

$$\|R_1 \otimes R_2 - \widehat{R}_1 \otimes \widehat{R}_2\|_\nu \le C'\epsilon,$$

where $\nu = \nu_1 \times \nu_2$ *is the joint invariance distribution.*

3) For n transition kernels $S_1, S_2, \ldots, S_n$, all with invariant measure ν, if $\|S_i - \widehat{S}_i\|_\nu \leq C\epsilon$ for $i = 1, \ldots, n$, then for $U = \frac{1}{n}\sum_{i=1}^{n} S_i$ and $\widehat{U} = \frac{1}{n}\sum_{i=1}^{s} \widehat{S}_i$, there is a constant C' so that

$$\|U - \widehat{U}\|_\nu \leq C'\epsilon.$$

The second lemma establishes proper bounds for the swapping transition.

Lemma 5.2. *Let Q be a transition probability of form:*

$$Q(x, S(x)) = a(x, S(x)), \quad Q(x, x) = 1 - a(x, S(x)),$$

where $S(x)$ is some given map. Suppose Q is reversible with a density ν, i.e.

$$\nu(x)Q(x, S(x)) = \nu(S(x))Q(S(x), x).$$

Similarly, let $\widehat{Q}$ denote the transition probability of the form

$$\widehat{Q}(x, S(x)) = \widehat{a}(x, S(x)), \quad \widehat{Q}(x, x) = 1 - \widehat{a}(x, S(x)),$$

reversible with $\widehat{\nu}$. If for some constant C, $(1 - C\epsilon)a(x, S(x)) \leq \widehat{a}(x, S(x)) \leq (1 + C\epsilon)a(x, S(x))$, then

$$\|(Q - \widehat{Q})f\|_\nu \leq 2C\epsilon\|f\|_\nu.$$

6. Numerical examples

In this section, we present some numerical examples based on the predator-prey system to illustrate the theoretical results developed in the preceding sections.

6.1. Predator-prey system

We consider inferring the parameters of a system of ordinary differential equations (ODEs) that models the predator-prey system [29]. Denoting the populations of prey and predator by (γ_p, γ_q), the populations change over time according to the pair of coupled ODEs:

$$\begin{aligned} \frac{d\gamma_p}{dt} &= r\gamma_p\left(1 - \frac{\gamma_p}{K}\right) - s\left(\frac{\gamma_p\,\gamma_q}{w + \gamma_p}\right), \\ \frac{d\gamma_q}{dt} &= u\left(\frac{\gamma_p\,\gamma_q}{w + \gamma_p}\right) - v\gamma_q, \end{aligned} \tag{17}$$

with initial conditions $\gamma_p(0)$ and $\gamma_q(0)$. r, K, a, s, u, and v are model parameters that control the dynamics of the populations of prey and predator. In the absence of the predator, the population of prey evolves according to the logistic equation, which is characterized by r and K. In the absence of the prey, the population of the predator has an exponential decay rate v. The additional parameters s, w, and u characterize the interaction between the predator population and the prey population.

In the inference problem, we want to estimate both the model parameters and the initial conditions. In this case, we have $d = 8$ and denote

$$\theta = (\gamma_p(0), \gamma_q(0), r, K, a, s, u, v).$$

A commonly used prior for this problem is a uniform distribution over a hypercube $(a_1, b_1) \times \cdots \times (a_d, b_d)$ (see, e.g., [33]). Here, we set $a_i = 10^{-3}$ and $b_i = 2 \times 10^2$ for all i. Noisy observations of both $\gamma_p(t; \theta)$ and $\gamma_q(t; \theta)$ at times regularly spaced at $m = 20$ time points in $t \in [2, 40]$ are used to infer θ. This defines a so-called forward model

$$F(\theta) = [\gamma_p(t_1; \theta), \gamma_q(t_1; \theta), \ldots, \gamma_p(t_m; \theta), \gamma_q(t_m; \theta)],$$

that maps a given parameter θ to the observables. The observables are perturbed with independent Gaussian observational errors with mean zero and variance 4. A "true" parameter

$$\theta_{\text{true}} = [50, 5, 0.6, 100, 1.2, 25, 0.5, 0.3]^\top$$

is used to generate the synthetic observed data set, which is denoted by y. The trajectories of $\gamma_p(t; \theta_{\text{true}})$ and $\gamma_q(t; \theta_{\text{true}})$ together with the synthetic data set are shown in Figure 1.

To avoid rejections caused by proposal samples that fall outside of the hypercube, we further consider the prior distribution as the pushforward of the standard Gaussian measure with the probability density function

$$p_0(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|x\|^2\right)$$

under a diffeomorphic transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ that maps each coordinate

$$\theta_i = T_i(x_i) = a_i + \frac{b_i - a_i}{\sqrt{2\pi}} \int_{-\infty}^{x_i} \exp\left(-\frac{1}{2}z^2\right) dz.$$

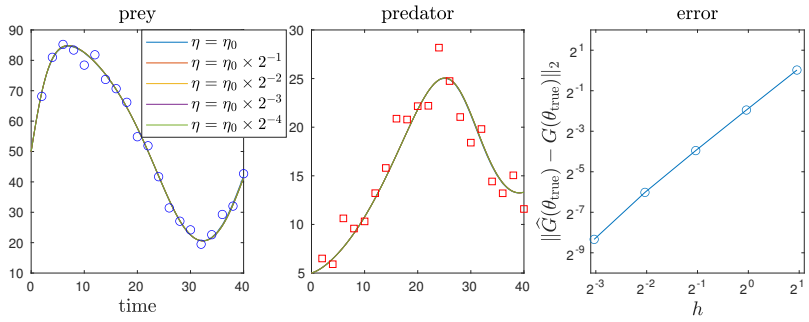

FIGURE 1: Left and middle: the trajectories of $\gamma_p(t;\theta_{\rm true})$ and $\gamma_q(t;\theta_{\rm true})$ computed using the second order Runge–Kutta method with different time step size η. Right: the L_2 error of the model outputs with different time step sizes η. Here $G(\theta_{\rm true})$ is computed using $\eta = \eta_0 \times 2^{-6}$. The trajectories computed by the time step size $\eta = \eta_0 \times 2^{-6}$ are used to generate a synthetic data set. The observed data sets of the prey and predator are shown as circles and squares, respectively.

In other words, $p_0(x)$ is the prior distribution for the transformed parameter $x = T^{-1}(\theta)$. Writing $G(x) = F(T(x))$, our goal is to characterize the posterior distribution

$$\pi(x) \propto p_0(x) \exp\left(-\frac{1}{8}\|G(x) - y\|^2\right).$$

The system of ODEs in (17), and hence the function $G(x)$, has to be numerically solved by some ODE solvers. Here we use the second order explicit Runge–Kutta method with time step size η to solve (17). As shown in Figure 1, the trajectories of $\gamma_p(t;\theta_{\rm true})$ and $\gamma_q(t;\theta_{\rm true})$ converge as $\eta \to 0$ at a rate of $O(\eta^2)$ (see [8] for a detailed analysis). The numerical solver, which is characterized by the step size η, defines the approximate model $\widehat{G}(x)$ and the approximate posterior density $\widehat{\pi}(x)$. Figure 2 shows the estimated marginal distributions (using Algorithm 1) of perturbed posteriors defined by various time step sizes. Here we observe that as h decreases, the estimated marginal distributions almost overlap each other, which suggests that the perturbed distributions converge as the discretized model $\widehat{G}$ converges.

6.2. MCMC results

To validate the theoretical results on Metropolis–Hasting MCMC on perturbed densities in Section 4, we first simulate the RWM algorithm with invariant densities $\widehat{\pi}(x)$ defined by various time step sizes as shown in Figure 1. All the Markov chains

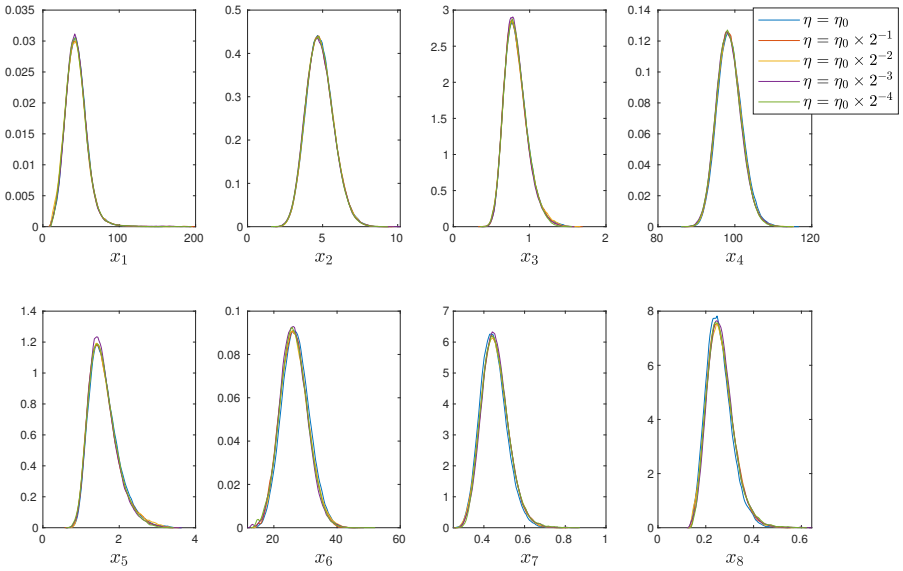

FIGURE 2: Marginal distributions of perturbed posteriors defined by various time step sizes.

in this set of simulation experiments are generated using the same Gaussian random walk proposal distribution. The box plots of the integrated autocorrelation times (IACTs) of the resulting Markov chains are shown in Figure 3. Then, we simulate MALA with invariant densities $\widehat{\pi}(x)$ defined by the same set of time step sizes. The box plots of resulting IACTs are shown in Figure 4. Again, all the Markov chains are generated using the same proposal distribution. For both algorithms, we simulate each Markov chain for 10^6 iterations after discarding burn-in samples and repeat the simulation for 20 times with different initial states to produce the box plots. As established in our theoretical analysis, for both algorithms, the resulting Markov chains targeting various approximate posterior densities produce similar IACTs. This provides empirical evidence that the spectral gaps of the approximate transition kernels defined by MRW or MALA converge as the discretization step size $\eta \to 0$.

To validate the theoretical results on the parallel tempering with perturbed densities in Section 5, we simulate Algorithm 1 with the same Gaussian random walk as in RWM. For each of the invariant densities $\widehat{\pi}(x)$ defined by various time step sizes, we set $K = 4$,

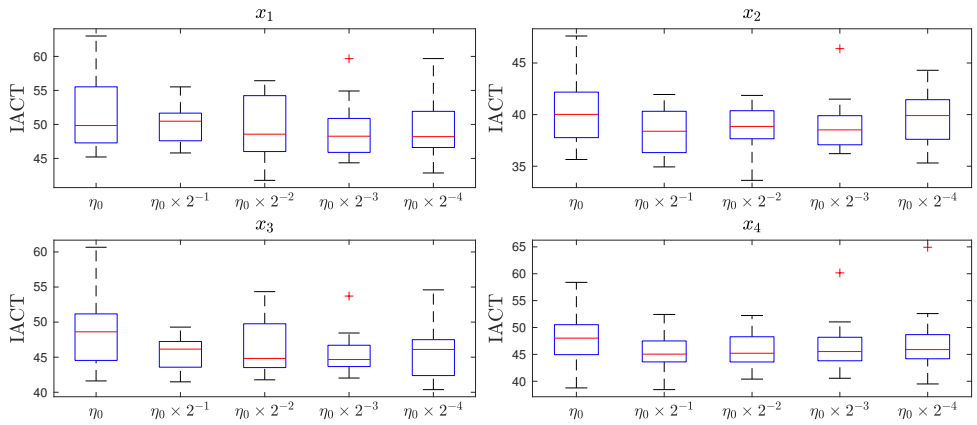

FIGURE 3: Box plots of integrated autocorrelation times (IACTs) of the first four parameter Markov chains simulated by the RWM algorithm. Here 20 realizations of Markov chains are used to estimate the IACT.

and the intermediate distributions take the form

$$\widehat{\pi}_k(x) \propto p_0(x) \exp\left(-\frac{\beta_k}{8}\|\widehat{G}(x) - y\|^2\right),$$

where $\beta_k = 1+\alpha^{-K}-\alpha^{-k}$ with $\alpha = 1.3$ and $k = \{0, 1, 2, 3, 4\}$. Here β_k is an increasing sequence such that $\beta_K = 1$. The same Gaussian random walk is used across all replicas to simulate the Markov chain. The box plots of the IACTs of resulting Markov chains are shown in Figure 5. Similar to the previous experiments, the resulting Markov chains targeting various approximate posterior densities produce similar IACTs. This provides empirical evidence that the spectral gaps of the approximated transition kernel induced by Algorithm 1 converge as the discretization step size $\eta \to 0$. Furthermore, we noticed that the IACTs of Algorithm 1 are smaller than those of the RMW and MALA algorithm, which suggests that Algorithm 1 has a better mixing rate.

7. Conclusion

In this paper, we quantify the convergence speed and the approximation accuracy of numerical MCMC samplers under two general frameworks: ergodicity and spectral gap. Our results can be easily applied to study the efficiency and accuracy of various sampling algorithms. In particular, we demonstrate how to apply our framework to study Metropolis Hasting MCMC algorithms and parallel tempering Monte Carlo

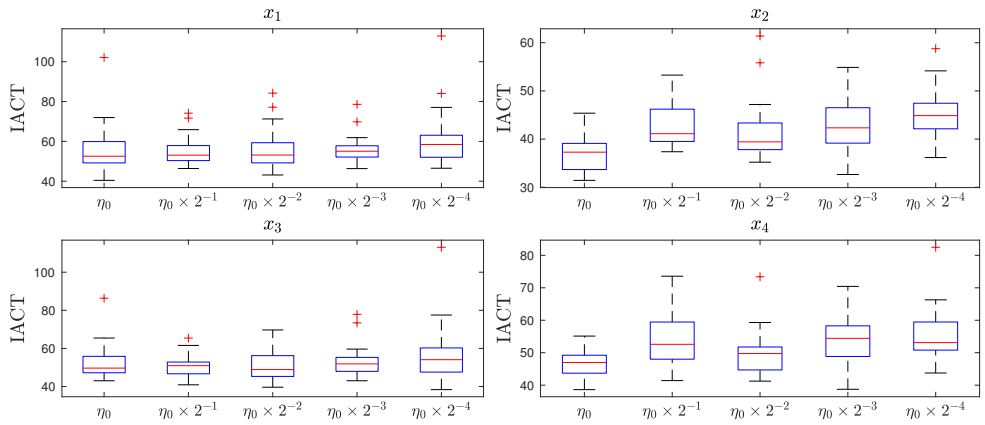

FIGURE 4: Box plots of integrated autocorrelation times (IACTs) of the first four parameter Markov chains simulated by the MALA algorithm. Here 20 realizations of Markov chains are used to estimate the IACT.

algorithms. These results are validated by numerical simulations on a Bayesian inverse problem based on the predator-prey model.

In the applications in Sections 4 and 5, we assume that we can approximate the target transition kernels with some uniform accuracy guarantees. In some problems with unbounded domains, this can be difficult to achieve. It would be interesting to relax such requirements to the ones assumed in [10].

References

[1] ANDRIEU, C., DE FREITAS, N., DOUCET, A. AND JORDAN, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning* **50,** 5–43.

[2] ANDRIEU, C. AND VIHOLA, M. (2015). Convergence properties of pseudo-marginal markov chain monte carlo algorithms. *The Annals of Applied Probability* **25,** 1030–1077.

[3] BAKRY, D., GENTIL, I., LEDOUX, M. ET AL. (2014). *Analysis and geometry of Markov diffusion operators* vol. 103. Springer.

[4] BARDENET, R., DOUCET, A. AND HOLMES, C. (2014). An adaptive subsampling approach for mcmc inference in large datasets. In *Proceedings of The 31st*

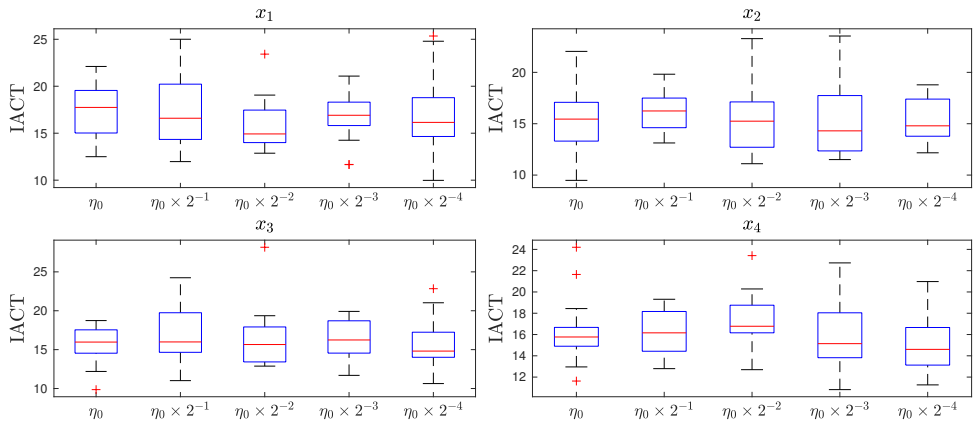

FIGURE 5: Box plots of integrated autocorrelation times (IACTs) of the first four parameters of Markov chains simulated by Algorithm 1. Here 20 realizations of Markov chains are used to estimate the IACT.

International Conference on Machine Learning.

[5] BESKOS, A., JASRA, A., LAW, K., MARZOUK, Y. AND ZHOU, Y. (2018). Multilevel sequential monte carlo with dimension-independent likelihood-informed proposals. *SIAM/ASA Journal on Uncertainty Quantification* **6,** 762–786.

[6] BREYER, L., ROBERTS, G. O. AND ROSENTHAL, J. S. (2001). A note on geometric ergodicity and floating-point roundoff error. *Statistics & probability letters* **53,** 123–127.

[7] BUI-THANH, T., GHATTAS, O., MARTIN, J. AND STADLER, G. (2013). A computational framework for infinite-dimensional bayesian inverse problems part i: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing* **35,** A2494–A2523.

[8] BUTCHER, J. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

[9] CHEN, X., DU, S. S. AND TONG, X. T. (2020). On stationary-point hitting time and ergodicity of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*.

[10] Cotter, S. L., Dashti, M. and Stuart, A. M. (2010). Approximation of bayesian inverse problems for pdes. *SIAM journal on numerical analysis* **48,** 322–345.

[11] Dong, J. and Tong, X. T. (2021). Replica exchange for non-convex optimization. *Journal of Machine Learning Research* **22,** 1–59.

[12] Dong, J. and Tong, X. T. (2022). Spectral gap of replica exchange langevin diffusion on mixture distributions. *Stochastic Processes and their Applications* **151,** 451–489.

[13] Dupuis, P., Liu, Y., Plattner, N. and Doll, J. D. (2012). On the infinite swapping limit for parallel tempering. *Multiscale Modeling & Simulation* **10,** 986–1022.

[14] Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability* **27,** 1551–1587.

[15] Dwivedi, R., Chen, Y., Wainwright, M. J. and Yu, B. (2018). Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on learning theory*. PMLR. pp. 793–797.

[16] Dwivedi, R., Chen, Y., Wainwright, M. J. and Yu, B. (2019). Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research* **20,** 1–42.

[17] Earl, D. J. and Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* **7,** 3910–3916.

[18] Ferré, D., Hervé, L. and Ledoux, J. (2013). Regular perturbation of v-geometrically ergodic markov chains. *Journal of applied probability* **50,** 184–194.

[19] Gallegos-Herrada, M. A., Ledvinka, D. and Rosenthal, J. S. (2022). Equivalences of geometric ergodicity of markov chains. *arXiv preprint arXiv:2203.04395*.

[20] Hairer, M. and Mattingly, J. C. (2008). Spectral gaps in wasserstein distances and the 2d stochastic navier–stokes equations. *The Annals of Probability* **36,** 2050–2091.

[21] Hairer, M., Stuart, A. M. and Vollmer, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability* **24,** 2455–2490.

[22] Hervé, L. and Ledoux, J. (2014). Approximating markov chains and v-geometric ergodicity via weak perturbation theory. *Stochastic Processes and their Applications* **124,** 613–638.

[23] Hoang, V. H., Schwab, C. and Stuart, A. M. (2013). Complexity analysis of accelerated mcmc methods for bayesian inversion. *Inverse Problems* **29,** 085010.

[24] Johndrow, J. E. and Mattingly, J. C. (2017). Coupling and decoupling to bound an approximating markov chain. *arXiv preprint arXiv:1706.02040*.

[25] Johndrow, J. E. and Mattingly, J. C. (2017). Error bounds for approximations of markov chains used in bayesian sampling. *arXiv preprint arXiv:1711.05382*.

[26] Jones, G. L. (2004). On the markov chain central limit theorem. *Probability surveys* **1,** 299–320.

[27] Joulin, A. and Ollivier, Y. (2010). Curvature, concentration and error estimates for markov chain monte carlo. *The Annals of Probability* **38,** 2418–2442.

[28] Lindvall, T. (2002). *Lectures on the coupling method*. Courier Corporation.

[29] Lotka, A. J. (1925). *Elements of physical biology*. Williams & Wilkins.

[30] Medina-Aguayo, F., Rudolf, D. and Schweizer, N. (2020). Perturbation bounds for monte carlo within metropolis via restricted approximations. *Stochastic processes and their applications* **130,** 2200–2227.

[31] Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

[32] Negrea, J. and Rosenthal, J. S. (2021). Approximations of geometrically ergodic reversible markov chains. *Advances in Applied Probability* **53,** 981–1022.

[33] Parno, M. D. and Marzouk, Y. M. (2018). Transport map accelerated markov chain monte carlo. *SIAM/ASA Journal on Uncertainty Quantification* **6,** 645–682.

[34] Pillai, N. S. and Smith, A. (2014). Ergodicity of approximate mcmc chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*.

[35] Roberts, G. O., Rosenthal, J. S. and Schwartz, P. O. (1998). Convergence properties of perturbed markov chains. *Journal of applied probability* **35,** 1–11.

[36] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli* **2,** 341–363.

[37] Rudolf, D. (2012). Explicit error bounds for markov chain monte carlo. *Dissertationes Math 485: 1-93*.

[38] Rudolf, D. and Schweizer, N. (2018). Perturbation theory for markov chains via wasserstein distance. *Bernoulli* **24,** 2610–2639.

[39] Shardlow, T. and Stuart, A. M. (2000). A perturbation theory for ergodic markov chains and application to numerical approximations. *SIAM journal on numerical analysis* **37,** 1120–1137.

[40] Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta numerica* **19,** 451–559.

[41] Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* **314,** 141–151.

[42] Tawn, N. G. and Roberts, G. O. (2019). Accelerating parallel tempering: Quantile tempering algorithm (quanta). *Advances in Applied Probability* **51,** 802–834.

[43] Tawn, N. G., Roberts, G. O. and Rosenthal, J. S. (2020). Weight-preserving simulated tempering. *Statistics and Computing* **30,** 27–41.

[44] Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics* 1701–1728.

[45] Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems* **32,**.

[46] Woodard, D. B., Schmidler, S. C. and Huber, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability* **19,** 617–640.

Appendix A. Proof for the Wasserstein convergence

Proof of Theorem 2.2. Let Q_x be the optimal coupled measure between $\delta_x P$ and $\delta_x \widehat{P}$ from the Kantorovich–Rubinstein theorem. Then,

$$\begin{aligned}|PV(x) - \widehat{P}V(x)| &= \left| \int Q_x(dx', dy')(V(x') - V(y')) \right| \\ &\leq \left| \int Q_x(dx', dy')(V(x') + V(y'))1_{x' \neq y'} \right| = \|\delta_x P - \delta_x \widehat{P}\|_V.\end{aligned}$$

Next, as $\|\delta_x P - \delta_x \widehat{P}\|_V \leq \epsilon V(x)$, we have

$$|PV(x) - \widehat{P}V(x)| \leq \epsilon V(x).$$

In addition, because V is a Lyapunov function under P,

$$\widehat{P}V(x) \leq PV(x) + \epsilon V(x) \leq (\lambda + \epsilon)V(x) + L.$$

As $(\lambda+\epsilon) \in (0, 1)$ for ϵ small enough, V is a Lyapunov function under $\widehat{P}$ with parameters $\lambda + \epsilon$ and L.

We next establish a bound for $\|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V$, $x \neq y$ using

$$\|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V \leq \|\delta_x \widehat{P}^n - \delta_x P^n\|_V + \|\delta_x P^n - \delta_y P^n\|_V + \|\delta_y P^n - \delta_y \widehat{P}^n\|_V. \quad (18)$$

For $\|\delta_x \widehat{P}^n - \delta_x P^n\|_V$, by Theorem 2.1, we have

$$\|\delta_x \widehat{P}^n - \delta_x P^n\|_V \leq \frac{C\epsilon}{1-\rho}\left(V(x) + \frac{L}{1-\lambda-\epsilon}\right) \leq C\epsilon V(x)$$

for some C, because $V(x) \geq 1$. A similar bound holds for $\|\delta_y P^n - \delta_y \widehat{P}^n\|_V$ as well, i.e.,

$$\|\delta_y P^n - \delta_y \widehat{P}^n\|_V \leq \frac{C\epsilon}{1-\rho}\left(V(y) + \frac{L}{1-\lambda-\epsilon}\right) \leq C\epsilon V(y).$$

For any $l \leq N$, Let $D_1 = \frac{C}{N\rho^{2N-1}}$. Then, (18) leads to

$$\begin{aligned} \|\delta_x \widehat{P}^{l+N} - \delta_y \widehat{P}^{l+N}\|_V &\leq \rho^{l+N} d_V(x,y) + C\epsilon(V(x)+V(y)) \\ &= \left(\rho^{l+N} + C\epsilon\right) d_V(x,y) \leq (\rho + D_1\epsilon)^{l+N} d_V(x,y), \end{aligned} \tag{19}$$

where the last inequality follows from $(a+b)^N > a^N + Na^{N-1}b$ for all $a, b > 0$, which comes from Taylor expansion.

Next, let $\widehat{Q}^k_{x,y}$ be the optimal coupled measure between $\delta_x \widehat{P}^{kN}$ and $\delta_y \widehat{P}^{kN}$. Then,

$$\begin{aligned} \|\delta_x \widehat{P}^{kN} - \delta_y \widehat{P}^{kN}\|_V &\leq \int \widehat{Q}^{k-1}_{x,y}(dx', dy') \|\delta_{x'} \hat{P}^N - \delta_{y'} \hat{P}^N\|_V \\ &\leq \int \widehat{Q}^{k-1}_{x,y}(dx', dy')(\rho + D_1\epsilon)^N d_V(x', y') \\ &\leq (\rho + D_1\epsilon)^{kN} d_V(x,y). \end{aligned} \tag{20}$$

For any $n \geq N$, we can write $n = kN + N + l$, for $k, l \in \mathbb{Z}_0^+$, and

$$\begin{aligned} \|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V &\leq \int \widehat{Q}^k_{x,y}(dx', dy') \|\delta_{x'} \hat{P}^{N+l} - \delta_{y'} \hat{P}^{N+l}\|_V \\ &\leq C_1(\rho + D_1\epsilon)^{N+l} \int \widehat{Q}^k_{x,y}(dx', dy') d_V(x', y') \text{ by (19)} \\ &\leq C_1(\rho + D_1\epsilon)^{l+kN+N} d_V(x,y) \text{ by (20)} \\ &= C_1(\rho + D_1\epsilon)^n d_V(x,y). \end{aligned}$$

Lastly, we show that $\widehat{P}$ has a unique invariant measure $\widehat{\pi}$. Fix a point x, consider a sequence $\{\delta_x \widehat{P}^n, n = 1, 2, \ldots\}$. Note that

$$\begin{aligned} \|\delta_x \widehat{P}^n - \delta_x \widehat{P}^{n+1}\|_V &\leq \int \|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V \widehat{P}(x, dy) \\ &\leq C_1(\rho + D_1\epsilon)^n \mathbb{E}[d_V(x, \widehat{X}_1)] \\ &\leq C_1(\rho + D_1\epsilon)^n [(\lambda + 1 + \epsilon)V(x) + L]. \end{aligned}$$

This implies that $\delta_x \widehat{P}^n$ is a Cauchy sequence in d_V and the total variation distance. Since the sublevel sets of V are compact, and $\delta_x \widehat{P}^n V$ remains bounded, it is a tight sequence. Therefore, the sequence has a limit, which we denote by π_x. Next, we show

that $\pi_x = \pi_y$:

$$\|\pi_x - \pi_y\|_V \le \|\pi_x - \delta_x \widehat{P}^n\|_V + \|\delta_y \widehat{P}^n - \pi_y\|_V + \|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V \to 0 \text{ as } n \to 0.$$

□

Proof of Proposition 2.1. For the first claim, let $Q^n_{x,y}$ be the the optimal coupled measure between $\delta_x \widehat{P}^n$ and $\delta_y \widehat{P}^n$ for any $x, y \in \mathbb{R}^d$. Then,

$$\begin{aligned} |\delta_x \widehat{P}^n f - \widehat{\pi} f| &= |\delta_x \widehat{P}^n f - \widehat{\pi} \widehat{P}^n f| \\ &\le \int |\delta_x \widehat{P}^n f - \delta_y \widehat{P}^n f| \widehat{\pi}(dy) \\ &\le \int \int Q^n_{x,y}(dx', dy') |f(x') - f(y')| \widehat{\pi}(dy) \\ &\le \int \|\delta_x \widehat{P}^n - \delta_y \widehat{P}^n\|_V \widehat{\pi}(dy) \\ &\le \widehat{\rho}^n \int d_V(x, y) \widehat{\pi}(dy) \le \widehat{\rho}^n (V(x) + \widehat{\pi} V). \end{aligned} \tag{21}$$

For the second claim, note that for any f with $\widehat{\pi} f = 0$, we have

$$\begin{aligned} \mathbb{E}_{\widehat{\pi}}[(\hat{f}_M - \widehat{\pi} f)^2] &= \frac{1}{M^2} \mathbb{E}_{\widehat{\pi}} \left[\sum_{j,k=1}^{M} f(\widehat{X}_j) f(\widehat{X}_k) \right] \\ &\le \frac{2}{M^2} \mathbb{E}_{\widehat{\pi}} \left[\sum_{j=1}^{M} |f(\widehat{X}_j)| \sum_{k=0}^{\infty} |f(\widehat{X}_{j+k})| \right] \\ &= \frac{2}{M} \mathbb{E}_{\widehat{\pi}} \left[|f(\widehat{X}_0)| \mathbb{E}\Big[\sum_{k=0}^{\infty} |f(\widehat{X}_k)| \Big| \widehat{X}_0 \Big] \right] \text{ since } \widehat{X}_j \sim \widehat{\pi} \\ &\le \frac{2}{M} \mathbb{E}_{\widehat{\pi}} \left[|f(\widehat{X}_0)| \sum_{k=0}^{\infty} \widehat{\rho}^k (V(\widehat{X}_0) + \widehat{\pi} V) \right] \text{ from (21)} \\ &\le \frac{2}{(1 - \widehat{\rho}) M} \mathbb{E}_{\widehat{\pi}} \Big[|f(\widehat{X}_0)| (V(\widehat{X}_0) + \widehat{\pi} V) \Big]. \end{aligned}$$

□

Appendix B. Proof for the χ^2 convergence

Proof of Theorem 3.1. We will write $\kappa = \kappa(P)$ for short.

For Claim 1, first note that

$$\begin{aligned}
\operatorname{var}_\pi(\widehat{P}f) =& \frac{1}{2}\int \left(\widehat{P}f(x) - \widehat{P}f(y)\right)^2 \pi(dx)\pi(dy) \\
=& \frac{1}{2}\int \left(Pf(x) - Pf(y) + (P-\widehat{P})f(y) - (P-\widehat{P})f(x)\right)^2 \pi(dx)\pi(dy) \\
\le& \left(\frac{1}{2} + \frac{a\kappa}{2}\right)\int \left(Pf(x) - Pf(y)\right)^2 \pi(dx)\pi(dy) \\
&+ \left(\frac{1}{2} + \frac{1}{2a\kappa}\right)\int \left((P-\widehat{P})f(x) - (P-\widehat{P})f(y)\right)^2 \pi(dx)\pi(dy) \\
\le& (1+a\kappa)\operatorname{var}_\pi(Pf) + \left(2 + \frac{2}{a\kappa}\right)\int ((P-\widehat{P})f(x))^2 \pi(dx).
\end{aligned} \tag{22}$$

Here we used $AB \le \frac{a\kappa}{2}A^2 + \frac{2}{a\kappa}B^2$ to get the first inequality.

Next, note that

$$(1+a\kappa)\operatorname{var}_\pi(Pf(X)) \le (1+a\kappa)(1-\kappa)\operatorname{var}_\pi f \le (1-(1-a)\kappa)\operatorname{var}_\pi f. \tag{23}$$

Let $\Delta f(x) = f(x) - \pi f$. Then,

$$\mathbb{E}_\pi\left[\left((P-\widehat{P})f(X)\right)^2\right] = \mathbb{E}_\pi\left[\left((P-\widehat{P})\Delta f(X)\right)^2\right] \le \epsilon^2\|\Delta f\|_\pi^2 = \epsilon^2 \operatorname{var}_\pi f. \tag{24}$$

Plugging the bounds (23) and (24) in (22), we have

$$\operatorname{var}_\pi(\widehat{P}f) \le \left(1 - (1-a)\kappa + \frac{(2a\kappa+2)\epsilon^2}{a\kappa}\right)\operatorname{var}_\pi f.$$

Using induction with $C = 2 + 2/\kappa$, we find

$$\operatorname{var}_\pi(\widehat{P}^n f) \le \left(1 - (1-a)\kappa + \frac{C\epsilon^2}{a}\right)^n \operatorname{var}_\pi f.$$

For Claim 2, first note that

$$\begin{aligned}
\left|\int \widehat{P}f(x)\pi(dx) - \pi f\right| &= \left|\int (\widehat{P}-P)f(x)\pi(dx)\right| \\
&= \left|\int (\widehat{P}-P)\Delta f(x)\pi(dx)\right| \text{ recall that } \Delta f(x) = f(x) - \pi f \\
&\le \left(\int \left((\widehat{P}-P)\Delta f(x)\right)^2 \pi(dx)\right)^{1/2} \\
&= \|(P-\widehat{P})\Delta f\|_\pi \le \epsilon\sqrt{\operatorname{var}_\pi f}.
\end{aligned}$$

Then, because $\int Pg(x)\pi(dx) = \int g(x)\pi(dx)$ for any function g,

$$\left|\int (\widehat{P}^{n+1} - \widehat{P}^n) f(x)\pi(dx)\right| = \left|\int (\widehat{P} - P)(\widehat{P}^n f)(x)\pi(dx)\right| \\ \leq \epsilon\sqrt{\mathrm{var}_\pi(\widehat{P}^n f)} \leq \epsilon(1-\widehat{\kappa})^{n/2}\sqrt{\mathrm{var}_\pi f},$$

from Claim 1. Let $\mu_n = \pi\widehat{P}^n$ and $f = \mathrm{sgn}(\mu_{n+1} - \mu_n)$. Because $\mathrm{var}_\pi f \leq \mathbb{E}_\pi |f|^2 = 1$,

$$\|\mu_{n+1} - \mu_n\|_{TV} = \left|\int (\widehat{P}^{n+1} - \widehat{P}^n) f(x)\pi(dx)\right| \leq \epsilon(1-\widehat{\kappa})^{n/2}\sqrt{\mathrm{var}_\pi f} \leq \epsilon(1-\widehat{\kappa})^{n/2}.$$

Thus, μ_n is a Cauchy sequence under the total variation metric, which implies that the sequence has a limit $\widehat{\pi}$ and

$$|(\widehat{\pi} - \mu_n)f| \leq \sum_{k=n}^{\infty} |(\mu_{k+1} - \mu_k)f| \leq \frac{\epsilon(1-\widehat{\kappa})^{n/2}}{1-(1-\widehat{\kappa})^{1/2}}\sqrt{\mathrm{var}_\pi f}$$

When letting $n = 0$, we have

$$|\widehat{\pi}f - \pi f| \leq \frac{\epsilon}{1-(1-\widehat{\kappa})^{1/2}}\sqrt{\mathrm{var}_\pi f} \tag{25}$$

Consider $f = \widehat{\pi}/\pi$. $D_{\chi^2}(\widehat{\pi}\|\pi) = |\widehat{\pi}f - \pi f| = \mathrm{var}_\pi f$. Combine this with (25), we have

$$D_{\chi^2}(\widehat{\pi}\|\pi) \leq \frac{\epsilon^2}{(1-(1-\widehat{\kappa})^{1/2})^2}.$$

□

Proof of Proposition 3.1. For the first claim, we note

$$|\nu\widehat{P}^n f - \pi\widehat{P}^n f| \leq \int \frac{\nu(x)}{\pi(x)}\pi(dx)|\widehat{P}^n f(x) - \pi\widehat{P}^n f| \\ \leq \sqrt{\int \left(\frac{\nu(x)}{\pi(x)}\right)^2 \pi(dx)}\sqrt{\int |\widehat{P}^n f(x) - \pi\widehat{P}^n f|^2\pi(dx)} \\ \leq \sqrt{D_{\chi^2}(\nu\|\pi) + 1} \times (1-\hat{\kappa})^{n/2}\sqrt{\mathrm{var}_\pi f} \text{ by Theorem 3.1.}$$

Meanwhile,

$$|(\widehat{\pi} - \pi\widehat{P}^n)f| \leq C\epsilon(1-\hat{\kappa})^{n/2}\sqrt{\mathrm{var}_\pi f} \text{ by Theorem 3.1.}$$

By triangular inequality,

$$|\nu\widehat{P}^n f - \widehat{\pi}f| \leq |\nu\widehat{P}^n f - \pi\widehat{P}^n f| + |\pi\widehat{P}^n f - \widehat{\pi}f| \\ \leq (1-\widehat{\kappa})^{n/2}\sqrt{\mathrm{var}_\pi f}\left(\sqrt{D_{\chi^2}(\nu\|\pi) + 1} + C\epsilon\right).$$

For the second part, we first note that for any f with $\widehat{\pi} f = 0$, we have

$$\begin{aligned}
\mathbb{E}_{\widehat{\pi}}[(\hat{f}_M - \widehat{\pi} f)^2] &= \frac{1}{M^2}\mathbb{E}_{\widehat{\pi}}\left[\sum_{j,k=1}^{M} f(\widehat{X}_j) f(\widehat{X}_k)\right] \\
&\leq \frac{2}{M^2}\mathbb{E}_{\widehat{\pi}}\left[\sum_{j=1}^{M} |f(\widehat{X}_j)| \sum_{k=0}^{\infty} |f(\widehat{X}_{j+k})|\right] \\
&= \frac{2}{M}\sum_{k=0}^{\infty} \mathbb{E}_{\widehat{\pi}}\left[|f(\widehat{X}_0)||f(\widehat{X}_k)|\right] \\
&\leq \frac{2}{M}\sqrt{\mathbb{E}_{\widehat{\pi}}[f(\widehat{X}_0)^2]}\sum_{k=0}^{\infty}\sqrt{\mathbb{E}_{\widehat{\pi}}[(\widehat{P}^k f(\widehat{X}_0))^2]}.
\end{aligned}$$

Next,

$$\begin{aligned}
\operatorname{var}_{\widehat{\pi}}(\widehat{P}^k f) &= \widehat{\pi}(\widehat{P}^k f)^2 \\
&\leq \widehat{\pi}(\widehat{P}^k f - \pi\widehat{P}^k f)^2 \\
&\leq \pi(\widehat{P}^k f - \pi\widehat{P}^k f)^2 + \frac{\epsilon}{1-(1-\widehat{\kappa})^{1/2}}\sqrt{\operatorname{var}_{\pi}[(\widehat{P}^k f - \pi\widehat{P}^k f)^2]} \text{ by (25)} \\
&\leq (1-\widehat{\kappa})^k \operatorname{var}_{\pi}(f) + \frac{\epsilon}{1-(1-\widehat{\kappa})^{1/2}}\sqrt{\operatorname{var}_{\pi}[(\widehat{P}^k f - \pi\widehat{P}^k f)^2]} \text{ by Theorem 3.1.}
\end{aligned}$$

Because $\sup_x |f(x)| \leq C$ for some $C \in (0,\infty)$, $\sup_x |\widehat{P}^k f(x)| \leq C$ and $\sup_x |(\widehat{P}^k - \pi\widehat{P}^k) f(x)| \leq 2C$. Then,

$$\pi(\widehat{P}^k f - \pi\widehat{P}^k f)^4 \leq 4C^2\pi(\widehat{P}^k f - \pi\widehat{P}^k f)^2 \leq 4C^2(1-\widehat{\kappa})^k \operatorname{var}_{\pi} f \text{ by Theorem 3.1.}$$

Thus,

$$\operatorname{var}_{\widehat{\pi}}(\widehat{P}^k f) \leq (1-\widehat{\kappa})^k \operatorname{var}_{\pi}(f) + \frac{\epsilon}{1-(1-\widehat{\kappa})^{1/2}} 2C(1-\widehat{\kappa})^{\frac{k}{2}}\sqrt{\operatorname{var}_{\pi} f}$$

and we can further find a constant C' such that

$$\mathbb{E}_{\widehat{\pi}}[(\hat{f}_M - \widehat{\pi} f)^2] \leq \frac{C'}{M(1-(1-\widehat{\kappa})^{1/4})}\sqrt{\operatorname{var}_{\widehat{\pi}} f \ \operatorname{var}_{\pi} f}.$$

□

Proof of Theorem 3.2. To simplify the notation, let κ denote the spectral gap of P and $\widehat{\kappa}$ denote the spectral gap of $\widehat{P}$. By the definition of spectral gap, i.e., (14), we have

$$\widehat{\kappa} = \min_f \frac{\langle f, (I - \widehat{P}^2) f\rangle_{\widehat{\pi}}}{\operatorname{var}_{\widehat{\pi}} f}.$$

First, note that

$$\begin{aligned}\operatorname{var}_{\widehat{\pi}} f &= \mathbb{E}_{\widehat{\pi}}[(f-\widehat{\pi}f)^2] \\ &\le \mathbb{E}_{\widehat{\pi}}[(f-\pi f)^2] \\ &\le (1+\epsilon)\mathbb{E}_{\pi}[(f-\pi f)^2] \\ &= (1+\epsilon)\operatorname{var}_{\pi} f.\end{aligned} \tag{26}$$

We next establish two useful bounds:

$$\begin{aligned}|\langle f,(I-P^2)f\rangle_{\widehat{\pi}} - \langle f,(I-P^2)f\rangle_{\pi}| &\le \int |f(x)(I-P^2)f(x)|\epsilon\pi(dx) \\ &\le \epsilon\|f\|_{\pi}\|(I-P^2)f\|_{\pi} \le \epsilon\|f\|_{\pi}^2,\end{aligned} \tag{27}$$

and

$$\begin{aligned}|\langle f,(\widehat{P}^2-P^2)f\rangle_{\widehat{\pi}}| &\le \|f\|_{\widehat{\pi}}\|(\widehat{P}^2-P^2)f\|_{\widehat{\pi}} \\ &\le (1+\epsilon)^2\|f\|_{\pi}\|(\widehat{P}^2-P^2)f\|_{\pi} \\ &\le (1+\epsilon)^2\|f\|_{\pi}(2\|(\widehat{P}-P)Pf\|_{\pi} + \|(\widehat{P}-P)^2f\|_{\pi}) \\ &\le (1+\epsilon)^2\|f\|_{\pi}(2\epsilon\|Pf\|_{\pi} + \epsilon\|(\widehat{P}-P)f\|_{\pi}) \\ &\le (1+\epsilon)^2\|f\|_{\pi}(2\epsilon\|f\|_{\pi} + \epsilon^2\|f\|_{\pi}) \\ &\le 3(1+\epsilon)^2\epsilon\|f\|_{\pi}^2 \\ &\le C\epsilon\|f\|_{\pi}^2.\end{aligned} \tag{28}$$

Then,

$$\begin{aligned}|\langle f,(I-\widehat{P}^2)f\rangle_{\widehat{\pi}}| &\ge |\langle f,(I-P^2)f\rangle_{\widehat{\pi}}| - |\langle f,(\widehat{P}^2-P^2)f\rangle_{\widehat{\pi}}| \text{ by triangular inequality} \\ &\ge |\langle f,(I-P^2)f\rangle_{\widehat{\pi}}| - C\epsilon\|f\|_{\pi}^2 \text{ by the bound in (28)} \\ &\ge |\langle f,(I-P^2)f\rangle_{\pi}| - |\langle f,(I-P^2)f\rangle_{\pi} - \langle f,(I-P^2)f\rangle_{\widehat{\pi}}| - C\epsilon\|f\|_{\pi}^2 \\ &\ge |\langle f,(I-P^2)f\rangle_{\pi}| - C\epsilon\|f\|_{\pi}^2 - C\epsilon\|f\|_{\pi}^2 \text{ by the bound in (27)} \\ &\ge \langle f,(I-P^2)f\rangle_{\pi} - C\epsilon\operatorname{var}_{\pi} f.\end{aligned} \tag{29}$$

Combining (26) and (29), we have $\widehat{\kappa} \ge \kappa - C\epsilon$.

□

Proof of Proposition 3.2. For the first claim,

$$\begin{aligned}
|\nu\widehat{P}^n f - \widehat{\pi} f| &= |\nu\widehat{P}^n f - \widehat{\pi}\widehat{P}^n f| \\
&\leq \int \frac{\nu(x)}{\widehat{\pi}(x)}\widehat{\pi}(x)|\widehat{P}^n f(x) - \widehat{\pi}\widehat{P}^n f|dx \\
&\leq \sqrt{\int \left(\frac{\nu(x)}{\widehat{\pi}(x)}\right)^2 \widehat{\pi}(x)dx}\sqrt{\int |\widehat{P}^n f(x) - \widehat{\pi}\widehat{P}^n f|^2\widehat{\pi}(x)dx} \\
&\leq \sqrt{D_{\chi^2}(\nu\|\widehat{\pi}) + 1} \times (1-\hat{\kappa})^{n/2}\sqrt{\mathrm{var}_{\widehat{\pi}} f}.
\end{aligned}$$

For the second part, we first note that if we replace f with $f - \widehat{\pi} f$, with a little abuse of notation, we have

$$\begin{aligned}
\mathbb{E}_{\widehat{\pi}}[(\hat{f}_M - \widehat{\pi} f)^2] &= \frac{1}{M^2}\mathbb{E}_{\widehat{\pi}}\left[\sum_{j,k=1}^{M} f(\widehat{X}_j)f(\widehat{X}_k)\right] \\
&\leq \frac{2}{M^2}\mathbb{E}_{\widehat{\pi}}\left[\sum_{j=1}^{M}|f(\widehat{X}_j)|\sum_{k=0}^{\infty}|f(\widehat{X}_{j+k})|\right] \\
&= \frac{2}{M}\sum_{k=0}^{\infty}\mathbb{E}_{\widehat{\pi}}\left[|f(\widehat{X}_0)||f(\widehat{X}_k)|\right] \\
&\leq \frac{2}{M}\sqrt{\mathbb{E}_{\widehat{\pi}}[f(\widehat{X}_0)^2]}\sum_{k=0}^{\infty}\sqrt{\mathbb{E}_{\widehat{\pi}}[(\widehat{P}^k f(\widehat{X}_0))^2]} \\
&\leq \frac{2}{M}\sqrt{\mathbb{E}_{\widehat{\pi}}[f(\widehat{X}_0)^2]}\sum_{k=0}^{\infty}(1-\widehat{\kappa})^{k/2}\sqrt{\mathrm{var}_{\widehat{\pi}} f} \\
&= \frac{2}{M(1-(1-\widehat{\kappa})^{1/2})}\,\mathrm{var}_{\widehat{\pi}} f
\end{aligned}$$

□

Proof of Proposition 3.3. Let Q_x be the optimal coupled measure between $\delta_x P$ and $\delta_x\widehat{P}$. Then

$$\begin{aligned}
|(\delta_x P - \delta_x\widehat{P})f| &\leq \int Q_x(dx', dy')|f(x') - f(y')| \\
&= \int Q_x(dx', dy')(|f(x') - f(y')|)1_{x'\neq y'}.
\end{aligned}$$

Next,

$$\begin{aligned}
\|(P-\hat{P})f\|_\pi^2 &= \int \pi(dx)|(\delta_x P - \delta_x \widehat{P})f|^2 \\
&\leq \int \pi(dx) \left(\int Q_x(dx', dy')(|f(x') - f(y')|) 1_{x' \neq y'} \right)^2 \\
&\leq \left(\int \pi(dx) Q_x(dx', dy') \left(2f(x')^2 + 2f(y')^2 \right) \right) \left(\int \pi(dx) Q_x(dx', dy') 1_{x' \neq y'} \right) \\
&\leq 2 \left(\langle \pi P, f^2 \rangle + \langle \pi \widehat{P}, f^2 \rangle \right) \left(\epsilon \int \pi(dx) V(x) \right) \\
&\leq 2\epsilon \left(\langle \pi, f^2 \rangle + a \langle \widehat{\pi} \widehat{P}, f^2 \rangle \right) (\pi V) \\
&\leq 2\epsilon \left(1 + a^2 \right) \|f\|_\pi^2 \|V\|_\pi.
\end{aligned}$$

□

Appendix C. Verification for Metropolis–Hastings MCMC

We first present an auxiliary lemma, which is well-known:

Lemma C.1. *Suppose a transition kernel P is and reversible with invariant measure π. Then, $\|P\|_\pi \leq 1$.*

Proof. We first note that

$$\begin{aligned}
&\int f(x)^2 \pi(dx) - \int \pi(dx) f(x) f(y) P^2(x, dy) \\
=&\frac{1}{2} \int \pi(dx) f(x)^2 P^2(x, dy) + \frac{1}{2} \int \pi(dx) f(y)^2 P^2(x, dy) - \int \pi(dx) f(x) f(y) P^2(x, dy) \\
=&\frac{1}{2} \int \pi(dx) (f(x) - f(y))^2 P^2(x, dy) \geq 0.
\end{aligned}$$

Thus,

$$\int f(x)^2 \pi(dx) \geq \int \pi(dx) f(x) f(y) P^2(x, dy) = \|Pf\|_\pi^2.$$

□

Proof of Lemma 4.1. For any density of form $\mu(x) = \nu(x)s(x)$, we have

$$\begin{aligned}
|\mu(P-\widehat{P})f| &\le \int \mu(x)|\alpha(x)-\hat{\alpha}(x)||f(x)|dx + \int \mu(x)|\beta(x,x')-\hat{\beta}(x,x')||f(x')|dx'dx \\
&\le C\epsilon \int \mu(x)|f(x)|dx + C\epsilon \int \mu(x)\beta(x,x')|f(x')|dx'dx \\
&\le C\epsilon \int \mu(x)|f(x)|dx + C\epsilon \int \mu(x)P(x,x')|f(x')|dx'dx \\
&= C\epsilon \int \nu(x)s(x)|f(x)|dx + C\epsilon \int \nu(x)s(x)P(x,x')|f(x')|dx'dx \\
&\le C\epsilon\|s\|_\nu\|f\|_\nu + C\epsilon\|s\|_\nu\|Pf\|_\nu \\
&\le 2C\epsilon\|s\|_\nu\|f\|_\nu \text{ by Lemma C.1.}
\end{aligned}$$

Next, take $\mu \propto |(P-\widehat{P})f|\nu$, we have

$$\|(P-\widehat{P})f\|_\nu^2 \le 2C\epsilon\|(P-\widehat{P})f\|_\nu\|f\|_\nu,$$

which further implies that there is a C_1, so that $\|P-\widehat{P}\|_\nu \le C_1\epsilon$. □

Proof of Proposition 4.1. We denote the acceptance probabilities for the original process and perturbed process as

$$b(x,x') = \frac{\pi(x')}{\pi(x)} \wedge 1 \text{ and } \hat{b}(x,x') = \frac{\hat{\pi}(x')}{\hat{\pi}(x)} \wedge 1$$

respectively. Since for any positive numbers a, b, c, d,

$$\min\{a/b, c/d\} \le \frac{a\wedge c}{b\wedge d} \le \max\{a/b, c/d\},$$

and since $\exp(-C\epsilon) \le \pi(x)/\hat{\pi}(x) \le \exp(C\epsilon)$, we have

$$\exp(-2C\epsilon)b(x,x') < \hat{b}(x,x') < \exp(2C\epsilon)b(x,x').$$

Using the fact that $\beta(x,x') = R(x,x')b(x,x')$ and $\hat{\beta}(x,x') = R(x,x')\hat{b}(x,x')$, we have

$$\exp(-2C\epsilon)\beta(x,x') < \hat{\beta}(x,x') < \exp(2C\epsilon)\beta(x,x').$$

In addition, for $\alpha(x) = \int R(x,x')(1-b(x,x'))dx'$ and $\hat{\alpha}(x) = \int R(x,x')(1-\hat{b}(x,x'))dx'$,

$$|\alpha(x)-\hat{\alpha}(x)| \le \int R(x,x')|b(x,x')-\hat{b}(x,x')|dx' \le C\epsilon.$$

By Lemma 4.1, we can find a C_1 so that

$$\|P_{RWM} - \widehat{P}_{RWM}\|_\pi \le C_1\epsilon.$$

□

Proof of Proposition 4.2. Note that

$$R(x,x') = \frac{1}{(4\pi h)^{d/2}} \exp\left(-\frac{1}{4h}\|x' - x - \nabla \log \pi(x) h\|^2\right),$$

and

$$\widehat{R}(x,x') = \frac{1}{(4\pi h)^{d/2}} \exp\left(-\frac{1}{4h}\|x' - x - \nabla \log \hat{\pi}(x) h\|^2\right).$$

As $|\nabla \log \widehat{\pi}(x) - \nabla \log \pi(x)| \le C\epsilon$ and the support is bounded, we can enlarge the value of C so that

$$(1 - C\epsilon)R(x,x') \le \widehat{R}(x,x') \le (1 + C\epsilon)R(x,x'),$$

Let the acceptance probability be

$$\begin{aligned} b(x,x') =& \frac{\pi(x') \exp\left(-\frac{1}{4h}\|x - x' + \nabla \log \pi(x') h\|^2\right)}{\pi(x) \exp\left(-\frac{1}{4h}\|x' - x + \nabla \log \pi(x) h\|^2\right)} \wedge 1 \\ =& \left\{ \exp\left(\log \pi(x') - \log \pi(x) - \frac{1}{2}\langle x - x', \nabla \log \pi(x') - \nabla \log \pi(x)\rangle\right)\right. \\ & \left.\times \exp\left(\frac{h}{4}\left[\|\nabla \log \pi(x')\|^2 - \|\nabla \log \pi(x)\|^2\right]\right)\right\} \wedge 1 \end{aligned}$$

Similarly, we define

$$\begin{aligned} \hat{b}(x,x') =& \frac{\hat{\pi}(x') \exp\left(-\frac{1}{4h}\|x - x' + \nabla \log \hat{\pi}(x') h\|^2\right)}{\hat{\pi}(x) \exp\left(-\frac{1}{4h}\|x' - x + \nabla \log \hat{\pi}(x) h\|^2\right)} \wedge 1 \\ =& \left\{ \exp\left(\log \hat{\pi}(x') - \log \hat{\pi}(x) - \frac{1}{2}\langle x - x', \nabla \log \hat{\pi}(x') - \nabla \log \hat{\pi}(x)\rangle\right)\right. \\ & \left.\times \exp\left(\frac{h}{4}\left[\|\nabla \log \hat{\pi}(x')\|^2 - \|\nabla \log \hat{\pi}(x)\|^2\right]\right)\right\} \wedge 1 \end{aligned}$$

Since $|\log \pi(x) - \log \widehat{\pi}_k(x)| \le C\epsilon$, $\|\nabla \log \pi(x) - \nabla \log \widehat{\pi}_k(x)\| \le C\epsilon$, and the support is bounded, we can further enlarge C, such that

$$(1 - C\epsilon)b(x,x') \le \hat{b}(x,x') \le (1 + C\epsilon)b(x,x').$$

Lastly, for $\beta(x,x') = R(x,x')b(x,x')$ and $\hat{\beta}(x,x') = \hat{R}(x,x')\hat{b}(x,x')$,

$$(1 - C\epsilon)\beta(x,x') \le \hat{\beta}(x,x') \le (1 + C\epsilon)\beta(x,x').$$

In addition, for $\alpha(x) = \int R(x,x')(1 - b(x,x'))dx'$ and $\hat{\alpha}(x) = \int \hat{R}(x,x')(1 - \hat{b}(x,x'))dx'$,

$$\begin{aligned} |\alpha(x) - \hat{\alpha}(x)| \le& \int |R(x,x') - \hat{R}(x,x')|(1 - b(x,x'))dx' \\ &+ \int \hat{R}(x,x')|b(x,x') - \hat{b}(x,x')|dx' \\ \le& C\epsilon \int R(x,x')dx' + C\epsilon \int \hat{R}(x,x')dx' = 2C\epsilon \end{aligned}$$

By Lemma 4.1, we have a constant C_1 so that

$$\|P_{MALA} - \widehat{P}_{MALA}\|_\pi \leq C_1\epsilon.$$

□

Proof of Proposition 4.3. The transition kernel of MALA takes the form

$$P(x,y) = \alpha(x)\delta_x(y) + \beta(x,y)$$

where

$$\beta(x,y) = \frac{\pi(y)}{\pi(x)}q(y,x) \wedge q(x,y) = a(x,y) \wedge q(x,y),$$

with

$$q(x,y) = \frac{1}{(2\pi h)^{d/2}} \exp\left(-\frac{1}{4h}\|y - x - \nabla \log \pi(x)h\|^2\right), a(x,y) = \frac{\pi(y)}{\pi(x)}q(y,x),$$

and $\alpha(x) = 1 - \int \beta(x,dy)$. Similarly, we can write $\widehat{P}(x,y) = \widehat{\alpha}(x)\delta_x(y) + \widehat{\beta}(x,y)$ when using the perturbed target density $\widehat{\pi}$.

We prove the proposition by showing that

$$\int |q(x,y) - \hat{q}(x,y)|dy \leq C\epsilon \text{ and } \int |a(x,y) - \widehat{a}(x,y)|dy \leq C\epsilon \exp(\delta x^2). \tag{30}$$

In particular, note that $a \wedge q - \widehat{a} \wedge \widehat{q} \in \{a - \hat{a}, q - \hat{q}, a - \hat{q}, \hat{a} - q\}$, which further implies that

$$|a \wedge q - \widehat{a} \wedge \widehat{q}| \leq |a - \widehat{a}| + |q - \widehat{q}|.$$

Thus, if the bounds in (30) hold, then

$$\begin{aligned}
\|\delta_x P - \delta_x \widehat{P}\|_{TV} &= \int |\beta(x,y) - \widehat{\beta}(x,y)|dy + |\alpha(x) - \widehat{\alpha}(x)| \\
&\leq 2\int |\beta(x,y) - \widehat{\beta}(x,y)|dy \\
&\leq 2\int |a(x,y) - \widehat{a}(x,y)|dy + 2\int |q(x,y) - \widehat{q}(x,y)|dy \\
&\leq 2C\epsilon(1 + \exp(\delta x^2)).
\end{aligned}$$

In order to obtain the first part of (30), note that by intermediate value theorem,

$|\exp(a)-\exp(b)| \le |\exp(a)+\exp(b)||a-b|$ holds for any a, b, so we can bound

$$\begin{aligned}|q(x,y)-\hat{q}(x,y)| \le & \frac{1}{4}|q(x,y)+\hat{q}(x,y)|\|\nabla \log \pi(x)-\nabla \log \widehat{\pi}(x)\| \\ & \left(\|y-x-h\nabla \log \pi(x)\|+\|y-x-h\nabla \log \widehat{\pi}(x)\|\right) \\ \le & \frac{C\epsilon}{4}|q(x,y)+\hat{q}(x,y)|\left(\|y-x-h\nabla \log \pi(x)\|+\|y-x-h\nabla \log \widehat{\pi}(x)\|\right).\end{aligned} \tag{31}$$

Note that $q(x,y)$ is the proposal density of y. Thus,

$$\begin{aligned}&\int q(x,y)\left(\|y-x-h\nabla \log \pi(x)\|+\|y-x-h\nabla \log \widehat{\pi}(x)\|\right)dy \\ &\le \int q(x,y)\left(2\|y-x-h\nabla \log \pi(x)\|+Ch\epsilon\right)dy \\ &\le Ch\epsilon+2\sqrt{\int q(x,y)\|y-x-h\nabla \log \pi(x)\|^2 dy}=Ch\epsilon+2\sqrt{2hd}.\end{aligned}$$

Similarly,

$$\int \hat{q}(x,y)\left(\|y-x-h\nabla \log \pi(x)\|+\|y-x-h\nabla \log \widehat{\pi}(x)\|\right)dy \le Ch\epsilon+2\sqrt{2hd}.$$

Therefore, we use (31) and find a larger C so that

$$\int |q(x,y)-\hat{q}(x,y)|dy \le C\epsilon.$$

To handle the second part of (30), we use $|\exp(a)-\exp(b)| \le |\exp(a)+\exp(b)||a-b|$ again and find

$$\begin{aligned}&|a(x,y)-\widehat{a}(x,y)| \\ =&\left|\frac{\pi(y)}{\pi(x)}q(y,x)-\frac{\widehat{\pi}(y)}{\widehat{\pi}(x)}\hat{q}(y,x)\right| \\ \le&\frac{1}{4}\left|\frac{\pi(y)}{\pi(x)}q(y,x)+\frac{\widehat{\pi}(y)}{\widehat{\pi}(x)}\hat{q}(y,x)\right|\Bigg(\left|\log \pi(x)-\log \widehat{\pi}(x)\right|+\left|\log \pi(y)-\log \widehat{\pi}(y)\right| \\ &+\|\nabla \log \pi(y)-\nabla \log \widehat{\pi}(y)\|\left(\|y+h\nabla \log \pi(y)-x\|+\|y+h\nabla \log \widehat{\pi}(y)-x\|\right)\Bigg) \\ \le&\frac{C\epsilon}{4}\left|\frac{\pi(y)}{\pi(x)}q(y,x)+\frac{\widehat{\pi}(y)}{\widehat{\pi}(x)}\hat{q}(y,x)\right|\left(2+\left(\|y+h\nabla \log \pi(y)-x\|+\|y+h\nabla \log \widehat{\pi}(y)-x\|\right)\right).\end{aligned}$$

Note that the first part can be bounded by

$$\begin{aligned}
&\int \frac{\pi(y)}{\pi(x)} q(y,x)\left(C + (\|y + h\nabla \log \pi(y) - x\| + \|y + h\nabla \log \widehat{\pi}(y) - x\|\right) dx \\
\leq &\int \frac{\pi(y)}{\pi(x)} q(y,x)\left(C + h\epsilon + 2\|y + h\nabla \log \pi(y) - x\|\right) dx \\
= &\int \frac{\pi(y)}{(2\pi h)^{d/4}\pi(x)} \sqrt{q(y,x)} \cdot (2\pi h)^{d/4}\sqrt{q(y,x)}\left(C + h\epsilon + 2\|y + h\nabla \log \pi(y) - x\|\right) dx
\end{aligned}$$

For $\frac{\pi(y)}{\pi(x)}(2\pi h)^{-d/4}\sqrt{q(y,x)}$, we can bound it by

$$\begin{aligned}
&\frac{1}{(2\pi h)^{d/4}} \frac{\pi(y)}{\pi(x)} \sqrt{q(y,x)} \\
= &\frac{1}{(2\pi h)^{d/2}} \exp\left(\log \pi(y) - \log \pi(x) - \frac{1}{8h}\|y + h\nabla \log \pi(y) - x\|^2\right) \\
\leq &\frac{1}{(2\pi h)^{d/2}} \exp\left(\langle \nabla \log \pi(w), y - x\rangle - \frac{1}{8h}\|y - x\|^2 - \frac{1}{4}\langle \nabla \log \pi(y), y - x\rangle\right) \text{ for some } w \\
\leq &\frac{1}{(2\pi h)^{d/2}} \exp\left(\frac{5L_\pi}{4}\|y - x\|(\|x\| + \|y - x\| + C) - \frac{1}{8h}\|y - x\|^2\right) \text{ by Lipschitzness of } \nabla \log \pi \\
\leq &\frac{1}{(2\pi h)^{d/2}} \exp\left(\left(\frac{5L_\pi}{16\delta} + \frac{5L_\pi}{4} - \frac{1}{8h}\right)\|y - x\|^2 + \delta\|x\|^2 + 10L_\pi^2 C^2\right) \\
\leq &\frac{1}{(2\pi h)^{d/2}} \exp\left(-\frac{1}{16h}\|y - x\|^2 + \delta\|x\|^2 + 10L_\pi^2 C^2\right) \text{ as } h < (\tfrac{5L_\pi}{\delta} + 20L_\pi)^{-1}.
\end{aligned}$$

For $(2\pi h)^{d/4}\sqrt{q(y,x)}\left(C + h\epsilon + 2\|y + h\nabla \log \pi(y) - x\|\right)$, first note that we can find a larger C so that

$$\begin{aligned}
&(2\pi h)^{d/4}\sqrt{q(y,x)}\|y + h\nabla \log \pi(y) - x\| \\
= &\exp\left(-\frac{1}{8h}\|y + h\nabla \log \pi(y) - x\|^2\right)\|y + h\nabla \log \pi(y) - x\| \leq C.
\end{aligned}$$

Combining these two upper bounds, we can find a C_1 so that

$$\int \frac{\pi(y)}{\pi(x)} q(y,x)\left(C + h\epsilon + 2\|y + h\nabla \log \pi(y) - x\|\right) dx \leq C_1 \exp(\delta\|x\|^2).$$

Similarly, we can show that

$$\begin{aligned}
&\int \frac{\widehat{\pi}(y)}{\widehat{\pi}(x)} \hat{q}(y,dx)\left(C + (\|y + h\nabla \log \pi(y) - x\| + \|y + h\nabla \log \widehat{\pi}(y) - x\|\right) \\
\leq &\int \frac{\widehat{\pi}(y)}{\widehat{\pi}(x)} \hat{q}(y,dx)\left(C + h\epsilon + 2\|y + h\nabla \log \widehat{\pi}(y) - x\|\right) \leq C\exp(\delta\|x\|^2).
\end{aligned}$$

Thus, $\int |a(x,y) - \widehat{a}(x,y)|dy \leq C\epsilon\exp(\delta x^2)$ for some C. This concludes the proof of (30) and our claim. □

Appendix D. Verification for the parallel tempering algorithm

Proof of Lemma 5.1. For claim 1), note that for any $\|f\|_\nu \le 1$,

$$\begin{aligned}\|RSf - \widehat{R}\widehat{S}f\|_\nu &\le \|R(S-\widehat{S})f\|_\nu + \|(R-\widehat{R})\widehat{S}f\|_\nu \\ &\le \|(S-\widehat{S})f\|_\nu + C\epsilon\|\widehat{S}f\|_\nu \text{ by Lemma C.1} \\ &\le C\epsilon\|f\|_\nu + C\epsilon\|(\widehat{S}-S)f\|_\nu + C\epsilon\|Sf\|_\nu \\ &\le (2C + C^2\epsilon)\epsilon\|f\|_\nu \text{ by Lemma C.1.}\end{aligned}$$

For claim 2), we first note that

$$R_1 \otimes R_2 = (R_1 \otimes I)(I \otimes R_2)$$

We will show that

$$\|(R_1 \otimes I) - (\widehat{R}_1 \otimes I)\|_\nu = \|((R_1 - \widehat{R}_1) \otimes I)\|_\nu \le C\epsilon.$$

For any $f(x,y)$, define

$$g(x,y) := ((R_1 - \widehat{R}_1) \otimes I)f(x,y)$$

Then for each fixed y, since $\|R_1 - \widehat{R}_1\|_{\nu_1} \le C\epsilon$,

$$\int g(x,y)^2\nu_1(x)dx \le C^2\epsilon^2 \int f(x,y)^2\nu_1(x)dx$$

Thus,

$$\|g\|_\nu^2 = \int g(x,y)^2\nu_1(x)\nu_2(y)dxdy \le C^2\epsilon^2 \int\int f(x,y)^2\nu_1(x)\nu_2(y)dxdy = C^2\epsilon^2\|f\|_\nu^2.$$

Similarly, we can show that

$$\|(I \otimes R_2) - (I \otimes \widehat{R}_2)\|_\nu = \|I \otimes (R_2 - \widehat{R}_2)\|_\nu \le C\epsilon.$$

From claim 1), we can find a C' so that

$$\|R_1 \otimes R_2\|_\nu = \|(R_1 \otimes I)(I \otimes R_2)\|_\nu \le C'\epsilon.$$

For claim 3), by triangular inequality, we have

$$\|U - \widehat{U}\|_\nu \le \frac{1}{n}\sum_{i=1}^n \|S_i - \widehat{S}_i\|_\nu \le C\epsilon.$$

□

Proof of Lemma 5.2. Denote $x' = S(x)$. For any density of form $\mu(x) = s(x)\nu(x)$, we have

$$\begin{aligned}
|\mu(Q - \widehat{Q})f| &\leq \int \mu(x)|a(x,x') - \widehat{a}(x,x')||f(x)|dx + \int \mu(x)|a(x,x') - \widehat{a}(x,x')||f(x')|dx \\
&\leq C\epsilon \int \mu(x)|f(x)|dx + C\epsilon \int \mu(x)a(x,x')|f(x')|dx \\
&\leq C\epsilon \int s(x)\nu(x)|f(x)|dx + C\epsilon \int s(x)\nu(x)(Q(x,x)|f(x)| + Q(x,x')|f(x')|)dx \\
&\leq C\epsilon \|s\|_\nu \|f\|_\nu + C\epsilon \|s\|_\nu \|Q|f|\|_\nu \\
&\leq 2C\epsilon \|s\|_\nu \|f\|_\nu.
\end{aligned}$$

Taking $\mu(x) \propto |(Q - \widehat{Q})f(x)|\nu(x)$, we have the result. □

Proof of Proposition 5.1. Recall that

$$P = MQ, \quad M = (M_0 \otimes \cdots \otimes M_K), \quad Q = \left(\frac{1}{K} \sum_{k \in \{0,\ldots,K-1\}} Q_{k,k+1} \right).$$

and

$$\widehat{P} = \widehat{M}\widehat{Q}, \quad \widehat{M} = \left(\widehat{M}_0 \otimes \cdots \otimes \widehat{M}_K\right), \quad \widehat{Q} = \left(\frac{1}{K} \sum_{k \in \{0,\ldots,K-1\}} \widehat{Q}_{k,k+1} \right).$$

Since M is a product of M_k, Lemma 5.1 claim 2) indicates that $\|M - \widehat{M}\|_\Pi \leq C_1\epsilon$ for some C_1. Then note that if $a \leq CA, b \leq CB$ then $\min\{a,b\} \leq C\min\{A,B\}$ so the acceptance probability of $Q_{k,k+1}$ and $\widehat{Q}_{k,k+1}$ satisfies

$$\frac{\alpha_k(x,x')}{\widehat{\alpha}_k(x,x')} \leq \sup_{x,x'} \left\{ \frac{\pi_k(x')\pi_{k+1}(x)\widehat{\pi}_k(x)\widehat{\pi}_{k+1}(x')}{\widehat{\pi}_k(x')\widehat{\pi}_{k+1}(x)\pi_k(x)\pi_{k+1}(x')} \right\} \leq (1 + C_1\epsilon)^4 \leq 1 + D\epsilon$$

for some constant D. Then Lemma 5.2 indicates that $\|Q_{k,k+1} - \widehat{Q}_{k,k+1}\|_\Pi \leq C_2\epsilon$ for some C_2. Then Lemma 5.1 claim 3) indicates that for some C_3

$$\|Q - \widehat{Q}\|_\Pi \leq C_3\epsilon.$$

Finally, we use claim 1) from Lemma 5.1 and find that $\|P - \widehat{P}\|_\Pi \leq C'\epsilon$ for some C'. □

Acknowledgement

We thank Daniel Rudolf for discussing of some the results and providing us with some references.

Funding information

TC is supported by the Australian Research Council grant DP210103092. JD is supported by NSF grant DMS-1720433. AJ is supported by KAUST baseline funding. XT is supported by the Singapore Ministry of Education (MOE) grant R-146-000-292-114.