Bernoulli **24**(4A), 2018, 2610–2639
https://doi.org/10.3150/17-BEJ938

Perturbation theory for Markov chains via Wasserstein distance

DANIEL RUDOLF[1] and NIKOLAUS SCHWEIZER[2]

[1] *Institut für Mathematische Stochastik, Universität Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany. E-mail: daniel.rudolf@uni-goettingen.de*
[2] *Department of Econometrics & OR, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: n.f.f.schweizer@uvt.nl*

Perturbation theory for Markov chains addresses the question of how small differences in the transition probabilities of Markov chains are reflected in differences between their distributions. We prove powerful and flexible bounds on the distance of the nth step distributions of two Markov chains when one of them satisfies a Wasserstein ergodicity condition. Our work is motivated by the recent interest in approximate Markov chain Monte Carlo (MCMC) methods in the analysis of big data sets. By using an approach based on Lyapunov functions, we provide estimates for geometrically ergodic Markov chains under weak assumptions. In an autoregressive model, our bounds cannot be improved in general. We illustrate our theory by showing quantitative estimates for approximate versions of two prominent MCMC algorithms, the Metropolis–Hastings and stochastic Langevin algorithms.

Keywords: big data; Markov chains; MCMC; perturbations; Wasserstein distance

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are one of the key tools in computational statistics. They are used for the approximation of expectations with respect to probability measures given by unnormalized densities. For almost all classical MCMC methods, it is essential to evaluate the target density. In many cases, this requirement is not an issue, but there are also important applications where it is a problem. This includes applications where the density is not available in closed form, see [27], or where an exact evaluation is computationally too demanding, see [2]. Problems of this kind lead to the approximation of Markov chains and to the question of how small differences in the transitions of two Markov chains affect the differences between their distributions.

In Bayesian inference when *big data* sets are involved an exact evaluation of the target density is typically very expensive. For instance, in each step of a Metropolis–Hastings algorithm the likelihood of a proposed state must be computed. Every observation in the underlying data set contributes to the likelihood and must be taken into account in the calculation. This may result in evaluating several terabytes of data in each step of the algorithm. These are the reasons for the recent interest in numerically cheaper approximations of classical MCMC methods, see [3,4,23,42,47]. A reduction of the computational costs can, for example, be achieved by relying on a moderately sized random subsample of the data in each step of the algorithm. The function value of the target density is thus replaced by an approximation. Naturally, subsampling and alternative

1350-7265 © 2018 ISI/BS

attempts at “cutting the Metropolis–Hastings budget” [23] induce additional biases. These biases can lead to dramatic changes in the properties of the algorithms as discussed in [6].

We thus need a better theoretical understanding of the behavior of such *approximate MCMC* methods. Indeed, a number of recent papers prove estimates of these biases, see [2,3,19,24,29, 35]. A key tool in these papers are perturbation bounds for Markov chains. One such result for uniformly ergodic Markov chains due to Mitrophanov [33] is used in [2]. A similar perturbation estimate implicitly appears in [3]. The focus on uniformly ergodic Markov chains is rather restrictive, especially for high-dimensional, non-compact state spaces such as $\mathbb{R}^m$. Working with Wasserstein distances has recently turned out to be a fruitful alternative in several contributions on high-dimensional MCMC algorithms, see [11,12,14,18,25].

We provide perturbation bounds based on Wasserstein distances, which lead to flexible quantitative estimates of the biases of approximate MCMC methods. Our first main result is the Wasserstein perturbation bound of Theorem 3.1. Under a Wasserstein ergodicity assumption, explained in Section 2, it provides an upper bound on the distance of the nth step distribution between an ideal and an approximating Markov chain in terms of the difference between their one-step transition probabilities. The result is well-suited for applications on a non-compact state space, since the difference of the one-step transition probabilities is measured by a weighted supremum with respect to a suitable Lyapunov function. For an autoregressive model, we show in Section 4.1 that the resulting perturbation bound cannot be improved in general. As a consequence of the Wasserstein approach, we also obtain perturbation estimates for geometrically ergodic Markov chains. We first adapt our Wasserstein perturbation bound to this setting. Then, as a second main result, Theorem 3.2, we prove a refined estimate for geometrically ergodic chains where the perturbation is measured by a weighted total variation distance. Our perturbation bounds, and earlier ones in [32,33], establish a direct connection between an exponential convergence property for Markov chains and their robustness to perturbations. In particular, fast convergence to stationarity implies insensitivity to perturbations in the transition probabilities. Geometric ergodicity has been studied extensively in the MCMC literature. Thus, our estimates can be used in combination with many existing convergence results for MCMC algorithms. In Section 4, we illustrate the applicability of both theorems by generalizing recent findings on approximate Metropolis–Hastings algorithms from [3] and on noisy Langevin algorithms for Gibbs random fields from [2].

1.1. Related literature

We refer to [20,21] for an overview of the classical literature on perturbation theory for Markov chains. However, as Stuart and Shardlow observed in [41], the classical assumptions on the perturbation might be too restrictive for many interesting applications. As a consequence, they develop a perturbation theory for geometrically ergodic Markov chains [41] which requires to control perturbations of iterated transition kernels in a weaker sense. In our bounds for geometrically ergodic Markov chains, we have similar flexibility in the perturbation due to the Lyapunov-type stability condition, and require only a control on the errors of one-step transition kernels.

Mitrophanov, in [33], considers uniformly ergodic Markov chains and provides the best estimates in those settings. In the geometrically ergodic case, there are further related results, see [13] and the references therein. Compared to [13], our focus is on non-asymptotic estimates with explicit constants, while their main focus is on qualitative results such as inheritance of geometric ergodicity by the perturbation. Earlier related results on perturbations induced by floating-point roundoff errors are shown in [7,38].

Finally, let us point out that our paper is complementary to the work of Pillai and Smith [35] who also present Wasserstein perturbation bounds for Markov chains. When moving beyond the uniformly ergodic Markov chain case, an important challenge is to handle the issue that in many applications suprema of relevant quantities over the whole state space are infinite. The authors of [35] guarantee finiteness of supremum norms by restricting attention to subsets of the state space. Their bounds thus involve exit probabilities from these subsets. Our approach circumvents these issues by relying on Lyapunov-type stability conditions for the approximate algorithm.

2. Wasserstein ergodicity

Let G be a Polish space and $\mathcal{B}(G)$ be the corresponding Borel σ-algebra. Let d be a metric, possibly different from the one which makes the space Polish, which is assumed to be lower semi-continuous with respect to the product topology of G. Let $\mathcal{P}$ be the set of all Borel probability measures on $(G, \mathcal{B}(G))$. Then, we define the Wasserstein distance of $\nu, \mu \in \mathcal{P}$ by

$$W(\nu, \mu) = \inf_{\xi \in M(\nu,\mu)} \int_G \int_G d(x, y) \, \mathrm{d}\xi(x, y),$$

where $M(\nu, \mu)$ is the set of all couplings of ν and μ, that is, all probability measures ξ on $G \times G$ with marginals ν and μ. Indeed, on $\mathcal{P}$ the Wasserstein distance satisfies the properties of a metric but is not necessarily finite, see [46], Chapter 6. For a measurable function $f : G \to \mathbb{R}$ we define

$$\|f\|_{\mathrm{Lip}} = \sup_{x,y \in G, x \neq y} \frac{|f(x) - f(y)|}{d(x, y)},$$

which leads to the well-known duality formula

$$W(\nu, \mu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \left| \int_G f(x) \big(\mathrm{d}\nu(x) - \mathrm{d}\mu(x)\big) \right|. \tag{2.1}$$

For details we refer to [45], Chapter 1.2. By δ_x we denote the probability measure concentrated at x. Hence $W(\delta_x, \delta_y) = d(x, y)$ is finite for $x, y \in G$.

Let P be a transition kernel on $(G, \mathcal{B}(G))$ which defines a linear operator $P : \mathcal{P} \to \mathcal{P}$ given by

$$\mu P(A) = \int_G P(x, A) \, \mathrm{d}\mu(x), \qquad \mu \in \mathcal{P}, A \in \mathcal{B}(G).$$

With this notation we have $\delta_x P(A) = P(x, A)$. Further, for a measurable function $f : G \to \mathbb{R}$ and $\mu \in \mathcal{P}$ we have

$$\int_G f(x) \, \mathrm{d}(\mu P)(x) = \int_G P f(x) \, \mathrm{d}\mu(x),$$

with $Pf(x) = \int_G f(y)P(x, \mathrm{d}y)$ whenever one of the integrals exist, see, for example, [40], Lemma 3.6. Now, by

$$\tau(P) := \sup_{x,y\in G, x\neq y} \frac{W(\delta_x P, \delta_y P)}{d(x,y)}$$

we define the *generalized ergodicity coefficient* of transition kernel P. This coefficient can be understood as a generalized Dobrushin ergodicity coefficient, see [8,9]. Dobrushin himself called $\tau(P)$ the Kantorovich norm of P, see [10], formula (14.34). Finally, $\tau(P)$ also provides a lower bound of the coarse Ricci curvature of P introduced in [34].

Two essential properties of the ergodicity coefficient are submultiplicativity and contractivity, see [10], Proposition 14.3 and Proposition 14.4.

Proposition 2.1. *For two transition kernels P and $\widetilde{P}$ on $(G, \mathcal{B}(G))$ and $\mu, \nu \in \mathcal{P}$, we have*

$$\tau(P\widetilde{P}) \leq \tau(P)\tau(\widetilde{P}) \qquad (\textit{Submultiplicativity}),$$

and

$$W(\nu P, \mu P) \leq \tau(P) W(\nu, \mu) \qquad (\textit{Contractivity}).$$

As an immediate consequence of this contractivity, we obtain the following corollary.

Corollary 2.1. *Let P be a transition kernel with stationary distribution π, that is, $\pi P = \pi$, and assume for some (and hence any) $x_0 \in G$ it holds that $\int_G d(x_0, x)\,\mathrm{d}\pi(x) < \infty$. Then*

$$\sup_{x\in G} \frac{W(\delta_x P, \pi)}{W(\delta_x, \pi)} \leq \tau(P). \tag{2.2}$$

Proof. Because of the assumption $\int_G d(x_0, x)\,\mathrm{d}\pi(x) < \infty$ we have that $W(\delta_x, \pi)$ is finite for any $x \in G$. Thus, the assertion follows by Proposition 2.1 and stationarity of π. □

Remark 2.1. For some special cases one also has an estimate of the form (2.2) in the other direction. To this end, consider the trivial metric $d(x,y) = 2 \cdot \mathbf{1}_{x\neq y}$ with indicator function

$$\mathbf{1}_{x\neq y} = \begin{cases} 1, & x \neq y, \\ 0, q & x = y. \end{cases}$$

Further, let

$$\|q\|_{\mathrm{tv}} := \sup_{\|f\|_\infty \leq 1} \left| \int_G f(y)\,\mathrm{d}q(y) \right| = 2 \sup_{A\in\mathcal{B}(G)} \left| q(A) \right|$$

be the total variation norm of a signed measure q on G. In this setting $W(\mu, \nu) = \|\mu - \nu\|_{\mathrm{tv}}$. For $x, y \in G$ with $x \neq y$, we have $\|\delta_x - \delta_y\|_{\mathrm{tv}} = d(x,y) = 2$ so that

$$\tau_1(P) = \frac{1}{2} \sup_{x,y\in G, x\neq y} \|\delta_x P - \delta_y P\|_{\mathrm{tv}}. \tag{2.3}$$

The "1" in the subscript of $\tau_1(P)$ indicates that we use the trivial metric. By applying the triangle inequality of the total variation norm we obtain $\tau_1(P) \le \sup_{x\in G} \|\delta_x P - \pi\|_{\text{tv}}$. If additionally π is atom-free, that is, $\pi(\{y\}) = 0$ for all $y \in G$, we have $\|\delta_y - \pi\|_{\text{tv}} = 2$. Then, the previous consideration and (2.2) lead to

$$\frac{1}{2} \sup_{x\in G} \|\delta_x P - \pi\|_{\text{tv}} \le \tau_1(P) \le \sup_{x\in G} \|\delta_x P - \pi\|_{\text{tv}}.$$

For the moment, let us assume that P is uniformly ergodic, that is, there exist numbers $\rho \in [0, 1)$ and $C \in (0, \infty)$ such that

$$\sup_{x\in G} \left\| \delta_x P^n - \pi \right\|_{\text{tv}} \le C\rho^n, \qquad n \in \mathbb{N}.$$

An immediate consequence of the uniform ergodicity is that $\tau_1(P^n) \le C\rho^n$.

Also note that if there is an $n_0 \in \mathbb{N}$ for which $\tau(P^{n_0}) < 1$ we have by the submultiplicativity, see Proposition 2.1, that $\tau(P^n)$ converges exponentially to zero. This motivates to impose the following assumption which contains the idea to measure convergence of $\delta_x P^n$ to π in terms of $\tau(P^n)$.

Assumption 2.1 (Wasserstein ergodicity). For the transition kernel P there exist numbers $\rho \in [0, 1)$ and $C \in (0, \infty)$ such that

$$\tau\left(P^n\right) = \sup_{x,y\in G, x\neq y} \frac{W(P^n(x, \cdot), P^n(y, \cdot))}{d(x, y)} \le C\rho^n, \qquad n \in \mathbb{N}. \tag{2.4}$$

For any probability measure $p_0 \in \mathcal{P}$, a transition kernel P with stationary distribution π and $p_n = p_0 P^n$ we have under the Wasserstein ergodicity condition that

$$W(p_n, \pi) \le C\rho^n W(p_0, \pi).$$

3. Perturbation bounds

By $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, we denote the non-negative integers and assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ mapping to a Polish space G equipped with a lower semi-continuous metric d. Let the sequence of random variables $(X_n)_{n\in\mathbb{N}_0}$ be a Markov chain with transition kernel P and initial distribution p_0, that is, we have almost surely

$$\mathbb{P}(X_n \in A \mid X_0, \dots, X_{n-1}) = \mathbb{P}(X_n \in A \mid X_{n-1}) = P(X_{n-1}, A), \qquad n \in \mathbb{N}$$

and $p_0(A) = \mathbb{P}(X_0 \in A)$ for any measurable set $A \subseteq G$. Assume that $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ is another Markov chain with transition kernel $\widetilde{P}$ and initial distribution $\widetilde{p}_0$. We denote by p_n the distribution of X_n and by $\widetilde{p}_n$ the distribution of $\widetilde{X}_n$. Throughout the paper, $(X_n)_{n\in\mathbb{N}}$ is considered to be the ideal, unperturbed Markov chain we would like to simulate while $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ is the perturbed Markov chain that we actually implement.

3.1. Wasserstein perturbation bound

Similar as in [33], Theorem 3.1, we show quantitative bounds on the difference of p_n and $\widetilde{p}_n$, but use the Wasserstein distance instead of total variation. Besides Assumption 2.1, the bounds depend on the difference of the initial distributions and on a suitably weighted one-step difference between P and $\widetilde{P}$.

Theorem 3.1 (Wasserstein perturbation bound). *Let Assumption 2.1 be satisfied with the numbers $C \in (0, \infty)$ and $\rho \in [0, 1)$, that is, $\tau(P^n) \le C\rho^n$. Assume that there are numbers $\delta \in (0, 1)$ and $L \in (0, \infty)$ and a measurable Lyapunov function $\widetilde{V} : G \to [1, \infty)$ of $\widetilde{P}$ such that*

$$(\widetilde{P}\widetilde{V})(x) \le \delta \widetilde{V}(x) + L. \tag{3.1}$$

Let

$$\gamma = \sup_{x \in G} \frac{W(\delta_x P, \delta_x \widetilde{P})}{\widetilde{V}(x)} \quad \text{and} \quad \kappa = \max\left\{ \widetilde{p}_0(\widetilde{V}), \frac{L}{1 - \delta} \right\}$$

with $\widetilde{p}_0(\widetilde{V}) = \int_G \widetilde{V}(x) \, \mathrm{d}\widetilde{p}_0(x)$. Then

$$W(p_n, \widetilde{p}_n) \le C\left(\rho^n W(p_0, \widetilde{p}_0) + \left(1 - \rho^n\right) \frac{\gamma\kappa}{1 - \rho} \right). \tag{3.2}$$

Proof. By induction one can show that

$$\widetilde{p}_n - p_n = (\widetilde{p}_0 - p_0)P^n + \sum_{i=0}^{n-1} \widetilde{p}_i (\widetilde{P} - P) P^{n-i-1}, \qquad n \in \mathbb{N}. \tag{3.3}$$

We have

$$W(\widetilde{p}_i P, \widetilde{p}_i \widetilde{P}) \le \int_G W(\delta_x P, \delta_x \widetilde{P}) \, \mathrm{d}\widetilde{p}_i(x) \le \gamma \int_G \widetilde{V}(x) \, \mathrm{d}\widetilde{p}_i(x).$$

Moreover, for $i \ge 0$ we have

$$\begin{aligned} \int_G \widetilde{V}(x) \, \mathrm{d}\widetilde{p}_i(x) &= \int_G \widetilde{P}^i \widetilde{V}(x) \, \mathrm{d}\widetilde{p}_0(x) \\ &\le \delta^i \widetilde{p}_0(\widetilde{V}) + \frac{L(1 - \delta^i)}{(1 - \delta)} \le \max\left\{ \widetilde{p}_0(\widetilde{V}), \frac{L}{1 - \delta} \right\} \end{aligned}$$

so that we obtain $W(\widetilde{p}_i P, \widetilde{p}_i \widetilde{P}) \le \gamma\kappa$. By this fact, we have

$$W\left(\widetilde{p}_i \widetilde{P} P^{n-i-1}, \widetilde{p}_i P P^{n-i-1}\right) \le \gamma\kappa \cdot \tau\left(P^{n-i-1}\right). \tag{3.4}$$

Then, by (3.3), (3.4) and the triangle inequality of the Wasserstein distance we have

$$\begin{aligned}W(p_n, \widetilde{p}_n) &\le W\big(p_0 P^n, \widetilde{p}_0 P^n\big) + \sum_{i=0}^{n-1} W\big(\widetilde{p}_i \widetilde{P} P^{n-i-1}, \widetilde{p}_i P P^{n-i-1}\big) \\ &\le W(p_0, \widetilde{p}_0)\tau\big(P^n\big) + \gamma\kappa \sum_{i=0}^{n-1} \tau\big(P^i\big).\end{aligned}$$

Finally, by (2.4) we obtain $\sum_{i=0}^{n-1} \tau(P^i) \le \frac{C(1-\rho^n)}{1-\rho}$, which allows us to complete the proof. □

Remark 3.1. The parameter κ is an upper bound on $\widetilde{p}_i(\widetilde{V})$. It can be interpreted as a measure for the stability of the perturbed Markov chain. The parameter γ quantifies with a weighted supremum norm the one-step difference between P and $\widetilde{P}$. The use of the Lyapunov function increases the flexibility of the resulting estimate, since larger values of $\widetilde{V}$ compensate larger values of the Wasserstein distance between the kernels. Notice that the existence of a Lyapunov function satisfying (3.1) is weaker than assuming $\widetilde{V}$-uniform ergodicity of $\widetilde{P}$ since it is not associated with a small set condition. In particular, the condition is satisfied for any $\widetilde{P}$ with the trivial choice $\widetilde{V}(x) = 1$ for all $x \in G$, see Corollary 3.2. As we will see in Section 4, allowing for non-trivial choices of $\widetilde{V}$ considerably increases the applicability of our results.

If $\widetilde{P}$ has a stationary distribution, say $\widetilde{\pi} \in \mathcal{P}$, as a consequence of the previous theorem, we obtain bounds on the difference between π and $\widetilde{\pi}$.

Corollary 3.1. *Let the assumptions of Theorem 3.2 be satisfied. Assume that $\widetilde{P}$ has a stationary distribution $\widetilde{\pi} \in \mathcal{P}$ and let $W(\pi, \widetilde{\pi})$ be finite. Then*

$$W(\pi, \widetilde{\pi}) \le \frac{\gamma C}{1-\rho} \cdot \frac{L}{1-\delta}. \tag{3.5}$$

Proof. By Theorem 3.2, we obtain with $p_0 = \pi$, $\widetilde{p}_0 = \widetilde{\pi}$, the stationarity of the distributions π, $\widetilde{\pi}$ and by letting $n \to \infty$ that

$$W(\pi, \widetilde{\pi}) \le \frac{C\gamma\kappa}{1-\rho}.$$

By the Lyapunov condition and [16], Proposition 4.24, it holds that

$$\widetilde{\pi}(\widetilde{V}) = \int_G \widetilde{V}(x)\, \mathrm{d}\widetilde{\pi}(x) \le \frac{L}{1-\delta}$$

which leads to $\kappa \le L/(1-\delta)$ and finishes the proof. □

Remark 3.2. It may seem artificial to assume $W(\pi, \widetilde{\pi}) < \infty$ but this is needed for the limit argument in the proof. This condition is often satisfied a priori. For example, it holds if the metric is bounded, that is, $\sup_{x,y\in G} d(x,y)$ is finite, or, more generally, if the distributions π and $\widetilde{\pi}$ possess a first moment in the sense that there exist $x_0, \widetilde{x}_0 \in G$ such that

$$\int_G d(x_0,x)\,\mathrm{d}\pi(x) < \infty, \qquad \int_G d(\widetilde{x}_0,x)\,\mathrm{d}\widetilde{\pi}(x) < \infty.$$

As pointed out in Remark 3.1, we do not need to impose condition (3.1) to obtain a non-trivial perturbation bound.

Corollary 3.2. *Assume that Assumption 2.1 holds with the numbers $C \in (0,\infty)$ and $\rho \in [0,1)$, that is, $\tau(P^n) \le C\rho^n$, and let*

$$\gamma := \sup_{x\in G} W(\delta_x P, \delta_x \widetilde{P}).$$

Then

$$W(p_n, \widetilde{p}_n) \le C\left(\rho^n W(p_0,\widetilde{p}_0) + \left(1-\rho^n\right)\frac{\gamma}{1-\rho}\right). \tag{3.6}$$

Proof. The statement follows by Theorem 3.1 with $\widetilde{V}(x) = 1$ and $L = 1-\delta$. □

Remark 3.3. For the trivial metric $d(x,y) = 2\cdot \mathbf{1}_{x\neq y}$ the last corollary states essentially the result of [33], Theorem 3.1, where instead of the general Wasserstein distance the total variation distance is used. There, the bound's dependence on C and ρ can be further improved by using the a priori bound $\tau_1(P^n) \le 1$ in addition to uniform ergodicity. For another metric d such an a priori bound is in general not available.

Remark 3.4. Table 1 provides a detailed comparison between our Theorem 3.1 and the related Wasserstein perturbation result of Pillai and Smith, [35], Lemma 3.3. An important ingredient in their estimate is a set $\widehat{G} \subseteq G$ which can be interpreted as the part of G where both Markov chains remain with high probability. When a good uniform upper bound on $W(\delta_x P, \delta_x \widetilde{P})$ for all $x \in G$ is available, we can choose $\widehat{G} = G$ in [35], Lemma 3.3, and $\widetilde{V}(x) = 1$ in Theorem 3.1. In that case, both results essentially simplify to Corollary 3.2. The results become entirely different when such a bound is not available or too rough. In our estimate, one then needs a non-trivial Lyapunov function for $\widetilde{P}$ and a uniform upper bound on $W(\delta_x P, \delta_x \widetilde{P})/\widetilde{V}(x)$. To apply their estimate, one needs a uniform bound on $W(\delta_x P, \delta_x \widetilde{P})$ for all $x \in \widehat{G}$. In addition, a bound on $\pi(G \setminus \widehat{G})$, Lyapunov functions and estimates of the exit probabilities from $\widehat{G}$ of both Markov chains need to be available. Finally, while [35], Lemma 3.3, requires slightly more regularity on the Lyapunov function, contractivity of the unperturbed transition kernel P (with $C = 1$) is not needed on the whole state space but only on $\widehat{G}$.

3.2. Perturbation bounds for geometrically ergodic Markov chains

In this section, we derive general perturbation bounds for geometrically ergodic Markov chains. First, we recall some results from [17], [26] and [36] which are helpful to apply our Wasserstein perturbation bounds in the geometrically ergodic case. Then we present the new estimates:

Table 1. Comparison of the Wasserstein perturbation bound of [35], Lemma 3.3, and Theorem 3.1. Here $\rho, \delta \in [0,1)$, $L, c_p, C, D \in (0,\infty)$, $V\colon G \to [0,\infty)$, $\widetilde{V}\colon G \to [1,\infty)$ and $E(x) = \int_G d(x,y)\,\mathrm{d}\pi(y)$.

	Assumptions of [35], Lemma 3.3	Assumptions of Theorem 3.1
Convergence property	$\exists \widehat{G} \subseteq G \quad \text{s.t.} \quad \sup_{x,y\in\widehat{G}} \frac{W(\delta_x P, \delta_y P)}{d(x,y)} \le \rho$	$\tau(P^n) \le C\rho^n$
Lyapunov function	$PV(x) \le \delta V(x) + L$ $\widetilde{P}V(x) \le \delta V(x) + L$	$\widetilde{P}\widetilde{V}(x) \le \delta \widetilde{V}(x) + L$
Drift regularity	$\mathbb{E}[V(X_{n+1} \mid X_n = x, X_{n+1} \notin \widehat{G})] \le C$ $\mathbb{E}[V(\widetilde{X}_{n+1} \mid \widetilde{X}_n = x, \widetilde{X}_{n+1} \notin \widehat{G})] \le C$ $\exists p \in \widehat{G} \quad \text{s.t.} \quad d(x,p) \le V(x) + c_p$	—
Perturbation error	$\widehat{\gamma} := \sup_{x\in\widehat{G}} W(\delta_x P, \delta_x \widetilde{P})$	$\gamma := \sup_{x\in G} \frac{W(\delta_x P, \delta_x \widetilde{P})}{\widetilde{V}(x)}$
Regularity of π	$\int_{G\setminus\widehat{G}} V(x)\mathrm{d}\pi(x) \le D$ $\pi(G \setminus \widehat{G})$ small	—
Conclusion: Upper bound of $W(\delta_x \widetilde{P}^n, \pi)$	$\rho^n E(x) + \frac{\widehat{\gamma}}{1-\rho} +$ $(\frac{2L}{1-\delta} + \delta^n (V(x) + D) + c_p)\pi(G \setminus \widehat{G}) +$ $2(1 - \mathbb{P}[\{X_j\}_{j=1}^{n-1} \cup \{\widetilde{X}_j\}_{j=1}^{n-1} \subseteq \widehat{G}])(C + \frac{L}{1-\delta} + c_p)$	$C\rho^n E(x) +$ $\frac{C\gamma}{1-\rho} \max\{\widetilde{V}(x), \frac{L}{1-\delta}\}$

- Corollary 3.3 is an application of Theorem 3.1 with Wasserstein distances replaced by V-norms of differences between measures.
- In Corollary 3.4, we show that having a Lyapunov function V for P is sufficient for our bounds if the transition kernels P and $\widetilde{P}$ are sufficiently close (in a suitable sense).
- In Theorem 3.2, we provide a quantitative perturbation bound which still applies if we can only control the total variation distance between $P(x,\cdot)$ and $\widetilde{P}(x,\cdot)$. To measure the perturbation in such a weak sense is new for geometrically ergodic Markov chains.

A transition kernel P with stationary distribution π is called geometrically ergodic if there is a constant $\rho \in [0,1)$ and a measurable function $C\colon G \to (0,\infty)$ such that for π-a.e. $x \in G$ we have

$$\left\| P^n(x,\cdot) - \pi \right\|_{\mathrm{tv}} \le C(x)\rho^n.$$

For ϕ-irreducible and aperiodic Markov chains, it is well known that geometric ergodicity is equivalent to V-uniform ergodicity, see [36], Proposition 2.1. Namely, if P is geometrically ergodic, then there exists a π-a.e. finite measurable function $V\colon G \to [1,\infty]$ with finite moments with respect to π and there are constants $\rho \in [0,1)$ and $C \in (0,\infty)$ such that

$$\left\| P^n(x,\cdot) - \pi \right\|_V := \sup_{|f|\le V} \left| \int_G f(y)\big(P^n(x,\mathrm{d}y) - \pi(\mathrm{d}y)\big) \right| \le C V(x)\rho^n, \qquad x \in G, n \in \mathbb{N}.$$

Thus,

$$\sup_{x\in G} \frac{\|P^n(x,\cdot) - \pi\|_V}{V(x)} \le C\rho^n. \tag{3.7}$$

The following result establishes the connection between V-norms and certain Wasserstein distances. It is basically due to Hairer and Mattingly [17], see also [26].

Lemma 3.1. *Assume that V is lower semi-continuous on G. For $x, y \in G$, let us define the metric*

$$d_V(x,y) = \big(V(x) + V(y)\big)\mathbf{1}_{x\neq y} = \begin{cases} V(x) + V(y), & x \neq y, \\ 0, & x = y. \end{cases}$$

Then, for any $\mu, \nu \in \mathcal{P}$ we have

$$\|\mu - \nu\|_V = W_{d_V}(\mu,\nu), \tag{3.8}$$

where W_{d_V} denotes the Wasserstein distance based on the metric d_V.

Lower semi-continuity of V implies lower semi-continuity of d_V, which leads to the duality formula (2.1) by [45], Theorem 1.14. We thus impose the standing assumption of lower semi-continuity of V whenever we speak of V-uniform ergodicity in the following. In principle, this requirement can be removed and (3.8) remains true, but we do not go into further detail in that direction. In applications, this is typically not restrictive since V is continuous anyway.

By similar arguments as in the proof of [26], Theorem 1.1, we observe that (3.7) implies a suitable upper bound on

$$\tau_V(P) = \sup_{x,y\in G, x\neq y} \frac{W_{d_V}(\delta_x P, \delta_y P)}{d_V(x,y)} = \sup_{x,y\in G, x\neq y} \frac{\|P(x,\cdot) - P(y,\cdot)\|_V}{V(x) + V(y)}.$$

Lemma 3.2. *If (3.7) is satisfied for the transition kernel P, then $\tau_V(P^n) \le C\rho^n$.*

Proof. For any positive real numbers a_1, a_2, b_1, b_2 we have the following elementary inequality

$$\frac{a_1 + a_2}{b_1 + b_2} \le \max\left\{\frac{a_1}{b_1}, \frac{a_2}{b_2}\right\}. \tag{3.9}$$

By (3.9), we obtain

$$\tau_V\big(P^n\big) = \sup_{x,y\in G, x\neq y} \frac{W_{d_V}(\delta_x P^n, \delta_y P^n)}{d_V(x,y)} \leq \sup_{x,y\in G, x\neq y} \frac{\|P^n(x,\cdot)-\pi\|_V + \|P^n(y,\cdot)-\pi\|_V}{V(x)+V(y)}$$

$$\leq \sup_{x,y\in G} \max\left\{\frac{\|P^n(x,\cdot)-\pi\|_V}{V(x)}, \frac{\|P^n(y,\cdot)-\pi\|_V}{V(y)}\right\} = \sup_{x\in G} \frac{\|P^n(x,\cdot)-\pi\|_V}{V(x)}.$$

Now, by using (3.7) we obtain the assertion. □

The lemmas above and Theorem 3.1 lead to the following new perturbation bound for geometrically ergodic Markov chains.

Corollary 3.3. *Let P be V-uniformly ergodic, that is, there are constants $\rho \in [0,1)$ and $C \in (0,\infty)$ such that*

$$\big\|P^n(x,\cdot)-\pi\big\|_V \leq CV(x)\rho^n, \qquad x \in G, n \in \mathbb{N}.$$

We also assume that there are numbers $\delta \in (0,1)$ and $L \in (0,\infty)$ and a measurable Lyapunov function $\widetilde{V} : G \to [1,\infty)$ of $\widetilde{P}$ such that

$$(\widetilde{P}\widetilde{V})(x) \leq \delta \widetilde{V}(x) + L. \tag{3.10}$$

Let

$$\gamma = \sup_{x\in G} \frac{\|P(x,\cdot)-\widetilde{P}(x,\cdot)\|_V}{\widetilde{V}(x)} \quad \text{and} \quad \kappa = \max\left\{\widetilde{p}_0(\widetilde{V}), \frac{L}{1-\delta}\right\}$$

with $\widetilde{p}_0(\widetilde{V}) = \int_G \widetilde{V}(x)\,\mathrm{d}\widetilde{p}_0(x)$. Then

$$\|p_n - \widetilde{p}_n\|_V \leq C\left(\rho^n\|p_0-\widetilde{p}_0\|_V + \big(1-\rho^n\big)\frac{\gamma\kappa}{1-\rho}\right). \tag{3.11}$$

Remark 3.5. In [41], Theorem 3.1, a related perturbation bound is proven. The convergence property of the unperturbed transition kernel is slightly weaker than our V-uniform ergodicity, but also based on a kind of Lyapunov function. More restrictively, there it is assumed that the difference of P^n and $\widetilde{P}^n$ for all $n > 0$ can be controlled. In addition, the perturbation error is measured with a weight given by the same Lyapunov function as in the convergence property of P, but by taking a supremum over a subset of test functions. With our approach, we can take the supremum over all test functions and obtain similar estimates by setting $p_0 = \pi$.

The next corollary demonstrates how the Lyapunov function of $\widetilde{P}$ can be replaced by a Lyapunov function of P, provided that the distance between the transition kernels is sufficiently small. Notice that assuming the existence of a Lyapunov function of P in addition to the V-uniform ergodicity is a definition of constants rather than an additional requirement, see, for example, [5].

Corollary 3.4. *Let P be V-uniformly ergodic, that is, there are constants $\rho \in [0,1)$ and $C \in (0,\infty)$ such that*

$$\left\| P^n(x,\cdot) - \pi \right\|_V \le C V(x) \rho^n, \qquad x \in G, n \in \mathbb{N}.$$

Moreover, $V : G \to [1,\infty)$ is a measurable Lyapunov function of P, such that

$$(PV)(x) \le \delta V(x) + L \tag{3.12}$$

with constants $\delta \in (0,1)$ and $L \in (0,\infty)$. Let

$$\gamma = \sup_{x\in G} \frac{\|P(x,\cdot) - \widetilde{P}(x,\cdot)\|_V}{V(x)} \quad \textit{and} \quad \kappa = \max\left\{ \widetilde{p}_0(V), \frac{L}{1-\delta-\gamma} \right\}$$

with $\widetilde{p}_0(V) = \int_G V(x)\,\mathrm{d}\widetilde{p}_0(x)$. If $\gamma + \delta < 1$, then

$$\|p_n - \widetilde{p}_n\|_V \le C\left(\rho^n \|p_0 - \widetilde{p}_0\|_V + \left(1-\rho^n\right) \frac{\gamma\kappa}{1-\rho} \right). \tag{3.13}$$

Proof. It suffices to show that

$$(\widetilde{P}V)(x) \le (\delta + \gamma)V(x) + L \tag{3.14}$$

and then to apply Corollary 3.3. We have

$$\big((\widetilde{P} - P)V\big)(x) \le \left| \big((\widetilde{P} - P)V\big)(x) \right| \le \left\| \widetilde{P}(x,\cdot) - P(x,\cdot) \right\|_V \le \gamma V(x)$$

which implies (3.14). The assertion follows by the assumption that $\delta + \gamma < 1$ and an application of Corollary 3.3. □

Remark 3.6. For discrete state spaces and under the requirement $p_0 = \widetilde{p}_0$, a result similar to the previous corollary is obtained in [21], Theorem 3, Corollary 3. The authors of [21] replace our constant κ by $\max_{0\le i\le n} \widetilde{p}_i(V)$. This we could do as well, see the proof of Theorem 3.1.

In the perturbation bound of Corollary 3.3, the function V plays two roles. In its first role, V appears in the V-uniform ergodicity condition and thus is used to quantify convergence of P. In its second role, V appears in the constant γ, with which we compare P and $\widetilde{P}$, as well as in the definition of the distance between p_n and $\widetilde{p}_n$. We can interpret γ of Corollary 3.3 as an operator norm of $P - \widetilde{P}$. To this end, let B_V be the set of all measurable functions $f : G \to \mathbb{R}$ with finite

$$|f|_V := \sup_{x\in G} \frac{|f(x)|}{V(x)}, \tag{3.15}$$

which means

$$B_V = \big\{ f : G \to \mathbb{R} \bigm| |f|_V < \infty \big\}.$$

It is easily seen that $(B_V, |\cdot|_V)$ is a normed linear space. In the setting of Corollary 3.3, we have

$$|||P - \widetilde{P}|||_{B_V \to B_{\widetilde{V}}} := \sup_{|f|_V \le 1} \left|(P - \widetilde{P})f\right|_{\widetilde{V}} = \gamma. \tag{3.16}$$

In Corollary 3.4, the more restrictive case $V = \widetilde{V}$ is considered. The corresponding operator norm $|||P - \widetilde{P}|||_{B_V \to B_V}$ appears in classical perturbation theory for Markov chains, see [20,21]. But as discussed in [41], page 1126, and [13] it might be too restrictive to measure the perturbation with this operator norm for $V = \widetilde{V}$.

By relying, for example, on [28], Proposition 2, we have some flexibility in the choice of V. There it is shown that, for $r \in (0, 1)$, V-uniform ergodicity implies V^r-uniform ergodicity. This leads to less favorable constants in the V^r-uniform ergodicity of P, but can relax the requirements on the similarity of P and $\widetilde{P}$. Namely, with a Lyapunov function $\widetilde{V}$ of $\widetilde{P}$ we can apply Corollary 3.3 with a V^r-uniformly ergodic P and $\gamma = |||P - \widetilde{P}|||_{B_{V^r} \to B_{\widetilde{V}}}$.

Unfortunately, this approach breaks down for $r = 0$. To see this, notice that V^r-uniform ergodicity with $r = 0$ is just uniform ergodicity which is not implied by geometric ergodicity. The next theorem overcomes this limitation by separating the two roles of the function V in the previous perturbation bounds. Roughly, we set $V = 1$ in the sense that we measure the distances between P and $\widetilde{P}$ as well as between p_n and $\widetilde{p}_n$ in the total variation distance. At the same time, we set $V = \widetilde{V}$ in the sense that we assume P is $\widetilde{V}$-uniformly ergodic with Lyapunov function $\widetilde{V}$.

Theorem 3.2. *Let P be $\widetilde{V}$-uniformly ergodic, that is, there are constants $\rho \in [0, 1)$ and $C \in (0, \infty)$ such that*

$$\left\|P^n(x, \cdot) - \pi\right\|_{\widetilde{V}} \le C\widetilde{V}(x)\rho^n, \qquad x \in G, n \in \mathbb{N}.$$

Moreover, $\widetilde{V}: G \to [1, \infty)$ is a measurable Lyapunov function of $\widetilde{P}$ and P, such that

$$(\widetilde{P}\widetilde{V})(x) \le \delta\widetilde{V}(x) + L \quad \text{and} \quad (P\widetilde{V})(x) \le \widetilde{V}(x) + L,$$

with constants $\delta \in (0, 1)$ and $L \in (0, \infty)$. Let

$$\gamma = \sup_{x \in G} \frac{\|P(x, \cdot) - \widetilde{P}(x, \cdot)\|_{\text{tv}}}{\widetilde{V}(x)} \quad \text{and} \quad \kappa = \max\left\{\widetilde{p}_0(\widetilde{V}), \frac{L}{1 - \delta}\right\} \tag{3.17}$$

with $\widetilde{p}_0(\widetilde{V}) = \int_G \widetilde{V}(x)\, \mathrm{d}\widetilde{p}_0(x)$. Then, for $\gamma \in (0, \exp(-1))$ we have

$$\|p_n - \widetilde{p}_n\|_{\text{tv}} \le C\rho^n\|p_0 - \widetilde{p}_0\|_{\widetilde{V}} + \frac{\kappa \exp(1)}{1 - \rho}\big(2C(L + 1)\big)^{\log(\gamma^{-1})^{-1}} \gamma \log\big(\gamma^{-1}\big). \tag{3.18}$$

Proof. From the proof of Theorem 3.2, we know that

$$\|\widetilde{p}_n - p_n\|_{\text{tv}} \le \left\|(\widetilde{p}_0 - p_0)P^n\right\|_{\text{tv}} + \sum_{i=0}^{n-1}\left\|\widetilde{p}_i(\widetilde{P} - P)P^{n-i-1}\right\|_{\text{tv}}, \qquad n \in \mathbb{N}.$$

By Lemma 3.2, we have

$$\big\|(\widetilde{p}_0 - p_0)P^n\big\|_{\mathrm{tv}} \le \big\|(\widetilde{p}_0 - p_0)P^n\big\|_{\widetilde{V}} \le C\rho^n \|\widetilde{p}_0 - p_0\|_{\widetilde{V}}.$$

Fix a real number $r \in (0,1)$ and let $s = 1-r$. By considering (2.3) one can see that $\tau_1(P) \le 1$. This leads to

$$\begin{aligned}\big\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\big\|_{\mathrm{tv}} &\le \big\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\big\|_{\mathrm{tv}}^r \big\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\big\|_{\widetilde{V}}^s \\ &\le \big\|\widetilde{p}_i(\widetilde{P}-P)\big\|_{\mathrm{tv}}^r \big\|\widetilde{p}_i(\widetilde{P}-P)\big\|_{\widetilde{V}}^s \tau_{\widetilde{V}}\big(P^{n-i-1}\big)^s.\end{aligned}$$

We also have

$$\begin{aligned}\big\|\widetilde{p}_i(\widetilde{P}-P)\big\|_{\mathrm{tv}} &\le \int_G \|\delta_x P - \delta_x \widetilde{P}\|_{\mathrm{tv}}\, \mathrm{d}\widetilde{p}_i(x) \le \gamma \int_G \widetilde{V}(x)\, \mathrm{d}\widetilde{p}_i(x), \\ \big\|\widetilde{p}_i(\widetilde{P}-P)\big\|_{\widetilde{V}} &\le \int_G W_{d_{\widetilde{V}}}(\delta_x P, \delta_x \widetilde{P})\, \mathrm{d}\widetilde{p}_i(x) \le \sup_{x\in G} \frac{W_{d_{\widetilde{V}}}(\delta_x P, \delta_x \widetilde{P})}{\widetilde{V}(x)} \int_G \widetilde{V}(x)\, \mathrm{d}\widetilde{p}_i(x).\end{aligned}$$

Moreover, for $i \ge 0$ we obtain

$$\int_G \widetilde{V}(x)\, \mathrm{d}\widetilde{p}_i(x) = \int_G \widetilde{P}^i \widetilde{V}(x)\, \mathrm{d}\widetilde{p}_0(x) \le \delta^i \widetilde{p}_0(\widetilde{V}) + \frac{L(1-\delta^i)}{(1-\delta)} \le \kappa,$$

and, by

$$\begin{aligned}W_{d_{\widetilde{V}}}(\delta_x P, \delta_x \widetilde{P}) &= \inf_{\xi \in M(\delta_x P, \delta_x \widetilde{P})} \int_G \int_G \big(\widetilde{V}(z) + \widetilde{V}(y)\big)\mathbf{1}_{z\neq y}\, \mathrm{d}\xi(y,z) \\ &\le P\widetilde{V}(x) + \widetilde{P}\widetilde{V}(x) \le (1+\delta)\widetilde{V}(x) + 2L,\end{aligned}$$

we have

$$\sup_{x\in G} \frac{W_{d_{\widetilde{V}}}(\delta_x P, \delta_x \widetilde{P})}{\widetilde{V}(x)} \le 2(L+1).$$

Then

$$\|\widetilde{p}_n - p_n\|_{\mathrm{tv}} \le C\rho^n \|\widetilde{p}_0 - p_0\|_{\widetilde{V}} + 2^s (L+1)^s \gamma^r \kappa \sum_{i=0}^{n-1} \tau_{\widetilde{V}}\big(P^i\big)^s.$$

Finally, by Lemma 3.2, we obtain

$$\sum_{i=0}^{n-1} \tau_{\widetilde{V}}\big(P^i\big)^s \le \frac{C^s(1-\rho^{ns})}{1-\rho^s} \le \frac{C^s}{1-\rho^s} \le \frac{C^s}{s(1-\rho)}.$$

For $\gamma \in (0, \exp(-1))$, we can choose the numbers $r = 1 + \log(\gamma)^{-1}$ and $s = \log(\gamma^{-1})^{-1}$. This yields $\gamma^r = \exp(1)\gamma$ and the proof is complete. □

Remark 3.7. Let $\widetilde{\pi} \in \mathcal{P}$ be a stationary distribution of $\widetilde{P}$. Notice that by the assumption that $\widetilde{V}$ is Lyapunov function of $\widetilde{P}$ and [16], Proposition 4.24, it follows that $\widetilde{\pi}(\widetilde{V}) \le L/(1-\delta)$. Further, by the $\widetilde{V}$-uniform ergodicity of P we also know that $\pi(\widetilde{V})$ is finite. Thus,

$$\|\pi - \widetilde{\pi}\|_{\widetilde{V}} \le \pi(\widetilde{V}) + \widetilde{\pi}(\widetilde{V}) < \infty.$$

Now, by Theorem 3.2, we can bound $\|\pi - \widetilde{\pi}\|_{\mathrm{tv}}$ with $p_0 = \pi$, $\widetilde{p}_0 = \widetilde{\pi}$ and by letting $n \to \infty$. We obtain

$$\|\pi - \widetilde{\pi}\|_{\mathrm{tv}} \le \frac{L(2C(L+1))^{\log(\gamma^{-1})^{-1}}}{(1-\delta)(1-\rho)} \exp(1)\gamma \log(\gamma^{-1}). \tag{3.19}$$

Remark 3.8. Let us comment on the dependence of γ. In Section 4.3, we apply Theorem 3.2 combined with (3.19) in a setting where we have $\gamma \le K \cdot \log(N)/N$ for a constant $K \ge 1$ and some parameter $N \in \mathbb{N}$ of the perturbed transition kernel. For $\varepsilon \in (0,1)$ and any $N > (K/\varepsilon)^{1/(1-\varepsilon)}$ we have $\gamma < \exp(-1)$. Then, with some simple calculations, we obtain for $p_0 = \widetilde{p}_0$ and $N > 6K^{3/2}$ the bound

$$\max\bigl\{\|p_n - \widetilde{p}_n\|_{\mathrm{tv}}, \|\pi - \widetilde{\pi}\|_{\mathrm{tv}}\bigr\} \le \frac{3\kappa(2C(L+1))^{2/\log(N)}}{1-\rho} \cdot \frac{K\log(N)^2}{N}.$$

Remark 3.9. In the setting of Theorem 3.2, we can also interpret γ as an operator norm. Namely,

$$|||P - \widetilde{P}|||_{B_1 \to B_{\widetilde{V}}} = \sup_{|f|_1 \le 1} \bigl|(P - \widetilde{P})f\bigr|_{\widetilde{V}} = \gamma. \tag{3.20}$$

Here the subscript "1" in $|f|_1$ indicates $V(x) = 1$ for all $x \in G$, see (3.15). For $\varepsilon_0 > 0$ and a family of perturbations $(\widetilde{P}_\varepsilon)_{|\varepsilon| \le \varepsilon_0}$ let $\gamma = |||P - \widetilde{P}_\varepsilon|||_{B_1 \to B_{\widetilde{V}}} \to 0$ for $\varepsilon \to 0$. This condition appears in [13], Theorem 1, condition (2), and is an assumption introduced by Keller and Liverani, see [22].

4. Applications

We illustrate our perturbation bounds in three different settings. We begin with studying an autoregressive process also considered in [13]. After this, we show quantitative perturbation bounds for approximate versions of two prominent MCMC algorithms, namely the Metropolis–Hastings and stochastic Langevin algorithms.

4.1. Autoregressive process

Let $G = \mathbb{R}$ and assume that $(X_n)_{n \in \mathbb{N}_0}$ is the autoregressive model defined by

$$X_n = \alpha X_{n-1} + Z_n, \qquad n \in \mathbb{N}. \tag{4.1}$$

Here X_0 is an $\mathbb{R}$-valued random variable, $\alpha \in (-1,1)$ and $(Z_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of random variables, independent of X_0. We also assume that the distribution of Z_1, say μ, admits a first moment. It is easily seen that $(X_n)_{n\in\mathbb{N}_0}$ is a Markov chain with transition kernel

$$P_\alpha(x,A) = \int_{\mathbb{R}} \mathbf{1}_A(\alpha x + y)\,\mathrm{d}\mu(y),$$

and it is well known that there exists a stationary distribution, say π_α, of P_α.

Now, let the transition kernel $P_{\widetilde{\alpha}}$ with $\widetilde{\alpha} \in (-1,1)$ be an approximation of P_α. For $x, y \in G$, let us consider the metric which is given by the absolute difference, that is, $d(x,y) = |x-y|$. We assume that $|\alpha - \widetilde{\alpha}|$ is small and study the Wasserstein distance, based on d, of $p_0 P_\alpha^n$ and $\widetilde{p}_0 P_{\widetilde{\alpha}}^n$ with two probability measures p_0 and $\widetilde{p}_0$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

We intend to apply Theorem 3.1. Notice that for $\widetilde{V}: \mathbb{R} \to [1,\infty)$ with $\widetilde{V}(x) = 1 + |x|$ we have

$$P_{\widetilde{\alpha}} \widetilde{V}(x) \le |\widetilde{\alpha}| \widetilde{V}(x) + 1 - |\widetilde{\alpha}| + \mathbb{E}|Z_1|$$

which guarantees that condition (3.1) is satisfied with $\delta = |\widetilde{\alpha}|$ and $L = 1 - |\widetilde{\alpha}| + \mathbb{E}|Z_1|$. Furthermore

$$W(\delta_x P_\alpha, \delta_y P_\alpha) \le \int_{\mathbb{R}} |\alpha x - z - \alpha y + z|\,\mathrm{d}\mu(z) \le |\alpha||x-y| = |\alpha| d(x,y),$$

leads to $\tau(P_\alpha^n) \le |\alpha|^n$. Similarly, one obtains

$$W(\delta_x P_\alpha, \delta_x P_{\widetilde{\alpha}}) \le \int_{\mathbb{R}} |\alpha x - z - \widetilde{\alpha} x + z|\,\mathrm{d}\mu(z) \le |x||\alpha - \widetilde{\alpha}|$$

which implies that

$$\sup_{x\in\mathbb{R}} \frac{W(\delta_x P_\alpha, \delta_x P_{\widetilde{\alpha}})}{\widetilde{V}(x)} \le |\alpha - \widetilde{\alpha}|.$$

We set

$$\kappa = 1 + \max\left\{ \int_{\mathbb{R}} |x|\,\mathrm{d}\widetilde{p}_0(x), \frac{\mathbb{E}|Z_1|}{1 - |\widetilde{\alpha}|} \right\}$$

and $p_{\alpha,n} = p_0 P_\alpha^n$, $\widetilde{p}_{\widetilde{\alpha},n} = \widetilde{p}_0 P_{\widetilde{\alpha}}^n$. Then, inequality (3.2) of Theorem 3.1 gives

$$W(p_{\alpha,n}, \widetilde{p}_{\widetilde{\alpha},n}) \le |\alpha|^n W(p_0, \widetilde{p}_0) + |\alpha - \widetilde{\alpha}| \frac{(1 - |\alpha|^n)\kappa}{1 - |\alpha|}, \tag{4.2}$$

and for $p_0 = \widetilde{p}_0$ we have

$$W(p_{\alpha,n}, \widetilde{p}_{\widetilde{\alpha},n}) \le |\alpha - \widetilde{\alpha}| \frac{(1 - |\alpha|^n)\kappa}{1 - |\alpha|}. \tag{4.3}$$

From the previous two inequalities one can see that if $\widetilde{\alpha}$ is sufficiently close to α, then the distance of the distribution $p_{\alpha,n}$ and $\widetilde{p}_{\widetilde{\alpha},n}$ is small. Let us emphasize here that we provide an explicit estimate rather than an asymptotic statement.

Note that by [16], Proposition 4.24, and the fact that $P_\beta g(x) \le |\beta| g(x) + \mathbb{E}|Z_1|$ with $g(x) = |x|$ and $\beta \in \{\alpha, \widetilde{\alpha}\}$ we obtain $\int_{\mathbb{R}} |x| \, \mathrm{d}\pi_\beta(x) < \infty$, which leads to a finite $W(\pi_\alpha, \pi_{\widetilde{\alpha}})$. As a consequence we obtain for the stationary distributions of P_α and $P_{\widetilde{\alpha}}$ by estimate (3.5) that

$$W(\pi_\alpha, \pi_{\widetilde{\alpha}}) \le |\alpha - \widetilde{\alpha}| \frac{1 - |\widetilde{\alpha}| + \mathbb{E}|Z_1|}{(1 - |\alpha|)(1 - |\widetilde{\alpha}|)}. \tag{4.4}$$

The dependence on $|\alpha - \widetilde{\alpha}|$ in the previous inequality cannot be improved in general. To see this, let us assume that $X_{0,\alpha}$ and $X_{0,\widetilde{\alpha}}$ are real-valued random variables with distribution π_α and $\pi_{\widetilde{\alpha}}$, respectively. Then, because of the stationarity we have that $X_{1,\alpha} = \alpha X_{0,\alpha} + Z_1$ and $X_{1,\widetilde{\alpha}} = \widetilde{\alpha} X_{0,\widetilde{\alpha}} + Z_1$ are also distributed according to π_α and $\pi_{\widetilde{\alpha}}$, respectively. Thus,

$$\mathbb{E}X_{0,\alpha} = \frac{\mathbb{E}Z_1}{1 - \alpha}, \qquad \mathbb{E}X_{0,\widetilde{\alpha}} = \frac{\mathbb{E}Z_1}{1 - \widetilde{\alpha}}.$$

Now, for $g : \mathbb{R} \to \mathbb{R}$ with $g(x) = x$, we have $\|g\|_{\mathrm{Lip}} \le 1$ and thus

$$\begin{aligned} W(\pi_\alpha, \pi_{\widetilde{\alpha}}) &= \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left| \int_G f(x) \big(\mathrm{d}\pi_\alpha(x) - \mathrm{d}\pi_{\widetilde{\alpha}}(x) \big) \right| \\ &\ge \left| \int_G x \big(\mathrm{d}\pi_\alpha(x) - \mathrm{d}\pi_{\widetilde{\alpha}}(x) \big) \right| = |\mathbb{E}X_{0,\alpha} - \mathbb{E}X_{0,\widetilde{\alpha}}| \\ &= |\alpha - \widetilde{\alpha}| \frac{|\mathbb{E}Z_1|}{|1 - \alpha||1 - \widetilde{\alpha}|}. \end{aligned}$$

Hence, whenever $\mathbb{E}Z_1 \neq 0$ we have a non-trivial lower bound with the same dependence on $|\alpha - \widetilde{\alpha}|$ as in the upper bound of (4.4). This fact shows that we cannot improve the upper bound.

Let us now discuss the application of Corollary 3.4 and Theorem 3.2. Under the additional assumption that μ, the distribution of Z_1, has a Lebesgue density h, it is shown in [15], Section 4, that the autoregressive model (4.1) is also $\widetilde{V}$-uniformly ergodic. Precisely, there is a constant $C \ge 1$ such that

$$\left\| P_\alpha^n(x, \cdot) - \pi_\alpha \right\|_{\mathrm{tv}} \le C |\alpha|^n \widetilde{V}(x).$$

Moreover, from [13], Example 1, we know that

$$\sup_{x \in \mathbb{R}} \frac{\|P_\alpha(x, \cdot) - P_{\widetilde{\alpha}}(x, \cdot)\|_{\widetilde{V}}}{\widetilde{V}(x)}$$

does not go to 0 when $\widetilde{\alpha} \downarrow \alpha$. Hence, Corollary 3.4 cannot quantify for small $|\widetilde{\alpha} - \alpha|$ whether the nth step distributions are close to each other. However, also in [13], Example 1, it is proven that

$$\sup_{x \in \mathbb{R}} \frac{\|P_\alpha(x, \cdot) - P_{\widetilde{\alpha}}(x, \cdot)\|_{\mathrm{tv}}}{\widetilde{V}(x)} \to 0 \qquad \text{if } \widetilde{\alpha} \to \alpha.$$

This indicates that Theorem 3.2 is applicable. By assuming in addition that h is *weakly unimodal*[1] and bounded from above by $h_{\max}$, we can quantify the result. Namely,

$$\begin{aligned}\sup_{x\in\mathbb{R}}\frac{\|P_\alpha(x,\cdot)-P_{\widetilde{\alpha}}(x,\cdot)\|_{\mathrm{tv}}}{\widetilde{V}(x)}&=\sup_{x\in\mathbb{R}}\frac{\|\mu(\cdot-\alpha x)-\mu(\cdot-\widetilde{\alpha}x)\|_{\mathrm{tv}}}{1+|x|}\\&=\sup_{x\in\mathbb{R}}\frac{\int_{\mathbb{R}}|h(z-\alpha x)-h(z-\widetilde{\alpha}x)|\,\mathrm{d}z}{1+|x|}\le 2|\alpha-\widetilde{\alpha}|h_{\max}.\end{aligned}$$

To see the final estimate, define $F(a)=\int_{\mathbb{R}}|h(z)-h(z-a)|\,\mathrm{d}z$ for $a\in\mathbb{R}$. By unimodality, there exists for any fixed $a\ge 0$ a constant c such that

$$\int_{\mathbb{R}}|h(z)-h(z-a)|\,\mathrm{d}z=\int_{-\infty}^{c}h(z)-h(z-a)\,\mathrm{d}z+\int_{c}^{\infty}h(z-a)-h(z)\,\mathrm{d}z.$$

The first summand on the right-hand side we can bound by

$$\int_{-\infty}^{c}h(z)\,\mathrm{d}z-\int_{-\infty}^{c}h(z-a)\,\mathrm{d}z=\int_{c-a}^{c}h(z)\,\mathrm{d}z\le ah_{\max}$$

and similarly for the second summand. Using that $F(a)=F(-a)$, we obtain $F(a)\le 2|a|h_{\max}$. Finally, by substitution we can write

$$\sup_{x\in\mathbb{R}}\frac{\int_{\mathbb{R}}|h(z-\alpha x)-h(z-\widetilde{\alpha}x)|\,\mathrm{d}z}{1+|x|}=|\alpha-\widetilde{\alpha}|\sup_{a\ge 0}\frac{F(a)}{a+|\alpha-\widetilde{\alpha}|}\le 2|\alpha-\widetilde{\alpha}|h_{\max}.$$

For simplicity set $p_0=\widetilde{p}_0$ and assume that $h_{\max}\le 1$ as well as $|\alpha-\widetilde{\alpha}|\in(0,\exp(-1)/2)$. Then, Theorem 3.2 implies

$$\max\bigl\{\|p_{\alpha,n}-\widetilde{p}_{\widetilde{\alpha},n}\|_{\mathrm{tv}},\|\pi_\alpha-\pi_{\widetilde{\alpha}}\|_{\mathrm{tv}}\bigr\}\le\frac{\kappa\exp(1)}{1-|\alpha|}\bigl(2C\bigl(\mathbb{E}|Z_1|+2\bigr)\bigr)|\alpha-\widetilde{\alpha}|\log\bigl(|\alpha-\widetilde{\alpha}|^{-1}\bigr)$$

which seems to be new.

4.2. Approximate Metropolis–Hastings algorithms

We apply our perturbation results to the approximate (or noisy) Metropolis–Hastings algorithms analyzed in [2–4,23,29,35]. We assume either that the unperturbed transition kernel of the Metropolis–Hastings algorithm satisfies the Wasserstein ergodicity condition stated in Assumption 2.1 or is geometrically ergodic. In particular, we do not assume that the transition kernel is uniformly ergodic. Let π be a probability distribution on $(G,\mathcal{B}(G))$ and assume that we are interested in sampling realizations from this distribution. Let Q be a transition kernel

[1] The function $h\colon\mathbb{R}\to[0,\infty)$ is called *weakly unimodal* if there exists $s\in\mathbb{R}$ such that $h(x)$ is nondecreasing for $x\in(-\infty,s)$ and nonincreasing for $x\in(s,\infty)$.

which serves as the proposal for the Metropolis–Hastings algorithm. From [44], Proposition 1, we know that there exists a set $S \subset G \times G$ such that we can define the "acceptance ratio" for $(x, y) \in G \times G$ as

$$r(x,y) := \begin{cases} \dfrac{\pi(\mathrm{d}y)Q(y,\mathrm{d}x)}{\pi(\mathrm{d}x)Q(x,\mathrm{d}y)}, & (x,y) \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

Then, let the acceptance probability be $\alpha(x, y) = \min\{1, r(x, y)\}$. With this notation the Metropolis–Hastings algorithm defines a transition kernel

$$P_\alpha(x, \mathrm{d}y) = Q(x, \mathrm{d}y)\alpha(x, y) + \delta_x(\mathrm{d}y)s_\alpha(x), \tag{4.6}$$

with

$$s_\alpha(x) = 1 - \int_G \alpha(x, y)Q(x, \mathrm{d}y).$$

We provide a step of a Markov chain $(X_n)_{n\in\mathbb{N}_0}$ with transition kernel P_α in algorithmic form.

Algorithm 4.1. A single transition from X_n to X_{n+1} of the Metropolis–Hastings algorithm works as follows:

1. Draw a sample $Y \sim Q(X_n, \cdot)$ and $U \sim \mathrm{Unif}[0, 1]$ independently, call the result y and u.
2. Set $r := r(X_n, y)$, with the ratio $r(\cdot, \cdot)$ defined in (4.5).
3. If $u < r$, then accept the proposal, and set $X_{n+1} := y$, else reject the proposal and set $X_{n+1} := X_n$.

Now, suppose we are unable to evaluate $r(x, y)$, so that we are forced to work with an approximation of $\alpha(x, y)$. The key idea behind approximate Metropolis–Hastings algorithms is to replace $r(x, y)$ by a non-negative random variable R with distribution, say $\mu_{x,y,u}$, depending on $x, y \in G$ and $u \in [0, 1]$. For concrete choices of the random variable R, we refer to [2–4,23]. We present a step of the corresponding Markov chain $(\widetilde{X}_n)_{n\in\mathbb{N}}$ in algorithmic form.

Algorithm 4.2. A single transition from $\widetilde{X}_n$ to $\widetilde{X}_{n+1}$ works as follows:

1. Draw a sample $Y \sim Q(\widetilde{X}_n, \cdot)$ and $U \sim \mathrm{Unif}[0, 1]$ independently, call the result y and u.
2. Draw a sample $R \sim \mu_{\widetilde{X}_n,y,u}$, call the result $\widetilde{r}$.
3. If $u < \widetilde{r}$, then accept the proposal, and set $\widetilde{X}_{n+1} := y$, else reject the proposal and set $\widetilde{X}_{n+1} := \widetilde{X}_n$.

The algorithm has acceptance probability

$$\widetilde{\alpha}(x, y) = \mathbb{E}\mathbf{1}_{[0,\min\{1,R\}]}(U) = \int_0^1 \int_0^\infty \mathbf{1}_{[0,\min\{1,\widetilde{r}\}]}(u)\, \mathrm{d}\mu_{x,y,u}(\widetilde{r})\, \mathrm{d}u$$

and the transition kernel of such a Markov chain is still of the form (4.6) with $\alpha(x, y)$ substituted by $\widetilde{\alpha}(x, y)$, that is, it is given by $P_{\widetilde{\alpha}}$. The following results hold in the slightly more general case where $\widetilde{\alpha}(x, y)$ is any approximation of the acceptance probability $\alpha(x, y)$.

The next lemma provides an estimate for the Wasserstein distance between transition kernels of the form (4.6) in terms of the acceptance probabilities.

Lemma 4.1. *Let Q be a transition kernel on $(G, \mathcal{B}(G))$ and let $\alpha\colon G \times G \to [0,1]$ and $\widetilde{\alpha}\colon G \times G \to [0,1]$ be measurable functions. By P_α and $P_{\widetilde{\alpha}}$ we denote the transition kernels of the form* (4.6) *with acceptance probabilities α and $\widetilde{\alpha}$. Then, for all $x \in G$, we have*

$$W(\delta_x P_\alpha, \delta_x P_{\widetilde{\alpha}}) \le \int_G d(x,y)\mathcal{E}(x,y)Q(x,\mathrm{d}y)$$

with $\mathcal{E}(x,y) = |\alpha(x,y) - \widetilde{\alpha}(x,y)|$.

Proof. By the use of the dual representation of the Wasserstein distance it follows that

$$\begin{aligned} W(\delta_x P_\alpha, \delta_x P_{\widetilde{\alpha}}) &= \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left| \int_G f(y)\big(P_\alpha(x,\mathrm{d}y) - P_{\widetilde{\alpha}}(x,\mathrm{d}y)\big) \right| \\ &= \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left| \int_G \big(f(y) - f(x)\big)\big(\alpha(x,y) - \widetilde{\alpha}(x,y)\big) Q(x,\mathrm{d}y) \right| \\ &\le \int_G d(x,y)\mathcal{E}(x,y)Q(x,\mathrm{d}y). \end{aligned}$$

□

By the previous lemma and Theorem 3.1, we obtain the following Wasserstein perturbation bound for the approximate Metropolis–Hastings algorithm.

Corollary 4.1. *Let Q be a transition kernel on $(G, \mathcal{B}(G))$ and let $\alpha\colon G \times G \to [0,1]$ and $\widetilde{\alpha}\colon G \times G \to [0,1]$ be measurable functions. By P_α and $P_{\widetilde{\alpha}}$ we denote the transition kernels of the form* (4.6) *with acceptance probabilities α and $\widetilde{\alpha}$. Let the following conditions be satisfied*:

- *Assumption 2.1 holds for the transition kernel P_α, that is, $\tau(P_\alpha^n) \le C\rho^n$ for $\rho \in [0,1)$ and $C \in (0,\infty)$.*
- *There are numbers $\delta \in (0,1)$, $L \in (0,\infty)$ and a measurable Lyapunov function $\widetilde{V}: G \to [1,\infty)$ of $P_{\widetilde{\alpha}}$, that is,*

$$(P_{\widetilde{\alpha}}\widetilde{V})(x) \le \delta\widetilde{V}(x) + L. \tag{4.7}$$

- *Let $\mathcal{E}(x,y) = |\alpha(x,y) - \widetilde{\alpha}(x,y)|$ and assume that*

$$\gamma = \sup_{x \in G} \frac{\int_G d(x,y)\mathcal{E}(x,y)Q(x,\mathrm{d}y)}{\widetilde{V}(x)} < \infty. \tag{4.8}$$

Then, for any $p_0 \in \mathcal{P}$ and finite $p_0(\widetilde{V}) = \int_G \widetilde{V}(x)\,\mathrm{d}p_0(x)$ we have

$$W\big(p_0 P_\alpha^n, p_0 P_{\widetilde{\alpha}}^n\big) \le \frac{\gamma\kappa C(1-\rho^n)}{1-\rho},$$

where $\kappa = \max\{p_0(\widetilde{V}), \frac{L}{1-\delta}\}$.

Let us point out several aspects of condition (4.7). Recall that (4.7) is always satisfied with $\widetilde{V}(x) = 1$ for all $x \in G$. However, in this case it seems more difficult to control γ. If some additional knowledge in form of a Lyapunov function $V : G \to [1, \infty)$ of P_α, that is, $P_\alpha V(x) \leq \delta V(x) + L$ for some $\delta \in (0, 1)$ and $L \in (0, \infty)$, is available, then a non-trivial candidate for $\widetilde{V}$ is V. For sufficiently small

$$\delta_V = \sup_{z \in G} \int_G \left(\frac{V(y)}{V(z)} + 1 \right) \mathcal{E}(z, y) Q(z, \mathrm{d}y)$$

this is indeed true. Namely, we have

$$\big|(P_\alpha - P_{\widetilde{\alpha}})V(x)\big| \leq \int_G V(y)\mathcal{E}(x, y)Q(x, \mathrm{d}y) + V(x) \int_G \mathcal{E}(x, y)Q(x, \mathrm{d}y) \leq V(x)\delta_V.$$

Then, $P_{\widetilde{\alpha}} V(x) \leq (\delta + \delta_V)V(x) + L$ and whenever $\delta + \delta_V < 1$ it is clear that condition (4.7) is verified.

To highlight the usefulness of a non-trivial Lyapunov function, we consider the following scenario which is related to a local perturbation of an independent Metropolis–Hastings algorithm.

Example 4.1. Let us assume that for P_α Assumption 2.1, as formulated in Corollary 4.1, is satisfied. For some probability measure μ on $(G, \mathcal{B}(G))$ define $Q(x, \cdot) = \mu$ and $p_0 = \widetilde{p}_0 = \mu$. For $\widetilde{G} \subseteq G$ let

$$\widetilde{\alpha}(x, y) = \min\big\{1, \alpha(x, y) + \mathbf{1}_{\widetilde{G}}(x)\big\}.$$

Hence, for $x \in \widetilde{G}$ the transition kernel $P_{\widetilde{\alpha}}(x, \cdot)$ accepts any proposed state and for $x \notin \widetilde{G}$ we have $P_{\widetilde{\alpha}}(x, \cdot) = P_\alpha(x, \cdot)$. It is easily seen that $\mathcal{E}(x, y) \leq \mathbf{1}_{\widetilde{G}}(x)$. For arbitrary $R > 0$ and $r \in (0, 1)$ set $\widetilde{V}(x) = 1 + R\mathbf{1}_{\widetilde{G}}(x)$ and note that

$$P_{\widetilde{\alpha}} \widetilde{V}(x) \leq r\widetilde{V}(x) + 1 - r + RP_{\widetilde{\alpha}}(x, \widetilde{G}) \leq r\widetilde{V}(x) + 1 - r + R\mu(\widetilde{G}).$$

The last inequality of the previous formula follows by distinguishing the cases $x \in \widetilde{G}$ and $x \notin \widetilde{G}$. Define $D(\widetilde{G}) = \sup_{x \in \widetilde{G}} \int_G d(x, y)\mu(\mathrm{d}y)$ and observe

$$\kappa = 1 + \frac{R\mu(\widetilde{G})}{1 - r} \quad \text{and} \quad \gamma \leq \frac{D(\widetilde{G})}{1 + R}.$$

Then, Corollary 4.1 leads to

$$W\big(p_0 P_\alpha^n, p_0 P_{\widetilde{\alpha}}^n\big) \leq \frac{C}{1 - \rho} \left(1 + \frac{R\mu(\widetilde{G})}{1 - r} \right) \frac{D(\widetilde{G})}{1 + R}$$

for arbitrary $R \in (0, \infty)$ and $r \in (0, 1)$. Under the assumption that $D(\widetilde{G})$ is finite and letting $R \to \infty$ as well as $r \downarrow 0$ we obtain

$$W\big(p_0 P_\alpha^n, p_0 P_{\widetilde{\alpha}}^n\big) \leq \frac{C\mu(\widetilde{G})D(\widetilde{G})}{1 - \rho},$$

which tells us that basically $\mu(\widetilde{G})$ measures the difference of the distributions. A small perturbation set $\widetilde{G}$ with respect to μ, thus implies a small bias. In contrast, with the trivial Lyapunov function $\widetilde{V} = 1$, and if there is $(x, y) \in \widetilde{G} \times G$ such that $\alpha(x, y) = 0$, we only obtain

$$\gamma\kappa = D(\widetilde{G}) \geq \inf_{x \in G} \int_G d(x, y)\mu(\mathrm{d}y).$$

The resulting upper bound on $W(p_0 P_\alpha^n, p_0 P_{\widetilde{\alpha}}^n)$ will typically be bounded away from zero regardless of the set $\widetilde{G}$.

Remark 4.1. The constant γ essentially depends on the distance $d(x, y)$ and the difference of the acceptance probabilities $\mathcal{E}(x, y)$. By applying the Cauchy–Schwarz inequality to the numerator of γ, we can separate the two parts, that is,

$$\int_G d(x, y)\mathcal{E}(x, y) Q(x, \mathrm{d}y) \leq \left(\int_G d(x, y)^2 Q(x, \mathrm{d}y) \cdot \int_G \mathcal{E}(x, y)^2 Q(x, \mathrm{d}y) \right)^{1/2}.$$

If both integrals remain finite, we see that an appropriate control of $\mathcal{E}(x, y)$ suffices for making the constant γ small.

Remark 4.2. By using a Hoeffding-type bound, in Bardenet *et al.* [3], Lemma 3.1, it is shown that for their version of the approximate Metropolis–Hastings algorithm with adaptive subsampling the approximation error $\mathcal{E}(x, y)$ is bounded uniformly in x and y by a constant $s > 0$. Moreover, s can be chosen arbitrarily small for the implementation of the algorithm.

Now we consider the case where the unperturbed transition kernel P_α is geometrically ergodic. Motivated by Remark 4.2, we also assume that $\mathcal{E}(x, y) \leq s$ for a sufficiently small number $s > 0$. The following corollary generalizes a main result of Bardenet *et al.* [3], Proposition 3.2, to the geometrically ergodic case.

Corollary 4.2. *Let Q be a transition kernel on $(G, \mathcal{B}(G))$ and let $\alpha\colon G \times G \to [0, 1]$ and $\widetilde{\alpha}\colon G \times G \to [0, 1]$ be measurable functions. By P_α and $P_{\widetilde{\alpha}}$ we denote the transition kernels of the form* (4.6) *with acceptance probabilities α and $\widetilde{\alpha}$. Let the following conditions be satisfied*:

- *The unperturbed transition kernel P_α is V-uniformly ergodic, that is,*

$$\left\| P_\alpha^n(x, \cdot) - \pi \right\|_V \leq C V(x) \rho^n, \qquad x \in G, n \in \mathbb{N}$$

for numbers $\rho \in [0, 1)$, $C \in (0, \infty)$ and a measurable function $V\colon G \to [1, \infty)$. Moreover, V is a Lyapunov function of P_α, that is,

$$(P_\alpha V)(x) \leq \delta V(x) + L \tag{4.9}$$

for numbers $\delta \in (0, 1)$ and $L \in (0, \infty)$.

- *A uniform bound $s > 0$ on the difference of the acceptance probabilities is given, that is, for all $x, y \in G$, we have*

$$\mathcal{E}(x, y) = |\alpha(x, y) - \widetilde{\alpha}(x, y)| \le s.$$

- *The constant λ satisfies*

$$\lambda = 1 + \sup_{x \in G} \int_G \frac{V(y)}{V(x)} Q(x, \mathrm{d}y) < \infty.$$

If $s < (1 - \delta)/\lambda$, then, for any $p_0 \in \mathcal{P}$ with finite $\kappa = \max\{p_0(V), \frac{L}{1-\delta-\lambda s}\}$ we have

$$\left\| p_0 P_\alpha^n - p_0 P_{\widetilde{\alpha}}^n \right\|_V \le \frac{\lambda s \kappa C(1 - \rho^n)}{1 - \rho}.$$

Proof. We consider the metric d_V, defined in Lemma 3.1, set $V = \widetilde{V}$ and use $\mathcal{E}(x, y) \le s$ so that it is easily seen that the constant γ from Corollary 4.1 satisfies $\gamma \le s\lambda$. From the proof of Corollary 3.4, we know that V is a Lyapunov function of $P_{\widetilde{\alpha}}$ provided that $\gamma + \delta < 1$. Thus, we have

$$P_{\widetilde{\alpha}} V(x) \le (\delta + \lambda s)V(x) + L. \tag{4.10}$$

Now if $s < (1 - \delta)/\lambda$, then $\delta + \lambda s < 1$ and the assertion follows from Corollary 4.1 by writing the Wasserstein distances in terms of V-norms as in Section 3.2. □

Remark 4.3. Without $V(x)$ in the denominator, that is, if we had relied on Corollary 3.2 instead of Theorem 3.1, the constant λ would often be infinite. Consider the following toy example: Let π be the exponential distribution with density $\exp(-x)$ on $G = [0, \infty)$ and assume that $Q(x, \mathrm{d}y)$ is a uniform proposal with support $[x - 1, x + 1]$. With $V(x) = \exp(x)$ it is well known that the Metropolis–Hastings algorithm is V-uniformly ergodic, see [30] or [37], Example 4. In this example,

$$\lambda \le 1 + \sup_{x \in [0,\infty)} \int_{x-1}^{x+1} \exp(y - x)\,\mathrm{d}y \le 1 + \exp(1)$$

whereas $\int_{x-1}^{x+1} \exp(y)\,\mathrm{d}y$ is unbounded in x. Notice that λ only depends on the unperturbed Markov chain so that a bound on λ can be combined with any approximation.

Remark 4.4. Let $P_{\widetilde{\alpha}}$ and P_α be ϕ-irreducible and aperiodic. Then, one can prove under the assumptions of Corollary 4.2 that $P_{\widetilde{\alpha}}$ is V-uniformly ergodic if s is sufficiently small. To see this, note that by [31], Theorem 16.0.1, the V-uniform ergodicity of P_α implies that P_α satisfies their drift condition (V4). By the arguments stated in the proof of Corollary 3.4, one obtains that $P_{\widetilde{\alpha}}$ also satisfies (V4) for sufficiently small s and this implies V-uniform ergodicity. In this case, clearly $P_{\widetilde{\alpha}}$ possesses a stationary distribution, say $\widetilde{\pi}$, and

$$\|\pi - \widetilde{\pi}\|_V \le \frac{\lambda s C}{1 - \rho} \cdot \frac{L}{1 - \delta - \lambda s}.$$

The previous inequality follows by (3.5) and the fact that

$$\|\pi - \widetilde{\pi}\|_V \le \pi(V) + \widetilde{\pi}(V) < \infty.$$

Here the finiteness of $\pi(V)$ follows by the V-uniform ergodicity of P and $\widetilde{\pi}(V) \le L/(1-\delta-\lambda s)$ follows by (4.10) and [16], Proposition 4.24.

4.3. Noisy Langevin algorithm for Gibbs random fields

An alternative to the Metropolis–Hastings algorithm is the Langevin algorithm, see [39]. Unfortunately, in its implementation one needs the gradient of the density of the target distribution. To overcome this problem, different approximate Langevin algorithms have been proposed and studied, see [1,2,43,47].

This section is mainly based on Alquier *et al.* [2], Section 3.4, where a noisy Langevin algorithm for Gibbs random fields is considered. We provide a quantitative version of [2], Theorem 3.2. The setting is as follows. Let $\mathcal{Y}$ be a finite set and with $M \in \mathbb{N}$ let $y = \{y_1, \dots, y_M\} \in \mathcal{Y}^M$ be an observed data set on nodes $\{1, \dots, M\}$ of a certain graph. The likelihood of y with parameter $\theta \in \mathbb{R}$ is defined by

$$\ell(y \mid \theta) = \frac{\exp(\theta s(y))}{\sum_{y \in \mathcal{Y}^M} \exp(\theta s(y))},$$

where $s \colon \mathcal{Y}^M \to \mathbb{R}$ is a given statistic. The density of the posterior distribution with respect to the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given the data $y \in \mathcal{Y}^M$ is determined by

$$\pi_y(\theta) := \pi(\theta \mid y) \propto \ell(y \mid \theta)\, p(\theta),$$

where the prior density $p(\theta)$ is the Lebesgue density of the normal distribution $\mathcal{N}(0, \sigma_p^2)$ with $\sigma_p > 0$.

We consider the Langevin algorithm, a first order Euler discretization of the SDE of the Langevin diffusion, see [39]. It is given by $(X_n)_{n \in \mathbb{N}_0}$ with

$$X_n = X_{n-1} + \frac{\sigma^2}{2} \nabla \log \pi_y(X_{n-1}) + Z_n, \qquad n \in \mathbb{N}. \tag{4.11}$$

Here X_0 is a real-valued random variable and $(Z_n)_{n \in \mathbb{N}}$ is an i.i.d. sequence of random variables, independent of X_0, with $Z_n \sim \mathcal{N}(0, \sigma^2)$ for a parameter $\sigma > 0$ which can be interpreted as the step size in the discretization of the diffusion. It is easily seen that $(X_n)_{n \in \mathbb{N}_0}$ is a Markov chain with transition kernel

$$P_\sigma(\theta, A) = \int_{\mathbb{R}} \mathbf{1}_A\left(\theta + \frac{\sigma^2}{2} \nabla \log \pi_y(\theta) + z\right) \mathcal{N}\big(0, \sigma^2\big)(\mathrm{d}z), \qquad A \in \mathcal{B}(\mathbb{R}).$$

In general π_y is not a stationary distribution of P_σ, but there exists a stationary distribution (see Proposition 4.1 below), say π_σ, which is close to π_y depending on σ. Let $z(\theta) =$

$\sum_{y\in\mathcal{Y}^M}\exp(\theta s(y))$ then, by the definition of π_y we have

$$\begin{aligned}\log\pi_y(\theta) &= \theta s(y) - \log z(\theta) + \log p(\theta) - \log\left(\int_{\mathbb{R}} \ell(y \mid z)p(z)\,\mathrm{d}z\right),\\ \nabla\log\pi_y(\theta) &= s(y) - \frac{z'(\theta)}{z(\theta)} + \nabla\log p(\theta) = s(y) - \frac{\sum_{z\in\mathcal{Y}^M} s(z)\exp(\theta s(z))}{\sum_{z\in\mathcal{Y}^M}\exp(\theta s(z))} - \frac{\theta}{\sigma_p^2}\\ &= s(y) - \mathbb{E}_{\ell(\cdot|\theta)}s(Y) - \frac{\theta}{\sigma_p^2},\end{aligned}$$

where Y is a random variable on $\mathcal{Y}^M$ distributed according the likelihood distribution determined by $\ell(\cdot \mid \theta)$. We do not have access to the exact value of the mean $\mathbb{E}_{\ell(\cdot|\theta)}s(Y)$ since in general we do not know the normalizing constant of the likelihood. We assume that we can use a Monte Carlo estimate. For $N\in\mathbb{N}$ let $(Y_i)_{1\le i\le N}$ be an i.i.d. sequence of random variables with $Y_i \sim \ell(\cdot \mid \theta)$ independent of $(Z_n)_{n\in\mathbb{N}}$ from (4.11). Then, $\frac{1}{N}\sum_{i=1}^N s(Y_i)$ is an approximation of $\mathbb{E}_{\ell(\cdot|\theta)}s(Y)$ which leads to an estimate of $\nabla\log\pi_y(\theta)$ given by

$$\widehat{\nabla}^N \log\pi_y(\theta) := s(y) - \frac{1}{N}\sum_{i=1}^N s(Y_i) - \frac{\theta}{\sigma_p^2}.$$

We substitute $\nabla\log\pi_y(\theta)$ by $\widehat{\nabla}^N\log\pi_y(\theta)$ in (4.11) and obtain a sequence of random variables $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ defined by

$$\begin{aligned}\widetilde{X}_n &= \widetilde{X}_{n-1} + \frac{\sigma^2}{2}\widehat{\nabla}^N\log\pi_y(\widetilde{X}_{n-1}) + Z_n\\ &= \left(1 - \frac{\sigma^2}{2\sigma_p^2}\right)\widetilde{X}_{n-1} + \frac{\sigma^2}{2}\left(s(y) - \frac{1}{N}\sum_{i=1}^N s(Y_i)\right) + Z_n.\end{aligned}$$

The sequence $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ is again a Markov chain with transition kernel

$$\begin{aligned}P_{\sigma,N}(\theta, A) = \int_{\mathbb{R}} \sum_{(y_1',\dots,y_N')\in\mathcal{Y}^{MN}} \mathbf{1}_A\left(\left(1 - \frac{\sigma^2}{2\sigma_p^2}\right)\theta + \frac{\sigma^2}{2}\left(s(y) - \frac{1}{N}\sum_{i=1}^N s(y_i')\right) + z\right)\\ \times \prod_{i=1}^N \ell(\theta \mid y_i')\mathcal{N}(0,\sigma^2)(\mathrm{d}z)\end{aligned}$$

for $\theta\in\mathbb{R}$ and $A\in\mathcal{B}(\mathbb{R})$. Let us state a transition of this noisy Langevin Markov chain according to $P_{\sigma,N}$ in algorithmic form.

Algorithm 4.3. A single transition from $\widetilde{X}_n$ to $\widetilde{X}_{n+1}$ works as follows:

1. Draw an i.i.d. sequence $(Y_i)_{1\le i\le N}$ with $Y_i \sim \ell(\cdot \mid \widetilde{X}_n)$, call the result $(y_1',\dots,y_N')$.

2. Calculate

$$\widehat{\nabla}^N \log \pi_y(\widetilde{X}_n) := s(y) - \frac{1}{N}\sum_{i=1}^{N} s\big(y_i'\big) - \frac{\widetilde{X}_n}{\sigma_p^2}.$$

3. Draw $Z_n \sim \mathcal{N}(0,\sigma^2)$, independent from step 1., call the result z_n. Set

$$\widetilde{X}_{n+1} = \widetilde{X}_n + \frac{\sigma^2}{2}\widehat{\nabla}^N \log \pi_y(\widetilde{X}_n) + z_n.$$

From [2], Lemma 3, and by applying arguments of [39], we obtain the following facts about the noisy Langevin algorithm.

Proposition 4.1. *Let $\|s\|_\infty = \sup_{z\in\mathcal{Y}^M}|s(z)|$ be finite with $\|s\|_\infty > 0$, let $V : \mathbb{R} \to [1,\infty)$ be given by $V(\theta) = 1 + |\theta|$ and assume that $\sigma^2 < 4\sigma_p^2$. Then*

1. *the function V is a Lyapunov function for P_σ and $P_{\sigma,N}$. We have*

$$P_\sigma V(\theta) \le \delta V(\theta) + L\mathbf{1}_I(\theta), \qquad P_{\sigma,N} V(\theta) \le \delta V(\theta) + L\mathbf{1}_I(\theta) \tag{4.12}$$

with $\delta = 1 - \frac{\sigma^2}{4\sigma_p^2}$, $L = \sigma + \sigma^2\|s\|_\infty + \frac{\sigma^2}{2\sigma_p^2}$ and the interval

$$I = \left\{\theta \in \mathbb{R} \;\middle|\; |\theta| \le 1 + 4\sigma_p^2\|s\|_\infty + \frac{4\sigma_p^2}{\sigma}\right\};$$

2. *there are distributions π_σ and $\pi_{\sigma,N}$ on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ which are stationary with respect to P_σ and $P_{\sigma,N}$, respectively*;
3. *the transition kernels P_σ and $P_{\sigma,N}$ are V-uniformly ergodic*;
4. *for $N > 4\max\{\|s\|_\infty^2\sigma^4, \|s\|_\infty^{-3}\sigma^{-6}\}$ we have*

$$\sup_{\theta\in\mathbb{R}}\big\|P_\sigma(\theta,\cdot) - P_{\sigma,N}(\theta,\cdot)\big\|_{\mathrm{tv}} \le 6\max\big\{\|s\|_\infty\sigma^2, \|s\|_\infty^{-2}\sigma^{-4}\big\}\frac{\log(N)}{N}. \tag{4.13}$$

Proof. We use the same arguments as in [39], Section 3.1. One can easily see that the Markov chains $(X_n)_{n\in\mathbb{N}_0}$ and $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ are irreducible with respect to the Lebesgue measure and weak Feller. Thus, all compact sets are petite, see [31], Proposition 6.2.8. Hence, for the existence of stationary distributions, say π_σ and $\pi_{\sigma,N}$, [31], Theorem 12.3.3, as well as for the V-uniform ergodicity [31], Theorem 16.0.1, it is enough to show that V satisfies (4.12). With $Z \sim \mathcal{N}(0,\sigma^2)$, we have

$$\begin{aligned} P_\sigma V(\theta) &\le \left(1 - \frac{\sigma^2}{2\sigma_p^2}\right)V(\theta) + \frac{\sigma^2}{2\sigma_p^2} + \frac{\sigma^2}{2}\big|s(y) - \mathbb{E}_{\ell(\cdot|\theta)}s(Y)\big| + \mathbb{E}|Z| \\ &\le \left(1 - \frac{\sigma^2}{2\sigma_p^2}\right)V(\theta) + \frac{\sigma^2}{2\sigma_p^2} + \sigma^2\|s\|_\infty + \sigma \end{aligned}$$

$$\leq \left(1-\frac{\sigma^2}{2\sigma_p^2}\right)V(\theta)+\max\left\{\frac{\sigma^2}{4\sigma_p^2}V(\theta), \frac{\sigma^2}{2\sigma_p^2}+\sigma^2\|s\|_\infty+\sigma\right\}$$
$$\leq \left(1-\frac{\sigma^2}{4\sigma_p^2}\right)V(\theta)+\left(\frac{\sigma^2}{2\sigma_p^2}+\sigma^2\|s\|_\infty+\sigma\right)\cdot \mathbf{1}_I(\theta).$$

By the fact that

$$\mathbb{E}\left[\left|s(y)-\frac{1}{N}\sum_{i=1}^N s(Y_i)\right| \mid \widetilde{X}_n=\theta\right]\leq 2\|s\|_\infty$$

we obtain with the same arguments that

$$P_{\sigma,N}V(\theta)\leq \delta V(\theta)+L\cdot \mathbf{1}_I(\theta).$$

Thus, the assertions from 1. to 3. are proven. The statement of 4. is a consequence of [2], Lemma 3. There it is shown that for $N>4\|s\|_\infty^2\sigma^4$ it holds that

$$\sup_{\theta\in\mathbb{R}}\left\|P_\sigma(\theta,\cdot)-P_{\sigma,N}(\theta,\cdot)\right\|_{\text{tv}}\leq \exp\left(\frac{\log(N)}{4N\|s\|_\infty^2\sigma^4}\right)-1+\frac{4\sqrt{\pi}\|s\|_\infty\sigma^2}{N}.$$

By using $\exp(\theta)-1\leq\theta\exp(\theta)$ and $N>4$, we further estimate the right-hand side by

$$\left(\frac{K_{N,s,\sigma}}{4\|s\|_\infty^2\sigma^4}+\frac{4\sqrt{\pi}\|s\|_\infty\sigma^2}{\log(5)}\right)\cdot\frac{\log(N)}{N}\qquad \text{with } K_{N,s,\sigma}=\exp\left(\frac{\log(N)}{4N\|s\|_\infty^2\sigma^4}\right).$$

Since $\log(N)\cdot N^{-1/3}<2$, we have the bound $K_{N,s,\sigma}\leq\exp(1)$ provided that $4N^{2/3}\|s\|_\infty^2\sigma^4\geq 2$ which follows from $N\geq\|s\|_\infty^{-3}\sigma^{-6}$. The assertion of (4.13) follows now by a simple calculation. □

By using the facts collected in the previous proposition, we can apply the perturbation bound of Theorem 3.2 and obtain a quantitative perturbation bound for the noisy Langevin algorithm.

Corollary 4.3. *Let p_0 be a probability measure on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ and set $p_n=p_0P_\sigma^n$ as well as $\widetilde{p}_{n,N}=p_0P_{\sigma,N}^n$. Suppose that $\sigma^2<4\sigma_p^2$. Then, there are numbers $\rho\in[0,1)$ and $C\in(0,\infty)$, independent of n, N, determining*

$$R:=\frac{18\max\{\|s\|_\infty\sigma^2,\|s\|_\infty^{-2}\sigma^{-4}\}}{1-\rho}\cdot\left(2+\max\left\{\mathbb{E}_{p_0}|X|, 4\sigma_p^2\left(\|s\|_\infty+\sigma^{-1}\right)\right\}\right)$$

with $\mathbb{E}_{p_0}|X|=\int_{\mathbb{R}}|\theta|\,\mathrm{d}p_0(\theta)$, so that for $N>90\max\{\|s\|_\infty^2\sigma^4,\|s\|_\infty^{-3}\sigma^{-6}\}$ we have

$$\max\left\{\|p_n-\widetilde{p}_{n,N}\|_{\text{tv}},\|\pi_\sigma-\pi_{\sigma,N}\|_{\text{tv}}\right\}\leq R\cdot\left(2C\left(\sigma+\sigma^2\|s\|_\infty+3\right)\right)^{2/\log(N)}\frac{\log(N)^2}{N}.$$

Proof. We have by Proposition 4.1 that P_σ is V-uniformly ergodic with $V(\theta) = 1 + |\theta|$, that is, there are numbers $\rho \in [0, 1)$ and $C \in (0, \infty)$ such that

$$\sup_{\theta\in\mathbb{R}} \frac{\|P_\sigma^n(\theta, \cdot) - \pi_\sigma\|_V}{V(\theta)} \le C\rho^n.$$

Now, by combining Theorem 3.2 and Remark 3.8 with the results from Proposition 4.1 we obtain the result. □

Remark 4.5. We want to point out that the assumptions imposed are the same as in [2], Theorem 3.2, but instead of the asymptotic result we provide an explicit estimate. The numbers $\rho \in [0, 1)$ and $C \in (0, \infty)$ are not stated in terms of the model parameters. In principle, these values can be derived from the drift condition (4.12) through [5], Theorem 1.1.

Acknowledgements

We thank Alexander Mitrophanov and the referees for their valuable comments which helped to improve the paper. D. Rudolf was supported by the DFG Research Training Group 2088.

References

[1] Ahn, S., Korattikara, A. and Welling, M. (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the* 29*th International Conference on Machine Learning*.

[2] Alquier, P., Friel, N., Everitt, R. and Boland, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Stat*. *Comput*. **26** 29–47. MR3439357

[3] Bardenet, R., Doucet, A. and Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the* 31*st International Conference on Machine Learning* 405–413.

[4] Bardenet, R., Doucet, A. and Holmes, C. (2015). On Markov chain Monte Carlo methods for tall data. arXiv:1505.02827.

[5] Baxendale, P.H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann*. *Appl*. *Probab*. **15** 700–738. MR2114987

[6] Betancourt, M. (2015). The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *Proceedings of the* 32*nd International Conference on Machine Learning* 533–540.

[7] Breyer, L., Roberts, G.O. and Rosenthal, J.S. (2001). A note on geometric ergodicity and floating-point roundoff error. *Statist*. *Probab*. *Lett*. **53** 123–127. MR1843871

[8] Dobrushin, R.L. (1956). Central limit theorem for non-stationary Markov chains. I. *Teor*. *Veroyatn*. *Primen*. **1** 72–89.

[9] Dobrushin, R.L. (1956). Central limit theorem for nonstationary Markov chains. II. *Teor*. *Veroyatn*. *Primen*. **1** 365–425. MR0097112

[10] Dobrushin, R.L. (1996). Perturbation methods of the theory of Gibbsian fields. In *Lectures on Probability Theory and Statistics*: *Ecole d'Eté de Probabilités de Saint-Flour XXIV* – 1994. *Lecture Notes in Mathematics* **1648** 1–66. Berlin: Springer. MR1600880

[11] Durmus, A. and Moulines, E. (2015). Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis Adjusted Langevin Algorithm. *Stat*. *Comput*. **25** 5–19.

[12] Eberle, A. (2014). Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Ann*. *Appl*. *Probab*. **24** 337–377. MR3161650

[13] Ferré, D., Hervé, L. and Ledoux, J. (2013). Regular perturbation of V-geometrically ergodic Markov chains. *J*. *Appl*. *Probab*. **50** 184–194. MR3076780

[14] Gibbs, A.L. (2004). Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stoch*. *Models* **20** 473–492. MR2094049

[15] Guibourg, D., Hervé, L. and Ledoux, J. (2012). Quasi-compactness of Markov kernels on weighted-supremum spaces and geometrical ergodicity. arXiv:1110.3240v5.

[16] Hairer, M. (2006). Ergodic properties of Markov processes. Lecture notes, Univ. Warwick. Available at http://www.hairer.org/notes/Markov.pdf.

[17] Hairer, M. and Mattingly, J.C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis*, *Random Fields and Applications VI*. *Progress in Probability* **63** 109–117. Basel: Birkhäuser/Springer Basel AG. MR2857021

[18] Hairer, M., Stuart, A.M. and Vollmer, S.J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann*. *Appl*. *Probab*. **24** 2455–2490.

[19] Johndrow, J., Mattingly, J.C., Mukherjee, S. and Dunson, D. (2015). Approximations of Markov chains and Bayesian inference. arXiv:1508.03387.

[20] Kartashov, N.V. (1986). Inequalities in stability and ergodicity theorems for Markov chains with a common phase space. I. *Theory Probab*. *Appl*. **30** 247–259.

[21] Kartashov, N.V. and Golomozyĭ, V. (2013). Maximal coupling procedure and stability of discrete Markov chains. I. *Theory Probab*. *Math*. *Statist*. **86** 93–104.

[22] Keller, G. and Liverani, C. (1999). Stability of the spectrum for transfer operators. *Ann*. *Sc*. *Norm*. *Super*. *Pisa Cl*. *Sci*. (4) **28** 141–152. MR1679080

[23] Korattikara, A., Chen, Y. and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *Proceedings of the* 31*st International Conference on Machine Learning* 181–189.

[24] Lee, A., Doucet, A. and Łatuszyński, K. (2014). Perfect simulation using atomic regeneration with application to sequential Monte Carlo. arXiv:1407.5770.

[25] Madras, N. and Sezer, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16** 882–908. MR2730652

[26] Mao, Y., Zhang, M. and Zhang, Y. (2013). A generalization of Dobrushin coefficient. *Chinese J*. *Appl*. *Probab*. *Statist*. **29** 489–494. MR3156322

[27] Marin, J.-M., Pudlo, P., Robert, C.P. and Ryder, R.J. (2012). Approximate Bayesian computational methods. *Stat*. *Comput*. **22** 1167–1180. MR2992292

[28] Mathé, P. (2004). Numerical integration using V-uniformly ergodic Markov chains. *J*. *Appl*. *Probab*. **41** 1104–1112.

[29] Medina-Aguayo, F.J., Lee, A. and Roberts, G.O. (2016). Stability of noisy Metropolis–Hastings. *Stat*. *Comput*. **26** 1187–1211. MR3538632

[30] Mengersen, K.L. and Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann*. *Statist*. **24** 101–121. MR1389882

[31] Meyn, S.P. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: Cambridge Univ. Press.

[32] Mitrophanov, A.Yu. (2003). Stability and exponential convergence of continuous-time Markov chains. *J*. *Appl*. *Probab*. **40** 970–979. MR2012680

[33] Mitrophanov, A.Yu. (2005). Sensitivity and convergence of uniformly ergodic Markov chains. *J*. *Appl*. *Probab*. **42** 1003–1014.

[34] Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **256** 810–864.

[35] Pillai, N. and Smith, A. (2015). Ergodicity of approximate MCMC chains with applications to large data sets. arXiv:1405.0182v2.

[36] Roberts, G.O. and Rosenthal, J.S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25. MR1448322

[37] Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565

[38] Roberts, G.O., Rosenthal, J.S. and Schwartz, P.O. (1998). Convergence properties of perturbed Markov chains. *J. Appl. Probab.* **35** 1–11. MR1622440

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

[40] Rudolf, D. (2012). Explicit error bounds for Markov chain Monte Carlo. Dissertationes Math. **485** 93 pp.

[41] Shardlow, T. and Stuart, A.M. (2000). A perturbation theory for ergodic Markov chains and application to numerical approximations. *SIAM J. Numer. Anal.* **37** 1120–1137. MR1756418

[42] Singh, S., Wick, M. and McCallum, A. (2012). Monte Carlo MCMC: Efficient inference by approximate sampling. In *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 1104–1113. Stroudsburg, PA: Association for Computational Linguistics.

[43] Teh, Y.W., Thiery, A.H. and Vollmer, S.J. (2016). Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.* **17** Art. ID 7. MR3482927

[44] Tierney, L. (1998). A note on the Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9.

[45] Villani, C. (2003). *Topics in Optimal Transportation. Graduate Studies in Mathematics* **58**. Providence, RI: Amer. Math. Soc. MR1964483

[46] Villani, C. (2009). *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften* **338**. Berlin: Springer.

[47] Welling, M. and Teh, Y.W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the* 28*th International Conference on Machine Learning* 681–688.

Received October 2015 and revised January 2017