Error Bounds for Approximations of Geometrically Ergodic Markov Chains

Jeffrey Negrea and Jeffrey S. Rosenthal

Department of Statistical Sciences, University of Toronto

(February 23, 2017)

Abstract

A common tool in the practice of Markov Chain Monte Carlo is to use approximating transition kernels to speed up computation when the true kernel is slow to evaluate. A relatively limited set of quantitative tools exist to determine whether the performance of such approximations will be well behaved and to assess the quality of approximation. We derive a set a tools for such analysis based on the Hilbert space generated by the stationary distribution we intend to sample, $L_2(\pi)$. The focus of our work is on determining whether the approximating kernel (i.e. perturbation) will preserve the geometric ergodicity of the chain, and whether the approximating stationary distribution will be close to the original stationary distribution. Our results directly generalise the results of [JMMD15] from the uniformly ergodic case to the geometrically ergodic case. We then apply our results to the class of 'Noisy MCMC' algorithms.

1 Introduction

The use of Markov Chain Monte Carlo (MCMC) arises from the need to sample from probabilistic models when simple Monte Carlo is not possible. The procedure is to simulate a positive recurrent Markov process where the stationary distribution is the model one intends to sample, so that the dynamics of the process converge to the distribution required. Temporally correlated samples may then be used to approximate the computation of various expectations; see e.g. [BGJM11] and the many references therein. Examples of common applications may be found in Hierarchical Models, Spatio-Temporal Models, Random Networks, Finance, Bionformatics, etc.

Often, however, the transition dynamics of the Markov Chain required to run this process exactly are too computationally expensive, either due to prohibitively large datasets, intractable likelihoods, etc. In such cases it is tempting to instead *approximate* the transition dynamics of the Markov process in question, either deterministically as in the Low-Rank Gaussian Approximation from [JMMD15], or stochastically as in the Noisy Metropolis Hastings procedure from [AFEB16]. It is important then to understand whether these approximations will yield stable and reliable results. This paper aims to provide quantitative tools for the analysis of these algorithms. Since the use of approximation for the transition dynamics may be interpreted as a *perturbation* of the transition kernel of the exact MCMC algorithm, we focus on bounds on the convergence of perturbations of Markov chains.

The primary purpose of this paper is to extend existing quantitative bounds on the errors of approximate Markov chains from the uniformly ergodic case in [JMMD15] to the geometrically ergodic case (a weaker condition, for which multiple equivalent definitions may be found in [RR97]). Our work will replicate all of the theorectical results of [JMMD15], replacing the total variation metric with L_2 distances, and relaxing the uniform contraction condition to $L_2(\pi)$-geometric ergodicity. In exchange, our results require that the approximating kernel be close in the operator norm induced by $L_2(\pi)$, which is more restrictive than the total variation closeness required by [JMMD15]. Thus, this paper's assumptions are not uniformly weaker nor stronger than those in [JMMD15].

1.1 Geometric Ergodicity

Since our results apply to geometrically ergodic Markov chains, we briefly digress to motivate the notion of geometric ergodicity and its usefulness in MCMC. When analysing the performance of exact MCMC algorithms, it is natural to decompose the error in approximation of expectations into a component for the 'burn-in' of the stochastic process and one for the Monte-Carlo approximation error. The former may be interpreted as the bias due to not having started the process in the stationary distribution. The geometric ergodicity condition essentially dictates that the 'burn-in' error of the n^{th} sample is $C\rho^n$ for some $0 < \rho < 1$, where the constant C depends on the

(suitable) initial distribution. (The chain is *uniformly* ergodic if C can be chosen independently of the initial distribution.) Geometric ergodicity is a desirable property as it ensures that cumulative 'burn-in' error asymptotically does not dominate the Monte-Carlo error, while being less restrictive than the uniform ergodicity condition.

When using approximate MCMC methods, one desires that the approximation preserves geometric ergodicity, so that convergence is still efficient and the 'burn-in' error goes to zero quickly.

1.2 Outline of the Paper

The outline of this paper is as follows. Section 2 reviews previous related work. Then Section 3 contains our main theoretical results and their proofs. Proposition 3 there demonstrates sufficient conditions for the stationary distribution of the perturbed chain to be a member of $L_2(\pi)$, and the resulting L_2 bound is strengthened in Proposition 5. Proposition 7 shows that the perturbed chain is $L_2(\pi)$-geometrically ergodic, and provides an associated geometric decay rate. Theorem 8 combines these results to give tight $L_2(\pi)$ bounds. Then, in Theorem 11, we provide sufficient conditions for the perturbed chain to also be L_1 and $L_2(\pi_\epsilon)$-geometrically ergodic (where π_ϵ is the stationary distribution of the perturbed chain), with the same geometric decay rate from Proposition 7. The remainder of Section 3 establishes the analogues of the main results from [JMMD15] in our geometrically ergodic context.

Finally, Section 4 considers Noisy Metropolis-Hastings algorithms. It provides sufficient conditions for our results from Section 3 to hold for this class of algorithms, in terms of bounding the operator norm of the differences (Proposition 23) and the estimation error bounds (Theorem 24).

2 Previous Related Work

We first present a brief review of other related work, discussing convergence of perturbed Markov chains in the uniformly ergodic and geometrically ergodic cases with varying metrics and additional assumptions.

Close to the present paper, Johndrow et al. [JMMD15] derive perturbation bounds to assess the robustness of approximate MCMC algorithms. The assumptions upon which their results rely are: the original chain is uniformly contractive in the total variation norm (this implies uniform ergodicity); and the perturbation is sufficiently small (in the operator norm induced by the total variation norm). The main results of their paper are: the perturbed kernel is uniformly contractive in the total variation norm; the perturbed stationary distribution is close to the original stationary distribution in total variation; explicit bounds on the total variation distance between finite time approximate sampling distributions and the original stationary distribution; explicit bounds on total variation difference between the original stationary distribution and the mixture of finite time approximate sampling distributions; and explicit bounds on the MSE for integral approximation using approximate kernel and the true kernel. The results derived by [JMMD15] are applied within the same paper to a wide variety of approximate MCMC problems including low rank approximation to Gaussian processes and subsampling approximations.

Further results on perturbations for uniformly ergodic chains may be found in Mitrophanov [Mit05]. This work is motivated in part by numerical rounding errors. Various applications of these results may be found in [AFEB16]. The only assumption of [Mit05] is that the original chain is uniformly ergodic. The paper is unique in that it makes no assumption regarding the proximity of the original and perturbed kernel, though the level of approximation error does still scale linearly with the total variation distance of the original and perturbed kernels. The main results are: explicit bounds on the total variation distance between finite time sampling distributions; and explicit bounds on the total variation distance between stationary distributions.

The work of Roberts et al. [RRS98] (see also [BRR01]) is also motivated by numerical rounding errors. The perturbed kernel is assumed to be derived from the original kernel by a *round-off function*, which e.g. maps the input to nearest multiple of 2^{-31}. In such cases, the new state space is at most countable while old state space may have been uncountable and so the resulting chains have mutually singular marginal distributions at all finite times and mutually singular stationary distributions (if they have stationary distributions at all). The results of [RRS98] require the analysis of Lyapunov drift conditions and drift functions (which we will avoid by working in an ap-

propriate L_2 space). The key assumptions in [RRS98] are: the original kernel is geometrically ergodic, and V is a Lyapunov drift function for the original kernel; the original and perturbed transition kernels are close in the V-norm; the perturbed kernel is defined via a round-off function with round-off error uniformly sufficiently small; and $\log V$ is uniformly continuous. The main results of the paper are: if the perturbed kernel is sufficiently close in the V-norm then geometric ergodicity is preserved; if the drift function, V, can be chosen so that $\log V$ is uniformly continuous and if the round-off errors can be made arbitrarily small then the kernels can be made arbitrarily close in the V-norm; explicit bounds on the total variation distance between the approximate finite-time sampling distribution and the true stationary distribution; and sufficient conditions for the approximating stationary distribution to be arbitrarily close in total variation to the true stationary distribution.

Pillai and Smith [PS14] provide bounds in terms of the Wasserstein topology (cf. [Gib04]). Their main focus is on approximate MCMC algorithms, especially approximation due to subsampling from a large dataset (e.g., when computing the posterior density). Their underlying assumtions are: the original and perturbed kernels satisfy a series of *drift-like conditions* with shared parameters; the original kernel has finite eccentricity for all states (where eccentricity of a state is defined as the expected distance between the state and a sample from the stationary distribution); the *Ricci curvature* of the original kernel has a non-trivial uniform lower bound on a positive measure subset of the state space; and the transition kernels are close in the Wasserstein metric, uniformly on the mentioned subset. Their main results under these assumptions are: explicit bounds on the Wasserstein distance between the approximate sampling distribution and the original stationary distribution; explicit bounds on the total variation distance of the original and perturbed stationary distributions and bounds on the mixing times of each chain; explicit bounds on the bias and L_1 error of Monte Carlo approximations; decomposition of the error from approximate MCMC estimation into components from *Burn-In*, *Asymptotic Bias*, and *Asymptotic Variance*; and rigorous discussion of the trade-off between the above error components.

Lastly, Rudolf and Schweizer [RS15] also use the Wasserstein topology. They focus on approximate MCMC algorithms, with applications to autoregressive processes and stochastic Langevin algorithms for Gibbs random fields. Their results use the following assumptions: the original kernel is Wasserstein ergodic; a Lyapunov drift condition for perturbed kernel is given,

with drift function $\tilde{V}$; $\tilde{V}$ has finite expectation under the initial distribution; and the perturbation operator is uniformly bounded in a $\tilde{V}$-normalised Wasserstein norm. Their main results are: explicit bounds on the Wasserstein distance between the original and perturbed finite time sampling distributions; and explicit bounds on the Wasserstein distance between stationary distributions.

Each of the above papers demonstrate bounds on various measures of error from using approximate finite-time sampling distributions and approximate ergodic distributions to calculate expectations of functions. On the other hand, the assumptions underlying the results vary dramatically. The results for uniformly ergodic chains are based on simpler and more intuitive assumptions than those for geometrically ergodic chains. Our work extends these results to geometrically ergodic chains and perturbations while preserving essentially the same level of simplicity in the assumptions.

3 Perturbation Bounds

This section extends the main results of [JMMD15] to the $L_2(\pi)$-geometrically ergodic case for reversible processes, assuming the perturbation $P - P_\epsilon$ has bounded $L_2(\pi)$ operator norm. We follow the derivation in [JMMD15] with minimal structural modification, though the technicalities must be handled differently and additional theoretical machinery is required. We use the fact that the existence of a spectral gap for the restriction of P to $L_2(\pi)$ yields an inequality of the same form as the uniform contractivity condition, but in the $L_2(\pi)$-norm as opposed to the total variation norm (cf. Theorem 2 of [RR97]).

3.1 Assumptions and Notation

We assume throughout that P is the transition kernel for a Markov chain on a countably generated state space $\mathcal{X}$ which is reversible with respect to a stationary probability distribution π. We further assume that P is $L_2(\pi)$-geometrically ergodic, with geometric convergence rate $0 < \rho = (1 - \alpha) < 1$. We let $\|\cdot\|_2$ denote the usual norm in $L_2(\pi)$, as well the corresponding operator norm on $\mathcal{B}(L_2(\pi))$.

We then assume that P_ϵ is a second ("perturbed") transition kernel, with $\|P - P_\epsilon\|_2 \leq \epsilon$ for some fixed $\epsilon > 0$. We assume that P_ϵ has its own stationary distribution, denoted π_ϵ. We assume throughout the technical condition that $\pi_\epsilon \ll \pi$, i.e. that π_ϵ is absolutely continuous with respect to the original stationary distribution π. Many of our results below (where indicated) also assume that $\pi \ll \pi_\epsilon$ (in addition to $\pi_\epsilon \ll \pi$), so that $\pi \equiv \pi_\epsilon$.

We shall write $\|\cdot\|_\epsilon$ for the norm on $L_2(\pi_\epsilon)$, as well as the corresponding operator norm. We also write $\|\cdot\|_1$ for the $L_1(\pi)$ norm, and $\|\cdot\|_{\mathrm{TV}}$ for the total variation norm. By convention we will use the version of the total variation norm which is equal to the $L_1(\pi)$-norm when restricted to $L_1(\pi)$, as opposed to the version which equals one-half of this. On the other hand, $\|\cdot\|_{\mathrm{TV}}$ applies to all bounded measures, while $\|\cdot\|_1$ applies only to the subspace of $L_1(\pi)$ measures. We note also that if $\pi \equiv \pi_\epsilon$, then the $L_1(\pi)$ and $L_1(\pi_\epsilon)$ norms and spaces are equal, so we make no distinction between them.

3.2 Preliminary Results

The following lemma is contained in the remark after Theorem 2 of [RR97]; we prove it here for completeness.

Lemma 1. *For any probability measure $\mu \in L_2(\pi)$,*

$$\|\mu - \pi\|_2^2 = \|\mu\|_2^2 - 1$$

Proof.

$$\begin{aligned} 0 \leq \|\mu - \pi\|_2^2 &= \int \left(\left(\frac{d\mu}{d\pi} \right) - 1 \right)^2 d\pi = \int \left(\left(\frac{d\mu}{d\pi} \right)^2 - 2\frac{d\mu}{d\pi} + 1 \right) d\pi \\ &= \int \left(\frac{d\mu}{d\pi} \right)^2 d\pi - 2\int d\mu + \int d\pi = \|\mu\|_2^2 - 1 \end{aligned}$$

□

We also have:

Proposition 2. *Under the assumptions of Section 3.1,*

$$\|\nu_1 P^n - \nu_2 P^n\|_2 \leq (1-\alpha)^n \|\nu_1 - \nu_2\|_2$$

for any probability distributions $\nu_1, \nu_2 \in L_2(\pi)$. In particular, taking $\nu_1 = \pi$,

$$\|\pi - \nu_2 P^n\|_2 \leq (1-\alpha)^n \|\pi - \nu_2\|_2 = (1-\alpha)^n \sqrt{\|\nu_2\|_2^2 - 1} < (1-\alpha)^n \|\nu_2\|_2$$

and applying Cauchy-Schwarz yields

$$\|\pi - \nu_2 P^n\|_1 \leq \|\pi - \nu_2 P^n\|_2 \leq (1-\alpha)^n \|\pi - \nu_2\|_2 = C_{\nu_2}(1-\alpha)^n$$

Proof. This statement follows from Theorem 2 of [RR97] and Lemma 1. □

3.3 Bounds in the Original Norm

We begin with a first result giving conditions under which the stationary distribution π_ϵ of the perturbed chain is in $L_2(\pi)$:

Proposition 3. *Under the assumptions of Section 3.1, if in addition $\epsilon < \alpha$ and $\pi \ll \pi_\epsilon$, then $\pi_\epsilon \in L_2(\pi)$ with $\|\pi_\epsilon\|_2 \leq \frac{\alpha}{\alpha - \epsilon}$*

Proof. Since $P_\epsilon^n(x, \cdot)$ converges in total variation to π_ϵ for π_ϵ-almost every x and since $\pi \ll \pi_\epsilon$, then $P_\epsilon^n(x, \cdot)$ converges in total variation to π_ϵ for π-almost every x. Thus, πP_ϵ^n converges in total variation to π_ϵ. Also since π and π_ϵ are equivalent measures, $\pi_\epsilon \in L_1(\pi)$. Let $Q = (P_\epsilon - P)$. We will use the fact that leading P's preserve π while Q maps π to a signed measure in $L_2(\pi)$ which integrates to 0. Compositions of P and Q applied to $L_2(\pi)$ signed measures which integrate to 0 yield $L_2(\pi)$ signed measures which integrate to 0. When restricted to $L_2(\pi)$ signed measures which integrate to 0, P has norm $(1-\alpha)$. Q has norm at most ϵ on all of $L_2(\pi)$. Let $\mathbf{2}^k = \{0,1\}^k$ for all $k \in \mathbb{N}$. We also note that $\|\pi\|_2 = 1$.

We complete the proof in two stages. First we show that $\{\pi P_\epsilon^n\}_{n \in \mathbb{N}}$ is an $L_2(\pi)$-Cauchy sequence, thus from completeness it must have an $L_2(\pi)$-limit, say $\hat{\pi}_\epsilon$. It will then be true that $\hat{\pi}_\epsilon = \pi_\epsilon$ because L_2 is a subspace of $L_1(\pi)$ and convergence in $L_2(\pi)$ implies convergence in $L_1(\pi)$ from Cauchy-Schwarz, which in turn implies convergence in total variation. Secondly we will prove the upper bound on the norm of π_ϵ. In both stages we will expand $(P+Q)^n$, and then group by the number of leading P's.

Let $m, n \in \mathbf{N}$ be arbitrary with $m \leq n$.

$$\begin{aligned}
\|\pi P_\epsilon^n - \pi P_\epsilon^m\|_2 &= \|\pi(P+Q)^n - \pi(P+Q)^m\|_2 \\
&= \left\| \pi \left[\left(\sum_{\mathbf{b}\in\mathbf{2}^n} \prod_{j=1}^{n} P^{b_j} Q^{1-b_j} \right) - \left(\sum_{\mathbf{b}\in\mathbf{2}^m} \prod_{j=1}^{m} P^{b_j} Q^{1-b_j} \right) \right] \right\|_2 \\
&= \left\| \pi \left[\left(P^n + \sum_{k=0}^{n-1} P^{n-k-1} Q \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} P^{b_j} Q^{1-b_j} \right) \right. \right. \\
&\qquad \left. \left. - \left(P^m + \sum_{k=0}^{m-1} P^{m-k-1} Q \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} P^{b_j} Q^{1-b_j} \right) \right] \right\|_2 \\
&= \left\| \left(\pi + \sum_{k=0}^{n-1} \pi Q \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} P^{b_j} Q^{1-b_j} \right) - \left(\pi + \sum_{k=0}^{m-1} \pi Q \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} P^{b_j} Q^{1-b_j} \right) \right\|_2 \\
&= \left\| \pi \sum_{k=m}^{n-1} Q \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} P^{b_j} Q^{1-b_j} \right\|_2 \\
&\leq \left(\epsilon \sum_{k=m}^{n-1} \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} (1-\alpha)^{b_j} \epsilon^{1-b_j} \right) \\
&= \left(\epsilon \sum_{k=m}^{n-1} (1-\alpha+\epsilon)^k \right) \\
&\leq \epsilon \frac{(1-\alpha+\epsilon)^m - (1-\alpha+\epsilon)^n}{\alpha - \epsilon}
\end{aligned}$$

Since this upper bound on $\|\pi P_\epsilon^n - \pi P_\epsilon^m\|_2$ decreases to 0 monotonically in $m = \min(m, n)$ then the sequence must be $L_2(\pi)$-Cauchy. As argued above, let $\hat{\pi}_\epsilon$ be the $L_2(\pi)$ limit of this sequence (which exists and belongs to $L_2(\pi)$ from completeness). Then, applying Cauchy-Schwarz, $\|\pi P_\epsilon^n - \hat{\pi}_\epsilon\|_1 \leq \|\pi P_\epsilon^n - \hat{\pi}_\epsilon\|_2$, so that since the right hand side converges to 0 then the left side must as well. Since the sequence can have only one limit in $L_1(\pi)$ we must have that $\pi_\epsilon = \hat{\pi}_\epsilon$ and hence $\pi_\epsilon \in L_2(\pi)$.

Now, we proceed to establish an upper bound on $\|\pi_\epsilon\|_2$. Let $n \in \mathbf{N}$ be

arbitrary. Then

$$\begin{aligned}
\|\pi P_\epsilon^n\|_2 &= \|\pi(P+Q)^n\|_2 \\
&= \left\| \pi \left(\sum_{\mathbf{b}\in\mathbf{2}^n} \prod_{j=1}^{n} P^{b_j} Q^{1-b_j} \right) \right\|_2 \\
&= \left\| \pi \left(P^n + \sum_{k=0}^{n-1} P^{n-k-1} Q \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} P^{b_j} Q^{1-b_j} \right) \right\|_2 \\
&\leq \left(1 + \epsilon \sum_{k=0}^{n-1} \sum_{\mathbf{b}\in\mathbf{2}^k} \prod_{j=1}^{k} (1-\alpha)^{b_j} \epsilon^{1-b_j} \right) \\
&= \left(1 + \epsilon \sum_{k=0}^{n-1} (1-\alpha+\epsilon)^k \right) \\
&\leq \frac{\alpha}{\alpha-\epsilon}\,.
\end{aligned}$$

From the continuity of norm, $\|\pi_\epsilon\|_2 \leq \sup_{n\in\mathbb{N}} \|\pi P_\epsilon^n\|_2 \leq \frac{\alpha}{\alpha-\epsilon}$ □

Remark 4. The total variation norm of π_ϵ is $\|\pi_\epsilon\|_1 = \|\pi_\epsilon\|_{TV} = 1$. By Lemma 1 and Cauchy-Schwarz, under the conditions of Proposition 3,

$$\|\pi_\epsilon - \pi\|_{TV} \leq \|\pi_\epsilon - \pi\|_2 \leq \frac{\epsilon}{\alpha-\epsilon}\sqrt{\frac{2\alpha}{\epsilon} - 1}$$

The $\sqrt{\frac{2\alpha}{\epsilon} - 1}$ term above grows without bound as $\epsilon \to 0$ for a fixed value of α. Hence this bound is asymptotically worse than the bound on the same quantity in [JMMD15], which equals $\frac{\epsilon}{\alpha}$. Let $b_0(\epsilon) = \frac{\epsilon}{\alpha}$ and let $b_1(\epsilon) = \frac{\epsilon}{\alpha-\epsilon}\sqrt{\frac{2\alpha}{\epsilon} - 1}$. We then have that $\frac{b_1(\epsilon)}{b_0(\epsilon)} = \frac{\alpha}{\alpha-\epsilon}\sqrt{\frac{2\alpha-\epsilon}{\epsilon}} = O(\epsilon^{-1/2})$ as $\epsilon \searrow 0$. The following result allows us to improve our bound.

Proposition 5. *Under the assumptions of Section 3.1, if in addition $\epsilon < \alpha$ and $\pi_\epsilon \in L_2(\pi)$ then*

$$1 \leq \|\pi_\epsilon\|_2 \leq \frac{\alpha}{\sqrt{\alpha^2 - \epsilon^2}}$$

and

$$0 \leq \|\pi - \pi_\epsilon\|_2 \leq \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}}$$

Proof. The two lower bounds are immediate from Lemma 1 and the positivity of norms:

$$0 \leq \|\pi - \pi_\epsilon\|_2^2 = \|\pi_\epsilon\|_2^2 - 1$$

To derive the first upper bound, we apply Lemma 1, our assumptions about the operators P and P_ϵ, and triangle inequality, to $\|\pi - \pi_\epsilon\|_2$:

$$\begin{aligned}\sqrt{\|\pi_\epsilon\|_2^2 - 1} = \|\pi - \pi_\epsilon\|_2 &= \|\pi P - \pi_\epsilon P + \pi_\epsilon P - \pi_\epsilon P_\epsilon\|_2 \\ &\leq \|\pi P - \pi_\epsilon P\|_2 + \|\pi_\epsilon P - \pi_\epsilon P_\epsilon\|_2 \\ &\leq (1-\alpha)\|\pi - \pi_\epsilon\|_2 + \epsilon\|\pi_\epsilon\|_2 \\ &= (1-\alpha)\sqrt{\|\pi_\epsilon\|_2^2 - 1} + \epsilon\|\pi_\epsilon\|_2\end{aligned}$$

Collecting the square roots and squaring both sides yields

$$\alpha^2 \left(\|\pi_\epsilon\|_2^2 - 1\right) \leq \epsilon^2 \|\pi_\epsilon\|_2^2$$

which implies that

$$\|\pi_\epsilon\|_2^2 \leq \frac{\alpha^2}{\alpha^2 - \epsilon^2}$$

Finally, the second upper bound is derived from the first one, again using Lemma 1:

$$\|\pi - \pi_\epsilon\|_2^2 = \|\pi_\epsilon\|_2^2 - 1 \leq \frac{\alpha^2}{\alpha^2 - \epsilon^2} - 1 = \frac{\epsilon^2}{\alpha^2 - \epsilon^2}$$

□

Remark 6. This upper bound for $\|\pi_\epsilon\|_2$ is tighter than the bound from Proposition 3 by a factor of $\sqrt{\frac{\alpha-\epsilon}{\alpha+\epsilon}} < 1$. By applying Cauchy-Schwarz again we have $\|\pi - \pi_\epsilon\|_1 \leq \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}$. This result thus extends the bound from the uniformly ergodic case with asymptotically no loss. That is, in the uniformly ergodic case, the bound in [JMMD15] is $\|\pi - \pi_\epsilon\|_1 \leq \frac{\epsilon}{\alpha}$. This compares with our result, $\|\pi - \pi_\epsilon\|_1 \leq \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}$. Indeed, let $b_0(\epsilon) = \frac{\epsilon}{\alpha}$ as in Remark 4, and let $b_2(\epsilon) = \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}$. Then the bounds b_0 and b_2 are asymptotically equivalent, since $\lim_{\epsilon \searrow 0} \frac{b_2(\epsilon)}{b_0(\epsilon)} = 1$.

We next observe that our assumptions imply that for small enough perturbations, the perturbed chain P_ϵ is geometrically ergodic in the $L_2(\pi)$ norm. (It is, however, awkward to use the $L_2(\pi)$ norm when studying P_ϵ; this is corrected in Section 3.4 below.)

Proposition 7. *Under the assumptions of Section 3.1, if $\epsilon < \alpha$ and $\pi_\epsilon \in L_2(\pi)$, then P_ϵ is $L_2(\pi)$-geometrically ergodic, with geometric contraction factor $\leq 1 - (\alpha - \epsilon)$.*

Proof. Suppose that $\nu \in L_2(\pi)$ with $\nu(\mathcal{X}) = 0$. Then

$$\|\nu P_\epsilon\|_2 \leq \|\nu(P_\epsilon - P)\|_2 + \|\nu P\|_2 \leq \epsilon\|\nu\|_2 + (1-\alpha)\|\nu\|_2 = (1 - (\alpha - \epsilon))\|\nu\|_2 \,.$$

Thus, for any probability measure $\mu \in L_2(\pi)$, since $\pi_\epsilon \in L_2(\pi)$, we have

$$\|\mu P_\epsilon^n - \pi_\epsilon\|_2 = \|(\mu - \pi_\epsilon)P_\epsilon^n\|_2 \leq (1 - (\alpha - \epsilon))^n \|\mu - \pi_\epsilon\|_2 \,.$$

□

Combining Propositions 3 and 5 and 7 together immediately yields:

Theorem 8. *Under the assumptions of Section 3.1, if in addition $\epsilon < \alpha$ and $\pi \ll \pi_\epsilon$, then $\pi_\epsilon \in L_2(\pi)$, and*

$$1 \leq \|\pi_\epsilon\|_2 \leq \frac{\alpha}{\sqrt{\alpha^2 - \epsilon^2}}$$

and

$$0 \leq \|\pi - \pi_\epsilon\|_2 \leq \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}} \,,$$

and also P_ϵ is $L_2(\pi)$-geometrically ergodic with geometric contraction factor $\leq 1 - (\alpha - \epsilon)$.

Remark 9. [JMMD15] require the stronger condition that $2\epsilon < \alpha$ for their corresponding result. However, this appears to be due to our defining ϵ as a bound on $L_2(\pi)$ differences, as opposed to TV differences.

3.4 Switching to the Perturbed Norm

The results of the previous section bound the perturbed chain P_ϵ in terms of the original norm $L_2(\pi)$. It would be more satisfying to demonstrate that P_ϵ is geometrically ergodic in the $L_2(\pi_\epsilon)$ norm, and this would also allow us to use the equivalences in [RR97]. We provide sufficient conditions for this extension below. First, we introduce the notion of a *hyper-small* set.

Definition 10. Following [RR97], a subset $S \subset \mathcal{X}$ is called *hyper-small* for the Markov kernel P if $\pi(S) > 0$ and there exists $\delta_S > 0$ and $k \in \mathbb{N}$ such that $\frac{dP^k(x,\cdot)}{d\pi} \geq \delta_S \mathbf{1}_S(x)$ or equivalently $P^k(x, A) \geq \delta_S \pi(A)$ for all $x \in S$ and $A \subset X$ measurable.

A main result of [JJ67] (see the discussion in [RR97]) is that on a countably generated state space (as we have assumed herein), every set of positive π measure contains a hyper-small subset. We will use this fact repeatedly in the proof of the following theorem. Also of importance to us is the $\langle i'' \Rightarrow i \rangle$ part of Proposition 1 of [RR97], which provides a characterisation of geometric ergodicity in terms of convergence to a hyper-small set.

Theorem 11. *Under the assumptions of Section 3.1, if $\epsilon < \alpha$ and $\pi \ll \pi_\epsilon$, then P_ϵ is total variation geometrically ergodic and L_1-geometrically ergodic with geometric contraction factor $\leq (1 - (\alpha - \epsilon))$. If P_ϵ is reversible, then P_ϵ is also $L_2(\pi_\epsilon)$-geometrically ergodic, with geometric contraction factor $\leq (1 - (\alpha - \epsilon))$.*

Proof. From [JJ67], there is a set R_0 which is hyper-small for P_ϵ. Since $\pi_\epsilon \ll \pi$ then $\pi_\epsilon(R_0) > 0 \Rightarrow \pi(R_0) > 0$, thus from [JJ67] again R_0 contains a hyper-small set, R_1, for P. Since $\pi(R_1) > 0$, and since by assumption $\pi \ll \pi_\epsilon$, then $\pi_\epsilon(R_1) > 0$ as well. Since any subset of a hyper-small set with positive measure is also hyper-small, R_1 is hyper-small for both P and P_ϵ. One may suppose that the smallness for the two chains comes with the same lower bound constant by taking the smaller of the respective constants. R_1 must contain a subset, S with $\pi_\epsilon(S) > 0$ where either $\pi(A) \leq \pi_\epsilon(A)$ for all measurable $A \subset S$ or $\pi(A) \geq \pi_\epsilon(A)$ for all measurable $A \subset S$, since one can partition R_1 based on whether the Radon-Nikodym derivative of π_ϵ with respect to π exceeds 1. We may then select S as a small set for both chains and the smaller of the two measures on S as a minorizing measure for

both chains. Now one may apply the Nummelin splitting technique yielding versions of the two chains which share a hyper-small set, S_1, with the same ergodic probability when restricted to the hyper-small set.

We will be using the notation from [MT09], where the components of the split chain are represented with a haček ($\check{\cdot}$) and subsets of the split state space are represented with a subscript of 0 indicating the original chain less the minorizor and 1 indicating the minorizor. The probability measure defined by restricting $\check{\pi}$ to S_1 and renormalizing is clearly in $L_2(\check{\pi})$ and has norm $\frac{1}{\sqrt{\check{\pi}(S_1)}}$. This measure is explicitly written as $\frac{\mathbf{1}_{S_1}}{\check{\pi}(S_1)}(A) = \frac{\check{\pi}(A \cap S_1)}{\check{\pi}(S_1)}$ and is clearly equal to the similar measure created by restricting $\check{\pi}_\epsilon$ to S_1 since $\check{\pi}$ and $\check{\pi}_\epsilon$ are equal on S_1. The notation defined is illuminative of the restricted measure's Radon-Nikodym derivative with respect to $\check{\pi}$. Also since the pre-split perturbed chain is $L_2(\pi)$-geometrically ergodic, the split chain is $L_2(\check{\pi})$-geometrically ergodic with the same convergence rate. Thus we can write

$$\begin{aligned}
\left\| \int_{S_1} \frac{\check{\pi}_\epsilon(dy)}{\check{\pi}_\epsilon(S_1)} \check{P}_\epsilon^n(y,\cdot) - \check{\pi}_\epsilon \right\|_{\mathrm{TV}} &= \left\| \int_{S_1} \frac{\check{\pi}(dy)}{\check{\pi}(S_1)} \check{P}_\epsilon^n(y,\cdot) - \check{\pi}_\epsilon \right\|_{\mathrm{TV}} \\
&= \left\| \frac{\mathbf{1}_{S_1}}{\check{\pi}(S_1)} \check{P}_\epsilon^n - \check{\pi}_\epsilon \right\|_{\mathrm{TV}} \\
&\leq \left\| \frac{\mathbf{1}_{S_1}}{\check{\pi}(S_1)} \check{P}_\epsilon^n - \check{\pi}_\epsilon \right\|_2 \\
&\leq (1-(\alpha-\epsilon))^n \left\| \frac{\mathbf{1}_{S_1}}{\check{\pi}(S_1)} - \check{\pi}_\epsilon \right\|_2 ,
\end{aligned}$$

where the first inequality comes from Cauchy-Schwarz, and the second comes from the previous proposition. Therefore, from Proposition 1 of [RR97], the post-split perturbed chain is $\check{\pi}_\epsilon$-almost everywhere geometrically ergodic.

As per [RT01], the $L_1(\pi_\epsilon)$ and $L_2(\pi_\epsilon)$-geometric contraction factors, ρ_1 and ρ_2 respectively, may be expressed as:

$$\rho_1 = \exp\left(\sup_{\mu \in b(\pi_\epsilon)} \lim_{n\to\infty} \frac{\log \|\mu P_\epsilon^n - \pi_\epsilon\|_1}{n} \right) \quad \rho_2 = \exp\left(\sup_{\mu \in b(\pi_\epsilon)} \lim_{n\to\infty} \frac{\log \|\mu P_\epsilon^n - \pi_\epsilon\|_2}{n} \right)$$

where $b(\pi_\epsilon) = \{\mu \in L_1(\pi_\epsilon) : \frac{d\mu}{d\pi_\epsilon} \text{ is bounded}\}$. We see that $b(\pi_\epsilon) \subset L_2(\pi)$, since for any $\mu \in b(\pi_\epsilon)$ there is an $M > 0$ such that $\frac{d\mu}{d\pi_\epsilon} \leq M$ so that

$$\|\mu\|_2^2 = \int \left(\frac{d\mu}{d\pi}\right)^2 d\pi = \int \left(\frac{d\mu}{d\pi_\epsilon}\frac{d\pi_\epsilon}{d\pi}\right)^2 d\pi \leq M^2 \int \left(\frac{d\pi_\epsilon}{d\pi}\right)^2 d\pi = M^2 \|\pi_\epsilon\|_2^2 < \infty$$

Hence, applying Proposition 7, we have that

$$\rho_1 \leq \exp\left(\sup_{\mu \in b(\pi_\epsilon)} \lim_{n\to\infty} \frac{\log C_\mu (1-(\alpha-\epsilon))^n}{n}\right) = (1-(\alpha-\epsilon))\,,$$

where $C_\mu = \frac{M\epsilon}{\sqrt{\alpha^2-\epsilon^2}}$. It then follows from the main result of [RT01] that if the chain P_ϵ is reversible, then $L_1(\pi_\epsilon)$-geometric ergodicity implies $L_2(\pi_\epsilon)$-geometric ergodicity with the same geometric contraction factor. □

Corollary 12. *If $\alpha < \epsilon$ and $\pi \ll \pi_\epsilon$ then for any probability measure $\mu \in L_2(\pi)$,*

$$\|\mu P_\epsilon^n - \pi\|_2 \leq (1-(\alpha-\epsilon))^n \|\mu - \pi_\epsilon\|_2 + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\,,$$

and for every $\nu \in L_1(\pi)$ there is $C_\nu < \infty$ with

$$\|\nu P_\epsilon^n - \pi\|_1 \leq (1-(\alpha-\epsilon))^n C_\nu + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\,.$$

If additionally $\|P_\epsilon - P\|_\epsilon < \epsilon$, and P_ϵ is reversible, then

$$\|\mu P_\epsilon^n - \pi\|_\epsilon \leq (1-(\alpha-\epsilon))^n \|\mu - \pi_\epsilon\|_\epsilon + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\,.$$

Proof. The first result follows from Theorem 8 and the triangle inequality, since

$$\|\mu P_\epsilon^n - \pi\|_2 \leq \|\mu P_\epsilon^n - \pi_\epsilon\|_2 + \|\pi - \pi_\epsilon\|_2 \leq (1-(\alpha-\epsilon))^n \|\mu - \pi_\epsilon\|_2 + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\,.$$

The second result follows similarly from the triangle inequality and Theorems 8 and 11. The third result then follows similarly by symmetry under the additional assumptions. □

3.5 Covariance Bounds

We next turn our attention to the covariance structure of the original and perturbed chains.

We define the class of functions $L_2'(\pi)$ as the collection of Radon-Nikodym derivatives of measures in $L_2(\pi)$ with respect to π. Let X_t and X_t^ϵ denote the original and perturbed chains run from some initial measure $\nu \in L_2(\pi)$.

Corollary 13. *Let f and g be in $L_2'(\pi)$. Then under the assumptions of Section 3.1,*

$$\mathbb{C}\mathrm{ov}[f(X_t), g(X_s)] \leq (1-\alpha)^{|t-s|}\|f\|_\star\|g\|_\star$$

and if $\epsilon < \alpha$

$$\mathbb{C}\mathrm{ov}[f(X_t^\epsilon), g(X_s^\epsilon)] \leq (1-(\alpha-\epsilon))^{|t-s|}\|f\|_\star\|g\|_\star\,,$$

where $\|h\|_\star = \|h - \pi(h)\|_2$, and $\pi(h)$ is the constant function equal to $\int h(s)\pi(ds)$ everywhere.

Proof. The proof of this result follows the proof of Corollary B.5 in [JMMD15]. We only show the proof for the original chain, however the proof for the perturbed chain is the same *mutatis mutandis*. Define the subspace $L_{2,0}'(\pi) = \{h \in L_2'(\pi) : \int h(s)\pi(ds) = 0\}$, and define the *forward operator*, $F \in \mathcal{B}(L_{2,0}'(\pi))$, by

$$[Ff](x) = \int P(x, dy)f(y) = \mathbb{E}[f(X_1)|X_0 = x]$$

From Lemma 12.6.4 of [Liu08],

$$\sup_{f,g\in L_2'(\pi)} \mathrm{corr}(f(X_0), g(X_t)) = \sup_{\substack{\|f\|_2=1=\|g\|_2 \\ f,g\in L_{2,0}'(\pi)}} \langle f, F^t g\rangle = \|F^t\|_2$$

Consider the canonical isomorphism between $L_2(\pi)$ and $L_2'(\pi)$. The restriction of this isomorphism (on the right) to elements of $L_{2,0}'(\pi)$ yields $L_{2,0}(\pi)$ on the left – the signed measures with total measure 0. The image of F under the restricted isomorphism is the adjoint operator of P restricted to $L_{2,0}(\pi)$. The adjoint of an operator has the same norm as the original operator, hence

$$\|F^t\|_2 \leq \|F\|_2^t = \|P\big|_{L_{2,0}(\pi)}\|_2^t \leq (1-\alpha)^t$$

Therefore

$$\sup_{f,g\in L_2'(\pi)} \mathbb{C}\mathrm{ov}(f(X_0), g(X_t)) \leq \|f\|_\star\|g\|_\star(1-\alpha)^t$$

Since this holds for any initial measure and since $\mathbb{C}\text{ov}$ is symmetric, the shifted and symmetrized result holds for any $f, g \in L_2'(\pi)$:

$$\mathbb{C}\text{ov}[f(X_t), g(X_s)] \leq (1-\alpha)^{|t-s|}\|f\|_\star\|g\|_\star$$

□

Remark 14. Note in Corollary 13 that $\|h\|_\star \leq \|h\|_2 + |\pi(h)| = \|h\|_2 + |\int h d\pi| \leq \|h\|_2 + \int |h| d\pi \leq \|h\|_2 + \sqrt{\int h^2 d\pi} = 2\|h\|_2$. Also note that $\|h\| \leq \|h - \pi(h)\|_2 + |\pi(h)|$.

3.6 Estimation Error Bounds for the Exact Chain

Finally, we turn our attention to bounds on the error of estimation measures of the form $\frac{1}{t}\sum_{k=0}^{t-1} \mu P^k$, and estimates of the form $\frac{1}{t}\sum_{k=0}^{t-1} f(X_k)$. We begin with:

Theorem 15. *Under the assumptions of Section 3.1, for any probability distribution $\mu \in L_2(\pi)$,*

$$\begin{aligned}\left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1}\mu P^k\right\|_2 &\leq \frac{1}{t}\sum_{k=0}^{t-1}\left\|\pi - \mu P^k\right\|_2 \\ &\leq \frac{1}{t}\sum_{k=0}^{t-1}(1-\alpha)^k\|\pi - \mu\|_2 \\ &= \frac{1-(1-\alpha)^t}{t\alpha}\|\pi - \mu\|_2\end{aligned}$$

and for any $\nu \in L_1(\pi_\epsilon)$ there is a $C_\nu > 0$ such that

$$\left\|\pi(\cdot) - \frac{1}{t}\sum_{k=0}^{t-1}[\nu P^k](\cdot)\right\|_1 \leq C_\nu \frac{1-(1-\alpha)^t}{t\alpha}$$

Proof. The first inequality is just the triangle inequality, the second inequality follows from Proposition 2, and the equality follows from direct algebra. The second statement follows from the equivalence of L_1 and L_2-geometric contraction factors for reversible chains from [RT01]. □

We then have:

Theorem 16. *Under the assumptions of Section 3.1, for any initial probability distribution $\mu \in L_2(\pi)$,*

$$\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k)\right)^2\right]$$
$$\leq \|f\|_2^2 \left(\frac{1-(1-\alpha)^t}{t\alpha}\|\pi - \mu\|_2\right)^2 + \|f\|_\star^2 \left(\frac{2-\alpha}{\alpha t} - \frac{2(1-\alpha)}{\alpha^2 t^2} + \frac{2(1-\alpha)^{t+1}}{\alpha^2 t^2}\right)$$

Proof. The proof proceeds by partitioning the MSE via the bias-variance decomposition then bounding the bias and variance terms respectively. We compute that

$$\begin{aligned}
&\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k)\right)^2\right] \\
&= \mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f) - \frac{1}{t}\sum_{k=0}^{t-1}(f(X_k) - [\mu P^k](f))\right)^2\right] \\
&= \left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f)\right)^2 + \mathbb{E}\left[\left(\frac{1}{t}\sum_{k=0}^{t-1}(f(X_k) - [\mu P^k](f))\right)^2\right] \\
&= \left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f)\right)^2 + \frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathbb{C}\mathrm{ov}(f(X_j), f(X_k))
\end{aligned}$$

The bias term is bounded by applying Corollary 13:

$$\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f)\right)^2 \leq \|f\|_2^2 \left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1}\mu P^k\right\|_2^2 \leq \|f\|_2^2 \left(\frac{1-(1-\alpha)^t}{t\alpha}\|\pi - \mu\|_2\right)^2$$

Next, the variance term is bounded using Theorem 15:

$$\begin{aligned}
\frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathbb{C}\mathrm{ov}(f(X_j),f(X_k)) &= \frac{\|f\|_\star^2}{t^2}\sum_{j=0}^{t-1}\left(\sum_{k=0}^{j}(1-\alpha)^{j-k}+\sum_{k=j+1}^{t-1}(1-\alpha)^{k-j}\right)\\
&= \frac{\|f\|_\star^2}{t^2}\sum_{j=0}^{t-1}\left(\frac{1-(1-\alpha)^{j+1}}{\alpha}+\frac{(1-\alpha)-(1-\alpha)^{t-j}}{\alpha}\right)\\
&= \frac{\|f\|_\star^2}{\alpha t^2}\sum_{j=0}^{t-1}\left(1+(1-\alpha)-(1-\alpha)^{j+1}-(1-\alpha)^{t-j}\right)\\
&= \frac{\|f\|_\star^2}{\alpha t^2}\left((2-\alpha)t-2\frac{(1-\alpha)-(1-\alpha)^{t+1}}{\alpha}\right)\\
&= \|f\|_\star^2\left(\frac{2-\alpha}{\alpha t}-\frac{2(1-\alpha)}{\alpha^2 t^2}+\frac{2(1-\alpha)^{t+1}}{\alpha^2 t^2}\right)
\end{aligned}$$

Putting these together yields the desired result. □

Remark 17. We note that, as per Remark 14, $\|f\|_\star \leq 2\|f\|_2$, and likewise $\|f\|_2 \leq \|f\|_\star + |\pi(f)|$. Also in the case that f is is π-essentially bounded, $\|f\|_2 \leq \|f\|_\infty$ and $\|f\|_\star \leq \|f - \mathrm{midrange}(f)\|_\infty$, and $\|f\|_\infty \leq 2\|f - \mathrm{midrange}(f)\|_\infty$, and $\|f - \mathrm{midrange}(f)\|_\infty \leq \|f\|_\infty$. These alternative norms may be substituted into the result as necessary in order to make the bounds tractable for a given application.

Remark 18. Comparing our above geometrically ergodic results to the L_1 results of [JMMD15] in the uniformly ergodic case, we see that the L_2 and L_1 bounds we establish above differ from the corresponding L_1 bound of [JMMD15] only by a factor, which is constant in time, but varies with the initial distribution (as is to be expected when moving from uniform ergodicity to geometric ergodicity). For the Mean-Squared-Error results, the $\|\cdot\|_\star$-norm in that paper is based on the midrange-centred infinity norm, which as per Remark 17 is an upper bound on what we have taken to be the $\|\cdot\|_\star$-norm. (Also, the bias term in our MSE bound decreases as $O(t^{-2})$, while in [JMMD15] it apparently decreases as just $O(t^{-1})$, but we believe this is simply due to their accidentally dropping a square on their TV norm in their calculation Section B.2.) Other than these differences, the bounds are essentially the same.

3.7 Error Bounds for the Perturbed Chain

We next turn our attention to the perturbed chain P_ϵ. We have:

Theorem 19. *Under the assumptions of Section 3.1, suppose $\epsilon < \alpha$ and $\pi_\epsilon \in L_2(\pi)$. Then for any probability distribution $\mu \in L_2(\pi)$,*

$$\begin{aligned}\left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1}\mu P_\epsilon^k\right\|_2 &\leq \frac{1}{t}\sum_{k=0}^{t-1}\left\|\pi - \mu P_\epsilon^k\right\|_2 \\ &\leq \frac{1}{t}\sum_{k=0}^{t-1}\left[(1-(\alpha-\epsilon))^k\|\pi_\epsilon - \mu\|_2 + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\right] \\ &= \frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)}\|\pi_\epsilon - \mu\|_2 + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\end{aligned}$$

If, in addition, $\pi \ll \pi_\epsilon$ then for any $\nu \in L_1(\pi) \equiv L_1(\pi_\epsilon)$ there is a $C_\nu^{(\epsilon)} > 0$ such that

$$\left\|\pi(\cdot) - \frac{1}{t}\sum_{k=0}^{t-1}[\nu P^k](\cdot)\right\|_1 \leq C_\nu^{(\epsilon)}\frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)} + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}$$

and if in addition $\|P - P_\epsilon\|_\epsilon < \epsilon$, P_ϵ is reversible, then the first result holds in the $\|\cdot\|_\epsilon$-norm.

Proof. Again, the first result is a direct consequence of the triangle inequality applied to previous propositions, the second result follows from the triangle inequality and Theorem 11, and the third result follows from symmetry in conjunction with Theorem 11. □

We then have:

Theorem 20. *Under the assumptions of Section 3.1, suppose $\epsilon < \alpha$. For*

any initial probability distribution $\mu \in L_2(\pi)$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k^\epsilon)\right)^2\right] \\
&\leq \|f\|_2^2 \left(\frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}} + \left(\frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)}\|\pi_\epsilon - \mu\|_2\right)\right)^2 \\
&\qquad + \|f\|_\star^2 \left(\frac{2-(\alpha-\epsilon)}{(\alpha-\epsilon)t} - \frac{2(1-(\alpha-\epsilon))}{(\alpha-\epsilon)^2 t^2} + \frac{2(1-(\alpha-\epsilon))^{t+1}}{(\alpha-\epsilon)^2 t^2}\right).
\end{aligned}$$

If in addition $\|P - P_\epsilon\|_\epsilon < \epsilon$ and $\pi \ll \pi_\epsilon$, then the same inequality holds in the $\|\cdot\|_\epsilon$ norm, and also in the $\|\cdot\|_{\epsilon\star}$ norm.

Proof. We again proceed via bias-variance decomposition, as in the corresponding result for the exact chain. However, now the bias under consideration is itself decomposed as the square of a sum of two components. The squared sum is expanded simultaneously with the bias-variance expansion. (And, Remark 17 regarding alternative norms for the exact chain holds here as well.) We compute that

$$\begin{aligned}
&\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k^\epsilon)\right)^2\right] \\
&= \mathbb{E}\left[\left(\pi(f) - \pi_\epsilon(f) + \frac{1}{t}\sum_{k=0}^{t-1}\left[\pi_\epsilon - \mu P_\epsilon^k\right](f) - \frac{1}{t}\sum_{k=0}^{t-1}(f(X_k^\epsilon) - [\mu P_\epsilon^k](f))\right)^2\right] \\
&= \left([\pi - \pi_\epsilon](f)\right)^2 + 2\left([\pi - \pi_\epsilon](f)\right)\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right) \\
&\qquad + \left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)^2 + \mathbb{E}\left[\left(\frac{1}{t}\sum_{k=0}^{t-1}(f(X_k^\epsilon) - [\mu P_\epsilon^k](f))\right)^2\right] \\
&= \left([\pi - \pi_\epsilon](f)\right)^2 + 2\left([\pi - \pi_\epsilon](f)\right)\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right) \\
&\qquad + \left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)^2 + \frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathbb{C}\mathrm{ov}(f(X_j^\epsilon), f(X_k^\epsilon))
\end{aligned}$$

We bound the first component of the bias term using Proposition 5:

$$([\pi - \pi_\epsilon](f))^2 \leq \|\pi - \pi_\epsilon\|_2^2 \|f\|_2^2 \leq \frac{\epsilon^2 \|f\|_2^2}{\alpha^2 - \epsilon^2}$$

We bound the third component of the bias term using Theorem 19:

$$\begin{aligned}\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)^2 &\leq \|f\|_2^2 \left\|\pi\epsilon - \frac{1}{t}\sum \mu P_\epsilon^k\right\|_2^2 \\ &\leq \|f\|_2^2 \left(\frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)}\|\pi_\epsilon - \mu\|_2\right)^2\end{aligned}$$

We bound the variance term using Corollary 13:

$$\begin{aligned}&\frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathbb{C}\mathrm{ov}(f(X_j^\epsilon), f(X\epsilon_k)) \\ &= \frac{\|f\|_\star^2}{t^2}\sum_{j=0}^{t-1}\left(\sum_{k=0}^{j}(1-\alpha)^{j-k} + \sum_{k=j+1}^{t-1}(1-(\alpha-\epsilon))^{k-j}\right) \\ &= \frac{\|f\|_\star^2}{t^2}\sum_{j=0}^{t-1}\left(\frac{1-(1-(\alpha-\epsilon))^{j+1}}{(\alpha-\epsilon)} + \frac{(1-(\alpha-\epsilon)) - (1-(\alpha-\epsilon))^{t-j}}{(\alpha-\epsilon)}\right) \\ &= \frac{\|f\|_\star^2}{(\alpha-\epsilon)t^2}\sum_{j=0}^{t-1}\left(1 + (1-(\alpha-\epsilon)) - (1-(\alpha-\epsilon))^{j+1} - (1-(\alpha-\epsilon))^{t-j}\right) \\ &= \frac{\|f\|_\star^2}{(\alpha-\epsilon)t^2}\left(t(2-(\alpha-\epsilon)) - 2\frac{(1-(\alpha-\epsilon)) - (1-(\alpha-\epsilon))^{t+1}}{(\alpha-\epsilon)}\right) \\ &= \|f\|_\star^2\left(\frac{2-(\alpha-\epsilon)}{(\alpha-\epsilon)t} - \frac{2(1-(\alpha-\epsilon))}{(\alpha-\epsilon)^2t^2} + \frac{2(1-(\alpha-\epsilon))^{t+1}}{(\alpha-\epsilon)^2t^2}\right)\end{aligned}$$

Finally, we bound the second bias term using Proposition 5 and Theorem 19:

$$2\left([\pi - \pi_\epsilon](f)\right)\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right) \leq \frac{\epsilon\|f\|_2}{\sqrt{\alpha^2-\epsilon^2}}\|f\|_2\left(\frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)}\|\pi_\epsilon - \mu\|_2\right)$$

Putting these together yields the desired result in the $\|\cdot\|$-norm. The $\|\cdot\|_\epsilon$ versions follow by symmetry. □

Remark 21. Comparing the above results to the corresponding uniform L_1 result of [JMMD15], we see that the burn-in bias part of our L_2 and L_1 bounds differ from their L_1 burn-in bias part only by a factor which is constant in time, but vary with the initial distribution (as is, again, to be expected when moving from uniform ergodicity to geometric ergodicity). The asymptotic-bias component of our L_2 and L_1 bounds are equivalent asymptotically as $\epsilon \to 0$, as in Remark 6. For the Mean-Squared-Error results, the $\|\cdot\|_\star$-norm in that paper is based on midrange-centred infinity norm, which again is an upper bound on what we have taken to be the $\|\cdot\|_\star$-norm. (Also, the bias terms again differ as $O(t^{-2})$ versus $O(t^{-1})$, as previously discussed in Remark 18 above.) Again, other than these differences, the bounds are essentially the same.

4 Application to Noisy MCMC

The Noisy Metropolis Hastings algorithm (nMH), as found in [AFEB16], is defined below along with the classical Metropolis Hastings (MH) algorithm. Note that the main difference between these algorithms is the acceptance ratio, α. Given the current state and the proposed next state, the acceptance ratio is deterministic for classical MH while it is stochastic for nMH. While the acceptance ratio has a generic formula for the MH algorithm, there are various expressions found in different types of nMH algorithms – hence our use of ambiguous notation in this case.

Algorithm 1 Metropolis Hastings

$x_0 \leftarrow$ sample ν_0
for $i = 1$ to N **do**
    $y_i \leftarrow$ sample $q(y_i|x_{i-1})$
    $\alpha_i \leftarrow \alpha(y_i|x_{i-1}) = \frac{\pi(y_i)q(x_{i-1}|y_i)}{\pi(x_{i-1})q(y_i|x_{i-1})}$
    $u_i \leftarrow$ sample unif$[0,1]$
    if $u_i \leq \alpha_i$ **then**
        $x_i \leftarrow y_i$
    else
        $x_i \leftarrow x_{i-1}$
    end if
end for

Algorithm 2 Noisy Metropolis Hastings

$x_0 \leftarrow$ sample ν_0
for $i = 1$ to N **do**
    $y_i \leftarrow$ sample $q(y_i|x_{i-1})$
    $z_i \leftarrow$ sample $f(z_i|y_i)$
    $\hat{\alpha}_i \leftarrow \hat{\alpha}(y_i|x_{i-1}, z_i)$
    $u_i \leftarrow$ sample unif$[0, 1]$
    if $u_i \leq \hat{\alpha}_i$ **then**
        $x_i \leftarrow y_i$
    else
        $x_i \leftarrow x_{i-1}$
    end if
end for

For our analysis of these algorithms, P will represent the transition kernel for the classical MH algorithm while $\hat{P}$ will represent the kernel for the corresponding nMH chain. The key will be to show the $L_2(\pi)$ closeness of the nMH transition kernel to the MH transition kernel. Again, $\|\cdot\|_2$ is the norm on $L_2(\pi)$ and the corresponding operator norm. We will assume that π and $\{q(\cdot|x)\}_{x\in\mathcal{X}}$ are all absolutely continuous with respect to the Lebesgue measure and have densities symbolised appropriately. All arguments below would follow identically if there were an arbitrary dominating measure in place of the Lebesgue measure. Let Q be the operator notation for the proposal kernel. We define the following functions for notational convenience. The conventions are that underbars represent the minimum of the quantity with 1, and primes denote signed quantities, and capitals denote linear operators on the space of signed measures.

$$\underline{\alpha}(y|x) = 1 \wedge \alpha(y|x) \qquad\qquad \underline{\hat{\alpha}}(y|x,z) = 1 \wedge \hat{\alpha}(y|x,z)$$

$$\delta(y|x) = \mathbb{E}_{z\sim f_y}|\alpha(y|x) - \hat{\alpha}(y|x,z)| = \int |\alpha(y|x) - \hat{\alpha}(y|x,z)| f_y(z)dz$$

$$\delta'(y|x) = \mathbb{E}_{z\sim f_y}\left(\underline{\alpha}(y|x) - \underline{\hat{\alpha}}(y|x,z)\right) = \int \left(\underline{\alpha}(y|x) - \underline{\hat{\alpha}}(y|x,z)\right) f_y(z)dz$$

$$\gamma(x) = \mathbb{E}_{y\sim q(y|x)}\delta(y|x) = \int \delta(y|x)q(y|x)dy \qquad [\nu\Gamma'](dy) = \nu(y)\gamma'(y)dy$$
$$\gamma'(x) = \mathbb{E}_{y\sim q(y|x)}\delta'(y|x) = \int \delta'(y|x)q(y|x)dy \quad [\nu Z'](dy) = \left[\int \delta'(y|x)q(y|x)\nu(x)dx\right] dy$$

Then $|\delta'(y|x)| \leq \delta(y|x)$ for all (x, y) from monotonicity and since $(1 \wedge \cdot)$ is Lipshitz with constant 1.

Lemma 22. $\left(P - \hat{P}\right) = (Z' - \Gamma')$

Proof. We first give expressions for the elements of measure for transitions of the original chain. The first formula is the element of measure for transition from an arbitrary, fixed initial point. It is defined for us by the mechanics of the Metropolis Hastings algorithm. The second expression is the element of measure for transition from a sample from an initial distribution, ν. It is derived from the first expression by integrating over the sample from ν.

$$P(x, dx') = \delta_x(dx')\left[1 - \int (\underline{\alpha}(y|x)q(y|x)dy\right] + \underline{\alpha}(x'|x)q(x'|x)dx'$$
$$\begin{aligned}[\nu P](dx') &= \int \left(\delta_x(dx')\left[1 - \int \underline{\alpha}(y|x)q(y|x)dy\right] + \underline{\alpha}(x'|x)q(x'|x)dx'\right)\nu(x)dx \\ &= \left[1 - \int \underline{\alpha}(y|x')q(y|x')dy\right]\nu(x')dx' + \left[\int \underline{\alpha}(x'|x)q(x'|x)\nu(x)dx\right]dx'\end{aligned}$$

The second form of the second expression is an application of Fubini's theorem. The exchange of the order of integration for the second term in the expression is immediately obvious and is 'safe' since the integrand for this term is non-negative. The exchange of the order of integration for the first term of the expression is less obvious, however it follows from the realization that for arbitrary non-negative functions f and g,

$$\int_s \int_t f(s)g(t)\delta_t(ds)dt = \int_t \int_s f(s)g(t)\delta_t(ds)dt = \int_t f(t)g(t)dt = \int_s f(s)g(s)ds$$

Where the first equality is Fubini's theorem, the second comes from integrating with respect to s, and the third comes from a change of dummy variable.

Similarly, the elements of measure for transitions from the approximating kernel are expressed below. The first expression, as above, is the element of measure for transition from an arbitrary, fixed initial point. It is defined for us by the mechanics of the Noisy Metropolis Hastings algorithm. The second expression is again derived by integrating the first against an initial sampling measure, ν.

$$\hat{P}(x, dx') = \delta_x(dx') \left[1 - \iint \hat{\underline{\alpha}}(y|x, z) q(y|x) f_y(z) dz dy\right] + \int \hat{\underline{\alpha}}(x'|x, z) q(x'|x) f_{x'}(z) dz dx'$$

$$\begin{aligned}\left[\nu\hat{P}\right](dx') &= \int \left(\delta_x(dx') \left[1 - \iint \hat{\underline{\alpha}}(y|x, z) q(y|x) f_y(z) dz dy\right]\right. \\ &\qquad \left. + \int \hat{\underline{\alpha}}(x'|x, z) q(x'|x) f_{x'}(z) dz dx'\right) \nu(x) dx \\ &= \left[1 - \iint \hat{\underline{\alpha}}(y|x', z) q(y|x') f_y(z) dz dy\right] \nu(x') dx' \\ &\qquad + \left[\iint \hat{\underline{\alpha}}(x'|x, z) q(x'|x) f_{x'}(z) \nu(x) dz dx\right] dx'\end{aligned}$$

The same applications of Fubini's theorem occur as above, however for triple integrals.

We may now leverage our notation defined above to simplify the difference of these elements of measure.

$$\begin{aligned}\left[\nu(P - \hat{P})\right](dx') &= \left[\iint \left(\hat{\underline{\alpha}}(y|x', z) - \underline{\alpha}(y|x')\right) q(y|x') f_y(z) dz dy\right] \nu(x') dx' \\ &\qquad + \left[\iint \left(\underline{\alpha}(x'|x) - \hat{\underline{\alpha}}(x'|x, z)\right) q(x'|x) f_{x'}(z) \nu(x) dz dx\right] dx' \\ &= \left[\int \delta'(x'|x) q(x'|x) \nu(x) dx\right] dx' - \left[\int \delta'(y|x') q(y|x') dy\right] \nu(x') dx' \\ &= [\nu(Z' - \Gamma')](dx')\end{aligned}$$

From this one may conclude that $\left(P - \hat{P} = Z' - \Gamma'\right)$ as operators. □

Proposition 23. *If* $\|Q\|_2 < \infty$ *and* $\sup_{x,y} \delta(y|x) \leq \delta$ *then* $\|\hat{P} - P\|_2 \leq \delta(1 + \|Q\|_2)$.

Proof. It is obvious that if $\delta(y|x) \leq \delta$ uniformly in (x, y) then $(\|Z'\|_2 \leq \delta\|Q\|_2)$, and $(\|\Gamma'\|_2 \leq \delta)$. By applying the previous lemma, given the assumptions stated, $\|P - \hat{P}\|_2 \leq \delta(1 + \|Q\|_2)$. □

Theorem 24. *If $\|Q\|_2 < \infty$, and $\sup_{x,y} \delta(y|x) \leq \delta$, and P is geometrically ergodic with geometric contraction factor $(1-\alpha)$, and $\delta(1+\|Q\|_2) < \alpha$, then*

$$\left\| \pi - \frac{1}{t} \sum_{k=0}^{t-1} \mu \hat{P}^k \right\|_2 \leq \frac{1 - (1 - (\alpha - \epsilon))^t}{t(\alpha - \epsilon)} \|\hat{\pi} - \mu\|_2 + \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}},$$

where $\epsilon = \delta(1 + \|Q\|_2)$, and $\hat{\pi}$ is the stationary distribution for $\hat{P}$.

The above theorem provides an explicit alternative to the analogous result of Corollary 2.3 from [AFEB16], relaxing the uniform ergodicity assumption by putting constraints on Q and δ. In particular, it requires that $Q \in \mathcal{B}(L_2(\pi))$ and that $\delta(1 + \|Q\|_2) < \alpha$. The first of these requirements is not dramatically limiting since the user has control over the choice of Q, but the user should be aware that proposal distributions with sufficiently heavy tails relative to π are likely to present issues in this regard. The second of these requirements is also not dramatically limiting as control over δ may be interpreted as limiting the amount of noise in the nMH algorithm and such control is required regardless in order to ensure the accuracy of approximation in both the geometrically ergodic and uniformly ergodic cases.

References

[AFEB16] Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47, 2016.

[BGJM11] S. Brooks, A. Gelmand, G.L. Jones, and X.-L. Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall, 2011.

[BRR01] L. Breyer, G.O. Roberts, and J.S. Rosenthal. A note on geometric ergodicity and floating-point roundoff error. *Statistics and Probability Letters*, 53:123–127, 2001.

[Gib04] A Gibbs. Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stochastic Models*, 20(4):473–492, 2004.

[JJ67] Naresh Jain and Benton Jamison. Contributions to Doeblin's theory of Markov processes. *Probability Theory and Related Fields*, 8(1):19–40, 1967.

[JMMD15] James E Johndrow, Jonathan C Mattingly, Sayan Mukherjee, and David Dunson. Approximations of Markov chains and high-dimensional Bayesian inference. *arXiv preprint arXiv:1508.03387*, 2015.

[Liu08] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[Mit05] A Yu Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, pages 1003–1014, 2005.

[MT09] Sean P Meyn and Richard L Tweedie. Markov chains and stochastic stability, 2009.

[PS14] Natesh S Pillai and Aaron Smith. Ergodicity of approximate MCMC chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*, 2014.

[RR97] Gareth O Roberts and Jeffrey S Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab*, 2(2):13–25, 1997.

[RRS98] Gareth O Roberts, Jeffrey S Rosenthal, and Peter O Schwartz. Convergence properties of perturbed Markov chains. *Journal of applied probability*, pages 1–11, 1998.

[RS15] Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *arXiv preprint arXiv:1503.04123*, 2015.

[RT01] Gareth O Roberts and Richard L Tweedie. Geometric L2 and L1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability*, pages 37–41, 2001.