Perturbation Bounds for Monte Carlo within Metropolis via Restricted Approximations

Felipe Medina-Aguayo[*] Daniel Rudolf[†]
Nikolaus Schweizer[‡]

Abstract

The Monte Carlo within Metropolis (MCwM) algorithm, interpreted as a perturbed Metropolis-Hastings (MH) algorithm, provides a simple approach for approximate sampling when the target distribution is doubly-intractable or contains latent variables. We assume that the associated unperturbed Markov chain is geometrically ergodic and show explicit estimates of the difference between the n-th step distributions of the perturbed MCwM and the unperturbed MH chains. These bounds are based on novel perturbation results for Markov chains which are of interest beyond the MCwM setting. To apply the bounds, we need to control the difference between the transition probabilities of the two Markov chains, at least in the center of the state-space. Moreover, we need to verify stability of the perturbed chain, either through a Lyapunov condition, or by restricting it to the center of the state space.

1 Introduction

The *Metropolis–Hastings* (MH) algorithm is a classical method for sampling approximately from a distribution of interest relying only on point-wise evaluations of an *unnormalized* density. However, when even this unnormalized

[*] Department of Mathematics and Statistics, University of Reading Whiteknights, PO Box 220, Reading RG6 6AX, United Kingdom, email: f.j.medinaaguayo@reading.ac.uk

[†] Institute for Mathematical Stochastics, Universität Göttingen & Felix-Bernstein-Institute for Mathematical Statistics, Goldschmidtstraße 7, 37077 Göttingen, Germany, email: daniel-rudolf@uni-goettingen.de

[‡] Department of Econometrics and OR, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands, email: n.f.f.schweizer@uvt.nl

density depends on unknown integrals and cannot easily be evaluated, then this approach is not feasible. A possible solution is to replace the required density evaluations in the MH acceptance ratio with suitable approximations. This idea is implemented in *Monte Carlo within Metropolis* (MCwM) algorithms which substitute the unnormalized density evaluations by Monte Carlo estimates for the intractable integrals.

Yet in general, replacing the exact MH acceptance ratio by an approximation leads to inexact algorithms in the sense that the resulting Markov chain does not have the distribution of interest as stationary one and might not converge to a stationary distribution at all. Nonetheless, these approximate or noisy methods, see e.g. [AFEB16, JM17b, JMMD15], have recently gained increased attention due to their applicability in certain intractable sampling problems. In this work we attempt to answer the following questions about the MCwM algorithm:

- Can one quantify the quality of MCwM algorithms?
- When might the MCwM algorithm fail and what can one do in such situations?

Regarding the first question, by using bounds on the difference of the n-th step distributions of MH and MCwM algorithm we give a positive answer which improves upon earlier results in the literature, see below. For the second question, we suggest a modification for stabilizing the MCwM approach by restricting the Markov chain to a suitably chosen set that contains the "essential part", or the "center", of the state space. We provide examples where this restricted version of MCwM converges towards the distribution of interest while the unrestricted version does not. Note also that in practical implementations of Markov chain Monte Carlo on a computer, simulated chains are effectively restricted to compact state spaces due to memory limitations. Our results on restricted approximations can also be read in this spirit.

Our overall approach is based on *perturbation theory* for Markov chains. Let $(X_n)_{n\in\mathbb{N}_0}$ be a Markov chain with transition kernel P and $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ be a Markov chain with transition kernel $\widetilde{P}$ on a common Polish space $(G, \mathcal{B}(G))$. We think of P and $\widetilde{P}$ as "close" to each other in a suitable sense and consider $\widetilde{P}$ as a perturbation of P. In order to quantify the difference of the distributions of X_n and $\widetilde{X}_n$, denoted by p_n and $\widetilde{p}_n$ respectively, we work with

$$\|p_n - \widetilde{p}_n\|_{\text{tv}}\,, \tag{1}$$

where $\|\cdot\|_{\text{tv}}$ denotes the total variation distance. The Markov chain $(X_n)_{n\in\mathbb{N}_0}$ can be interpreted as unavailable, unperturbed, ideal, while $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ is a perturbation that is available for simulation. We focus on the case where the ideal Markov chain is *geometrically ergodic*, more precisely *V-uniformly ergodic*, and satisfies a *Lyapunov condition* of the form

$$PV(x) \leq \delta V(x) + L, \qquad x \in G,$$

for some function $V\colon G \to [1,\infty)$ and numbers $\delta \in [0,1), L \in [1,\infty)$.

To obtain estimates of (1) we need two assumptions which can be informally explained as follows:

1. *Closeness of $\widetilde{P}$ and P:*
   The difference of $\widetilde{P}$ and P is measured by controlling either a uniformly weighted total variation distance or a uniformly weighted V-norm of $P(x,\cdot) - \widetilde{P}(x,\cdot)$, see $\varepsilon_{\text{tv},W}$, $\varepsilon_{V,W}$ in Lemma 4 and Theorem 8. In Theorem 10, uniformity refers at least to the essential region of the state space.

2. *Stability of $\widetilde{P}$:*
   A stability condition on $\widetilde{P}$ is satisfied either in the form of a Lyapunov condition as in Theorem 8 or by restriction to the center of the state space determined by V as in Theorem 10.

Explicit bounds on (1) are provided in Section 3 in Proposition 7, Theorem 8 and Theorem 10. More precisely, in Theorem 7 a Lyapunov condition on $\widetilde{P}$ is imposed whereas in Theorem 8 a *restricted approximation* $\widetilde{P}$ is considered. We now briefly comment on related literature to such perturbation estimates for V-uniformly ergodic Markov chains.

In contrast to the V-uniform ergodicity assumption we impose on the ideal Markov chain, the results in [AFEB16, JMMD15, Mit05] only cover perturbations of uniformly ergodic Markov chains. Nonetheless, perturbation theoretical questions for geometrically ergodic Markov chains have been studied before, see e.g. [BRR01, FHL13, MLR16, NR17, RRS98, RS18, SS00] and the references therein. A crucial aspect where those papers differ from each other is how one measures the closeness of the transitions of the unperturbed and perturbed Markov chains to have applicable estimates, see the discussion about that in [SS00, FHL13, RS18]. Our Theorem 7 and Theorem 8 refine and extend the results of [RS18, Theorem 3.2]. In particular, in Theorem 8 we take a restriction to the center of the state space into account. Let us also mention here that [PS14, RS18] contain related results under Wasserstein ergodicity assumptions. More recently, [JM17a] studies approximate chains using notions of maximal couplings, [NR17] extends the uniformly ergodic setting from [JMMD15] to using L_2 norms instead of total variation, and [JM17b] explores bounds on the approximation error of time averages.

In Section 4 we apply our perturbation bounds in the context of approximate sampling via MCwM. The goal is to (approximately) sample from a target distribution π on G, which is determined by an unnormalized density function $\pi_u\colon G \to [0,\infty)$ w.r.t a reference measure μ, that is,

$$\pi(A) = \frac{\int_A \pi_u(x)\,\mathrm{d}\mu(x)}{\int_G \pi_u(x)\,\mathrm{d}\mu(x)}, \quad A \in \mathcal{B}(G).$$

Classically the method of choice is to construct a Markov chain $(X_n)_{n\in\mathbb{N}_0}$ based on the MH algorithm for approximate sampling of π. This algorithm crucially relies on knowing (at least) the ratio $\pi_u(y)/\pi_u(x)$ for arbitrary $(x,y) \in G^2$, e.g., because $\pi_u(x)$ and $\pi_u(y)$ can readily be computed. However, in some scenarios, only approximations of $\pi_u(x)$ and $\pi_u(y)$ are available. Replacing the true unnormalized density π_u in the MH algorithm by an approximation yields a perturbed, "inexact" Markov chain $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$. If the approximation is based on a Monte Carlo method, the perturbed chain is called MCwM chain.

Two particular settings where approximations of π_u may rely on Monte Carlo estimates are *doubly-intractable distributions* and *latent variables*. Examples of the former occur in Markov or Gibbs random fields, where the function values of the unnormalized density $\pi_u(x)$ itself are only known up to a factor $Z(x)$. This means that

$$\pi_u(x) = \rho(x)/Z(x), \qquad x \in G, \tag{2}$$

where only values of $\rho(x)$ can easily be computed while the computational problem lies in evaluating

$$Z(x) = \int_{\mathcal{Y}} \bar{\rho}(x,y) r_x(\mathrm{d}y),$$

where $\mathcal{Y}$ denotes an auxiliary variable space, $\bar{\rho}\colon G \times \mathcal{Y} \to [0,\infty)$ and r_x is a σ-finite measure. We study the MCwM algorithm, which in every transition uses an iid sequence of random variables $(Y_i^{(x)})_{1\leq i\leq N}$, with $Y_1^{(x)} \sim r_x$, to approximate $Z(x)$ by $\widehat{Z}_N(x) := \frac{1}{N}\sum_{i=1}^N \bar{\rho}(x, Y_i^{(x)})$ (and $Z(y)$ by $\widehat{Z}_N(y)$, respectively).

Let us emphasize that the doubly-intractable case can also be approached algorithmically from various other perspectives. For instance, instead of estimating the normalizing constant $Z(x)$ in (2), one could estimate unbiasedly $(Z(x))^{-1}$ whenever exact simulation from the Markov or Gibbs random field is possible. In this case, $\pi_u(x)$ turns into a Monte Carlo estimate which can formally be analyzed with exactly the same techniques as the latent variable scenario described below. Yet another algorithmic possibility is explored in the *noisy exchange* algorithm of [AFEB16], where ratios of the form $Z(x)/Z(y)$ are approximated by a single Monte Carlo estimate. Their algorithm is motivated by the *exchange algorithm* [MGM06] which, perhaps surprisingly, can avoid the need for evaluating the ratio $Z(x)/Z(y)$ and targets the distribution π exactly, see [EJREH17] for a brief review of these and related methods. However, in some cases the exchange algorithm performs poorly, see [AFEB16]. Then approximate sampling methods for distributions of the form (2) might prove useful as long as the introduced bias is not too large. As a final remark in this direction, the recent work [ADYC18] considers a correction of the noisy exchange algorithm which produces a Markov chain with stationary distribution π.

The second setting we study arises from *latent variables.* Here, $\pi_u(x)$ cannot be evaluated since it takes the form

$$\pi_u(x) = \int_{\mathcal{Y}} \bar{\rho}(x, y)\, r_x(\mathrm{d}y), \tag{3}$$

where r_x is a distribution on a measurable space $\mathcal{Y}$ of latent variables y, and $\bar{\rho}\colon G \times \mathcal{Y} \to [0, \infty)$ is a non-negative density function. In general, no explicit computable expression of the above integral is at hand and the MCwM idea is to substitute $\pi_u(x)$ in the MH algorithm by a Monte Carlo estimate based on iid sequences of random variables $(Y_i^{(x)})_{1\le i\le N}$ and $(Y_i^{(y)})_{1\le i\le N}$ with $Y_1^{(x)} \sim r_x$, $Y_1^{(y)} \sim r_y$. The resulting MCwM algorithm has been studied before in [AR09, MLR16]. This MCwM approach should not be confused with the pseudo-marginal method, see [AR09]. The pseudo-marginal method constructs a Markov chain on the extended space $G \times \mathcal{Y}$ that targets a distribution with π as its marginal on G.

In both intractability settings, the corresponding MCwM Markov chains depend on the parameter $N \in \mathbb{N}$ which denotes the number of samples used within the Monte Carlo estimates. As a consequence, any bound on (1) is

N-dependent, which allows us to control the dissimilarity to the ideal MH based Markov chain. In Corollary 16 and the application of Corollary 17 to the examples considered in Section 4 we provide informative rates of convergence as $N \to \infty$. Note that with those estimates we relax the requirement of uniform bounds on the approximation error introduced by the estimator for π_u, which is essentially imposed in [MLR16, AFEB16]. In contrast to this requirement, we use (if available) the Lyapunov function as a counterweight for a second as well as inverse second moment and can therefore handle situations where uniform bounds on the approximation error are not available. If we do not have access to a Lyapunov function for the MCwM transition kernel we suggest to restrict it to a subset of the state space, i.e., use restricted approximations. This subset is determined by V and usually corresponds to a ball with some radius R that increases as the approximation quality improves, that is, $R(N) \to \infty$ as $N \to \infty$.

Our analysis of the MCwM algorithm is guided by some stylized facts we observe in simple illustrations, in particular, we consider a log-normal example discussed in Section 4.1. In this example, we encounter a situation where the mean squared error of the Monte Carlo approximation grows exponentially in the tail of the target distribution. We observe *empirically* that (unrestricted) MCwM works well whenever the growth behavior is dominated by the decay of the (Gaussian) target density in the tail. The application of Corollary 17 to the log-normal example shows that the restricted approximation converges towards the true target density in the number of samples N at least like $(\log N)^{-1}$ independent of *any* growth of the error. However, the convergence is better, at least like $\frac{\log N}{N}$, if the growth is dominated by the decay of the target density.

Note that our perturbation bounds of Corollary 16 and Corollary 17 hold generally in the doubly-intractable as well as in the latent variable case. Corollary 17 handles the situation of not uniformly bounded approximation error and Corollary 16 the situation of uniformly bounded error. In Section 4.2 we consider the latent variable case and provide an illustrative normal-normal example where the restricted approximation proves to be useful.

2 Preliminaries

Let G be a Polish space, where $\mathcal{B}(G)$ denotes its Borel σ-algebra. Assume that P is a transition kernel with stationary distribution π on G. For a signed measure q on G and a measurable function $f\colon G \to \mathbb{R}$ we define

$$qP(A) := \int_G P(x,A)\,\mathrm{d}q(x), \quad Pf(x) := \int_G f(y)\,P(x,\mathrm{d}y), \quad x \in G, A \in \mathcal{B}(G).$$

For a distribution μ on G we use the notation $\mu(f) := \int_G f(x)\,\mathrm{d}\mu(x)$. For a measurable function $V\colon G \to [1,\infty)$ and two probability measures μ, ν on G define

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

For the constant function $V = 1$ this is the total variation distance, i.e.,

$$\left\|\mu - \nu\right\|_{\mathrm{tv}} := \sup_{|f|\leq 1} \left|\mu(f) - \nu(f)\right|.$$

The next, well-known theorem defines geometric ergodicity and states a useful equivalent condition. The proof follows by [RR97, Proposition 2.1] and [MT09, Theorem 16.0.1].

Theorem 1. For a ϕ-irreducible and aperiodic transition kernel P with stationary distribution π defined on G the following statements are equivalent:

- The transition kernel P is *geometrically ergodic*, that is, there exists a number $\alpha \in [0,1)$ and a measurable function $C\colon G \to [1,\infty)$ such that for π-a.e. $x \in G$ we have
$$\left\|P^n(x,\cdot) - \pi\right\|_{\mathrm{tv}} \leq C(x)\alpha^n, \quad n \in \mathbb{N}. \tag{4}$$
- There is a π-a.e. finite measurable function $V\colon G \to [1,\infty]$ with finite moments with respect to π and there are constants $\alpha \in [0,1)$ and $C \in [1,\infty)$ such that
$$\left\|P^n(x,\cdot) - \pi\right\|_V \leq CV(x)\alpha^n, \quad x \in G,\ n \in \mathbb{N}. \tag{5}$$

In particular, the function V can be chosen such that a *Lyapunov condition* of the form

$$PV(x) \leq \delta V(x) + L, \qquad x \in G, \tag{6}$$

for some $\delta \in [0,1)$ and $L \in (0,\infty)$, is satisfied.

Remark 2. We call a transition kernel *V-uniformly ergodic* if it satisfies (5) and note that this condition can be be rewritten as

$$\sup_{x\in G} \frac{\left\|P^n(x,\cdot) - \pi\right\|_V}{V(x)} \leq C\alpha^n. \tag{7}$$

3 Quantitative perturbation bounds

Assume that $(X_n)_{n\in\mathbb{N}_0}$ is a Markov chain with transition kernel P and initial distribution p_0 on G. We define $p_n := p_0 P^n$, i.e., p_n is the distribution of X_n. The distribution p_n is approximated by using another Markov chain $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ with transition kernel $\widetilde{P}$ and initial distribution $\widetilde{p}_0$. We define $\widetilde{p}_n := \widetilde{p}_0 \widetilde{P}^n$, i.e., $\widetilde{p}_n$ is the distribution of $\widetilde{X}_n$. The idea throughout the paper is to interpret $(X_n)_{n\in\mathbb{N}_0}$ as some ideal, unperturbed chain and $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ as an approximating, perturbed Markov chain.

In the spirit of the doubly-intractable distribution and latent variable case considered in Section 4 we think of the unperturbed Markov chain as "nice", where convergence properties are readily available. Unfortunately since we cannot simulate the "nice" chain we try to approximate it with a perturbed Markov chain, which is, because of the perturbation, difficult to analyze directly. With this in mind, we make the following standing assumption on the unperturbed Markov chain.

Assumption 3. Let $V\colon G \to [1,\infty)$ be a measurable function and assume that P is V-uniformly ergodic, i.e.,

$$\|P^n(x,\cdot) - \pi\|_V \leq C V(x) \alpha^n$$

for some numbers $C \in [1,\infty)$ and $\alpha \in [0,1)$.

We start with an auxiliary estimate of $\|p_n - \widetilde{p}_n\|_{\text{tv}}$ which is interesting on its own and proved in Appendix A.1.

Lemma 4. Let Assumption 3 be satisfied and for a measurable function $W\colon G \to [1,\infty)$ define

$$\varepsilon_{\text{tv},W} := \sup_{x\in G} \frac{\left\|P(x,\cdot) - \widetilde{P}(x,\cdot)\right\|_{\text{tv}}}{W(x)},$$

$$\varepsilon_{V,W} := \sup_{x\in G} \frac{\left\|P(x,\cdot) - \widetilde{P}(x,\cdot)\right\|_{V}}{W(x)}.$$

Then, for any $r \in (0,1]$,

$$\|p_n - \widetilde{p}_n\|_{\text{tv}} \leq C\alpha^n \|p_0 - \widetilde{p}_0\|_V + \varepsilon_{\text{tv},W}^{1-r}\, \varepsilon_{V,W}^{r}\, C^r \sum_{i=0}^{n-1} \widetilde{p}_i(W) \alpha^{(n-i-1)r}. \quad (8)$$

Remark 5. If the initial distribution of the perturbed and unperturbed Markov chain differ, then, this appears also in the estimate. Yet, for n sufficiently large the influence of the difference of the initial distributions vanishes exponentially quickly.

Remark 6. The quantities $\varepsilon_{\mathrm{tv},W}$ and $\varepsilon_{V,W}$ measure the difference between P and $\widetilde{P}$. Note that we can interpret them as operator norms

$$\varepsilon_{\mathrm{tv},W} = \left\| P - \widetilde{P} \right\|_{B^{(1)} \to B^{(W)}} \quad \text{and} \quad \varepsilon_{V,W} = \left\| P - \widetilde{P} \right\|_{B^{(V)} \to B^{(W)}},$$

where

$$B^{(W)} = \left\{ f \colon G \to \mathbb{R} \mid \|f\|_{\infty,W} := \sup_{x \in G} \frac{|f(x)|}{W(x)} < \infty \right\}. \tag{9}$$

It is also easily seen that $\varepsilon_{\mathrm{tv},W} \leq \max\{2, \varepsilon_{V,W}\}$ which implies that $\varepsilon_{V,W}$ is a stronger criterion for measuring the perturbation. In (8) an additional parameter r appears which can be used to tune the estimate. Namely, if one is not able to bound $\varepsilon_{V,W}$ sufficiently well but has a good estimate of $\varepsilon_{\mathrm{tv},W}$ one can optimize over r. On the other hand, if there is a satisfying estimate of $\varepsilon_{V,W}$ one can just set $r = 1$.

In the previous lemma we proved an upper bound of $\|p_n - \widetilde{p}_n\|_{\mathrm{tv}}$ which still contains an unknown quantity given by

$$\sum_{i=0}^{n-1} \widetilde{p}_i(W) \alpha^{(n-i-1)r}$$

which measures, in a sense, stability of the perturbed chain through a weighted sum of expectations of the Lyapunov function W under $\widetilde{p}_i$. To control this term, we impose additional assumptions on the perturbed chain. In the following, we consider two assumptions of this type, a Lyapunov condition and a bounded support assumption.

3.1 Lyapunov condition

We start with a simple version of our main estimate which illustrates already some key aspects of the approach via the Lyapunov condition. Here the intuition is as follows: By Theorem 1 we know that the function V of Assumption 3 can be chosen such that a Lyapunov condition for P is satisfied.

Since we think of $\widetilde{P}$ as being close to P, it might be possible to show also a Lyapunov condition with V of $\widetilde{P}$. If this is the case, the following proposition is applicable.

Proposition 7. Let Assumption 3 be satisfied. Additionally, let $\delta \in [0,1)$ and $L \in (0,\infty)$ such that

$$\widetilde{P}V(x) \leq \delta V(x) + L, \qquad x \in G. \tag{10}$$

Assume that $p_0 = \widetilde{p}_0$ and define $\kappa := \max\left\{\widetilde{p}_0(V), \frac{L}{1-\delta}\right\}$, as well as (for simplicity)

$$\varepsilon_{\mathrm{tv}} := \varepsilon_{\mathrm{tv},V}, \qquad \varepsilon_V := \varepsilon_{V,V}.$$

Then, for any $r \in (0,1]$,

$$\|p_n - \widetilde{p}_n\|_{\mathrm{tv}} \leq \varepsilon_{\mathrm{tv}}^{1-r} \varepsilon_V^r \, \frac{C^r \kappa}{(1-\alpha)r}. \tag{11}$$

Proof. We use Lemma 4 with $W = V$. By (10) follows that

$$\widetilde{p}_i(V) = \int_G \widetilde{P}^i V(x)\, \widetilde{p}_0(\mathrm{d}x) \leq \delta^i \widetilde{p}_0(V) + (1-\delta^i)\frac{L}{1-\delta} \leq \kappa. \tag{12}$$

The final estimate is obtained by a geometric series and $1-\alpha^r \geq r(1-\alpha)$. □

Now we state a more general theorem. In particular, in this estimate the dependence on the initial distribution can be weakened. In the perturbation bound of the previous estimate, the initial distribution is only forgotten if $\widetilde{p}_0(V) < L/(1-\delta)$. Yet, intuitively, for long-term stability results $\widetilde{p}_0(V)$ should not matter at all. This intuition is confirmed by the theorem.

Theorem 8. Let Assumption 3 be satisfied. Assume also that $W\colon G \to [1,\infty)$ is a measurable function which satisfies with $\delta \in [0,1)$ and $L \in (0,\infty)$ the Lyapunov condition

$$\widetilde{P}W(x) \leq \delta W(x) + L, \qquad x \in G. \tag{13}$$

Define $\varepsilon_{\mathrm{tv},W}$, $\varepsilon_{V,W}$ as in Lemma 4 and $\gamma := \frac{L}{1-\delta}$. Then, for any $r \in (0,1]$ with

$$\beta_{n,r}(\delta,\alpha) := \begin{cases} n\alpha^{(n-1)r}, & \alpha^r = \delta, \\ \frac{|\alpha^{rn}-\delta^n|}{|\alpha^r-\delta|}, & \alpha^r \neq \delta, \end{cases}$$

we have

$$\|\widetilde{p}_n - p_n\|_{\mathrm{tv}} \le C\alpha^n \|\widetilde{p}_0 - p_0\|_V + \varepsilon_{\mathrm{tv},W}^{1-r}\ \varepsilon_{V,W}^r\ C^r \left[\widetilde{p}_0(W)\, \beta_{n,r}(\delta,\alpha) + \frac{\gamma}{(1-\alpha)r}\right]. \tag{14}$$

Proof. Here we use Lemma 4 with possibly different W and V. By (13) we have $\widetilde{p}_i(W) \le \delta^i \widetilde{p}_0(W) + \gamma$ and by

$$\sum_{i=0}^{n-1} \delta^i \alpha^{(n-i-1)r} = \beta_{n,r}(\delta,\alpha)$$

we obtain the assertion by a geometric series and $1-\alpha^r \ge r(1-\alpha)$. □

Remark 9. We consider an illustrating example where Theorem 8 leads to a considerably sharper bound than Proposition 7. This improvement is due to the combination of two novel properties of the bound of Theorem 8:

1. In the Lyapunov condition (13) the function W can be chosen differently from V.

2. The fact that $\beta_{n,r}(\delta,\alpha)$ is bounded from above by $\max\{\delta,\alpha^r\}^{n-1}$. Thus $\beta_{n,r}(\delta,\alpha)$ converges almost exponentially fast to zero in n. This implies that for n sufficiently large the dependence of $\widetilde{p}_0(W)$ vanishes. Nevertheless, the leading factor n can capture situations in which the perturbation error is increasing in n for small n.

Illustrating example. Let $G = \{0,1\}$ and assume $p_0 = \widetilde{p}_0 = (0,1)$. Here state “1” can be interpreted as the “tail” while state “0” is the “center”. Define

$$P = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \widetilde{P} = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Thus, the unperturbed Markov chain $(X_n)_{n\in\mathbb{N}_0}$ moves from “1” to “0” right away, while the perturbed one $(\widetilde{X}_n)_{n\in\mathbb{N}_0}$ takes longer. Both transition matrices have the same stationary distribution $\pi = (1,0)$. Obviously, $\|p_0 - \widetilde{p}_0\|_{\mathrm{tv}} = 0$ and for $n \in \mathbb{N}$ holds

$$\|p_n - \widetilde{p}_n\|_{\mathrm{tv}} = 2\mathbb{P}(X_n \ne \widetilde{X}_n) = \frac{1}{2^{n-1}}.$$

The unperturbed Markov chain is uniformly ergodic, such that we can choose $V = 1$ and (5) is satisfied with $C = 1$ and $\alpha = 0$. In particular, in this setting

ε_{tv} and ε_V from Proposition 7 coincide, we have $\varepsilon_{\text{tv}} = 1$. Thus, the estimate of Proposition 7 gives

$$\|p_n - \widetilde{p}_n\|_{\text{tv}} \leq \varepsilon_{\text{tv}} = 1.$$

This bound is optimal in the sense that it is best possible for $n = 1$. But for increasing n it is getting worse. Notice also that a different choice of V cannot really remedy this situation: The chains differ most strongly at $n = 1$ and the bound of Proposition 7 is constant over time. Now choose the function $W(x) = 1 + v \cdot \mathbf{1}_{\{x=1\}}$ for some $v \geq 0$. The transition matrix $\widetilde{P}$ satisfies the Lyapunov condition

$$\widetilde{P}W(x) \leq \frac{1}{2} W(x) + \frac{1}{2},$$

i.e., $\delta = L = \frac{1}{2}$. Moreover, we have $\widetilde{p}_0(W) = 1 + v$ and $\varepsilon_{V,W} = \varepsilon_{\text{tv},W} = 1/(1+v)$. Thus, in the bound from Theorem 8 we can set $r = 1$ and $\gamma = 1$ such that

$$\|p_n - \widetilde{p}_n\|_{\text{tv}} \leq \frac{1}{v+1} + \frac{1}{2^{n-1}}.$$

Since v can be chosen arbitrarily large, it follows that

$$\|p_n - \widetilde{p}_n\|_{\text{tv}} \leq \frac{1}{2^{n-1}},$$

which is best possible for all $n \in \mathbb{N}$.

The previous example can be seen as a toy model of a situation where the transition probabilities of a perturbed and unperturbed Markov chain are very similar in the "center" of the state space but differ considerably in the "tail". When the chains start both in the same point in the "tail", considerable differences between distributions can build up along the initial transient and then vanish again. Earlier perturbation bounds as for example in [Mit05, PS14, RS18] take only an initial error and a remaining error into account. Thus, those are worse for situations where this transient error captured by $\beta_{n,r}$ dominates. A very similar term also appears in the very recent error bounds due to [JM17b]. Here we also see that a function W different from V is advantageous.

3.2 Restricted approximation

In the previous section, we have seen that a Lyapunov condition of the perturbation helps to control the long-term stability of approximating a V-

uniformly ergodic Markov chain. In this section we assume that the perturbed chain is restricted to a "large" subset of the state space. In this setting a sufficiently good approximation of the unperturbed Markov chain on this subset leads to a perturbation estimate.

For the unperturbed Markov chain we assume that transition kernel P is V-uniformly ergodic. Then, for $R \geq 1$ define the "large subset" of the state space as

$$B_R = \{x \in G \mid V(x) \leq R\}.$$

If V is chosen as a monotonic transformation of a norm on G, B_R is simply a ball around 0. The *restriction of* P to the set B_R, given as P_R, is defined as

$$P_R(x, A) = P(x, A \cap B_R) + \mathbf{1}_A(x) P(x, B_R^c), \quad A \in \mathcal{B}(G),\, x \in G.$$

In other words, whenever P would make a transition from $x \in B_R$ to $G \setminus B_R$, P_R remains in x. Otherwise, P_R is the same as P. We obtain the following perturbation bound for approximations whose stability is guaranteed through a restriction to the set B_R.

Theorem 10. Under the V-uniform ergodicity of Assumption 3 let $\delta \in [0, 1)$ and $L \in [1, \infty)$ be chosen in such a way that

$$PV(x) \leq \delta\, V(x) + L, \quad x \in G.$$

For the perturbed transition kernel $\widetilde{P}$ assume that it is restricted to B_R, i.e., $\widetilde{P}(x, B_R) = 1$ for all $x \in G$, and that $R \cdot \Delta(R) \leq (1 - \delta)/2$ with

$$\Delta(R) := \sup_{x \in B_R} \frac{\left\| P_R(x, \cdot) - \widetilde{P}(x, \cdot) \right\|_{\mathrm{tv}}}{V(x)}.$$

Then, with $p_0 = \widetilde{p}_0$ and

$$\kappa := \max\left\{ \widetilde{p}_0(V), \frac{L}{1 - \delta} \right\}$$

we have for $R \geq \exp(1)$ that

$$\|p_n - \widetilde{p}_n\|_{\mathrm{tv}} \leq \frac{33C(L+1)\kappa}{1 - \alpha} \cdot \frac{\log R}{R}. \tag{15}$$

The proof of the result is stated in Appendix A.1. Notice that while the perturbed chain is restricted to the set B_R, we do not place a similar restriction on the unperturbed chain. The estimate (15) compares the restricted, perturbed chain to the unrestricted, unperturbed one.

Remark 11. In the special case where $\widetilde{P}(x,\cdot) = P_R(x,\cdot)$ for $x \in B_R$ we have $\Delta(R) = 0$. For example

$$\widetilde{P}(x,A) = \mathbf{1}_{B_R}(x) P_R(x,A) + \mathbf{1}_{B_R^c}(x)\delta_{x_0}(A), \quad A \in \mathcal{B}(G),$$

with $x_0 \in B_R$ satisfies this condition. The resulting perturbed Markov chain is simply a restriction of the unperturbed Markov chain to B_R and Theorem 10 provides a quantitative bound on the difference of the distributions.

4 Monte Carlo within Metropolis

In Bayesian statistics it is of interest to sample with respect to a distribution π on $(G, \mathcal{B}(G))$. We assume that π admits a possibly *unnormalized density* $\pi_u \colon G \to [0,\infty)$ with respect to a reference measure μ, for example the counting, Lebesgue or some Gaussian measure. The Metropolis-Hastings (MH) algorithm is often the method of choice to draw approximate samples according to π:

Algorithm 1. For a *proposal transition kernel* Q a transition from x to y of the MH algorithm works as follows.

1. Draw $U \sim \mathrm{Unif}[0,1]$ and a proposal $Z \sim Q(x,\cdot)$ independently, call the result u and z, respectively.

2. Compute
$$a(x,z) := \min\left\{1, r(x,z)\right\},$$
where
$$r(x,z) := \frac{\pi(\mathrm{d}z)Q(z,\mathrm{d}x)}{\pi(\mathrm{d}x)Q(x,\mathrm{d}z)} = \frac{\pi_u(z)}{\pi_u(x)} \frac{\mu(\mathrm{d}z)Q(z,\mathrm{d}x)}{\mu(\mathrm{d}x)Q(x,\mathrm{d}z)} \tag{16}$$
is the *acceptance ratio*, that is, the density of the measure $\pi(\mathrm{d}z)Q(z,\mathrm{d}x)$ w.r.t. $\pi(\mathrm{d}x)Q(x,\mathrm{d}z)$, see [Tie98].

3. If $u < a(x,z)$, then accept the proposal, and return $y := z$, otherwise reject the proposal and return $y := x$.

The transition kernel of the MH algorithm with proposal Q, stationary distribution π and acceptance probability $a(x,z)$ is given by

$$M_a(x,\mathrm{d}z) := a(x,z)Q(x,\mathrm{d}z) + \delta_x(\mathrm{d}z)\left(1 - \int_G a(x,y)Q(x,\mathrm{d}y)\right). \qquad (17)$$

For the MH algorithm in the computation of $r(x,z)$ one uses $\pi_u(z)/\pi_u(x)$, which might be known from having access to function evaluations of the unnormalized density π_u. However, when it is expensive or even impossible to compute function values of π_u, then it may not be feasible to sample from π using the MH algorithm. Here are two typical examples of such scenarios:

- **Doubly-intractable distribution:** For models such as *Markov or Gibbs random fields*, the unnormalized density $\pi_u(x)$ itself is typically only known up to a factor $Z(x)$, that is,

  $$\pi_u(x) = \rho(x)/Z(x), \qquad x \in G \qquad (18)$$

  where functions values of ρ can be computed, but function values of Z cannot. For instance, Z might be given in the form

  $$Z(x) = \int_{\mathcal{Y}} \bar{\rho}(x,y)\, r_x(\mathrm{d}y),$$

  where $\mathcal{Y}$ denotes an auxiliary variable space, $\bar{\rho}\colon G \times \mathcal{Y} \to [0,\infty)$ and r_x is a σ-finite measure.

- **Latent variables:** Here $\pi_u(x)$ cannot be evaluated, since it takes the form

  $$\pi_u(x) = \int_{\mathcal{Y}} \bar{\rho}(x,y)\, r_x(\mathrm{d}y) \qquad (19)$$

  with a distribution r_x on a measurable space $\mathcal{Y}$ of *latent variables* y and a non-negative function $\bar{\rho}\colon G \times \mathcal{Y} \to [0,\infty)$.

In the next sections, we study in both of these settings the perturbation error of an approximating MH algorithm. A fair assumption in both scenarios is that the infeasible, unperturbed MH algorithm is V-uniformly ergodic:

Assumption 12. For some function $V\colon G \to [1,\infty)$ let the transition kernel M_a of the MH algorithm be V-uniformly ergodic, that is,

$$\|M_a^n(x,\cdot) - \pi\|_V \le CV(x)\alpha^n$$

with $C \in [1,\infty)$ and $\alpha \in [0,1)$, and additionally, assume that the Lyapunov condition

$$M_a V(x) \leq \delta V(x) + L,$$

for some $\delta \in [0,1)$ and $L \in [1,\infty)$ is satisfied.

We have the following standard proposition (see e.g. [RS18, Lemma 4.1] or [AFEB16, BDH14, JM17b, MLR18, PS14]) which leads to upper bounds on ε_{tv}, ε_V and $\Delta(R)$ (see Lemma 4 and Theorem 10) for two MH type algorithms M_b and M_c with acceptance probability functions $b, c\colon G \times G \to [0,1]$.

Proposition 13. Let $b, c\colon G \times G \to [0,1]$ and let $V\colon G \to [1,\infty)$ be such that $\sup\limits_{x\in G} \frac{M_b V(x)}{V(x)} \leq T$ for a constant $T \geq 1$. Assume that there are functions $\eta, \xi\colon G \to [0,\infty)$ and a set $B \subseteq G$ such that either for $z = x$ or $z = y$ holds

$$|b(x,y) - c(x,y)| \leq \mathbf{1}_B(y)\xi(z)(\eta(x)+\eta(y))b(x,y), \quad \forall x,y \in G. \tag{20}$$

Then we have

$$\sup_{x\in B} \frac{\|M_b(x,\cdot) - M_c(x,\cdot)\|_V}{V(x)} \leq 4T\,\|\eta \cdot \mathbf{1}_B\|_\infty \,\|\xi \cdot \mathbf{1}_B\|_\infty\,,$$

and, for any $\beta \in (0,1)$,

$$\sup_{x\in B} \frac{\|M_b(x,\cdot) - M_c(x,\cdot)\|_{\text{tv}}}{V(x)} \leq 4T\,\|\eta \cdot \mathbf{1}_B\|_{\infty,V^\beta} \,\|\xi \cdot \mathbf{1}_B\|_{\infty,V^{1-\beta}}\,.$$

The proposition provides a tool for controlling the distance between the transition kernels of two MH type algorithms with identical proposal and different acceptance probabilities. The specific functional form for the dependence of the upper bound in (20) on x and y is motivated by the applications below. The set B indicates the "essential" part of G where the difference of the acceptance probabilities matter. The parameter β is used to shift weight between the two components ξ and η of the approximation error. For the proof of the proposition, we refer to Appendix A.2.

4.1 Doubly-intractable distributions

In the case where π_u takes the form (18), we can approximate $Z(x)$ by a Monte Carlo estimate

$$\widehat{Z}_N(x) := \frac{1}{N}\sum_{i=1}^{N} \bar{\rho}(x, Y_i^{(x)}),$$

under the assumption that we have access to an iid sequence of random variables $(Y_i^{(x)})_{1\le i\le N}$ where each $Y_i^{(x)}$ is distributed according to r_x. Then, the idea is to substitute the unknown quantity $Z(x)$ by the approximation $\widehat{Z}_N(x)$ within the acceptance ratio. Define a function $W_N \colon G \to \mathbb{R}$ by $W_N(x) := \frac{\widehat{Z}_N(x)}{Z(x)}$, then the acceptance probability given $W_N(x)$, $W_N(z)$ modifies to

$$\widetilde{a}(x, z, W_N) := \min\left\{1, r(x, z) \cdot \frac{W_N(x)}{W_N(z)}\right\}, \tag{21}$$

where the random variables $W_N(x)$, $W_N(z)$ are assumed to be independent from each other. This leads to a *Monte Carlo within Metropolis* (MCwM) algorithm:

Algorithm 2. For a given proposal transition kernel Q, a transition from x to y of the MCwM algorithm works as follows.

1. Draw $U \sim \mathrm{Unif}[0, 1]$ and a proposal $Z \sim Q(x, \cdot)$ independently, call the result u and z, respectively.

2. Based on independent samples compute $W_N(x)$, $W_N(z)$ and then calculate $\widetilde{a}(x, z, W_N)$.

3. If $u < \widetilde{a}(x, z, W_N)$, then accept the proposal, and return $y := z$, otherwise reject the proposal and return $y := x$.

Given the current state $x \in G$ and a proposed state $z \in G$ the overall acceptance probability is

$$a_N(x, z) := \mathbb{E}[\widetilde{a}(x, z, W_N)],$$

which leads to the corresponding transition kernel of the form M_{a_N}, see (17).

The quality of the MCwM algorithm depends on the error of the approximation of $Z(x)$. The root mean squared error of this approximation can be quantified by the use of W_N, that is,

$$(\mathbb{E}\,|W_N(x) - 1|^2)^{1/2} = \frac{s(x)}{\sqrt{N}} \qquad x \in G,\ N \in \mathbb{N}, \tag{22}$$

where

$$s(x) := (\mathbb{E}\,|W_1(x) - 1|^2)^{1/2}$$

is determined by the second moment of $W_1(x)$. In addition, due to the appearance of the estimator $W_N(z)$ in the denominator of $\widetilde{a}$, we need some control of its distribution near zero. To this end, we define, for $z \in G$ and $p > 0$, the inverse moment function

$$i_{p,N}(z) := \left(\mathbb{E} W_N(z)^{-p}\right)^{\frac{1}{p}}.$$

With this notation we obtain the following estimate, which is proved in Appendix A.2.

Lemma 14. Assume that there exists $k \in \mathbb{N}$ such that $i_{2,k}(x)$ and $s(x)$ are finite for all $x \in G$. Then, for all $x, z \in G$ and $N \geq k$ we have

$$|a(x,z) - a_N(x,z)| \leq a(x,z) \frac{1}{\sqrt{N}} i_{2,k}(z)(s(x) + s(z)).$$

Remark 15. One can replace the boundedness of the second inverse moment $i_{2,k}(x)$ for any $x \in G$ by boundedness of a lower moment $i_{p,m}(x)$ for $p \in (0,2)$ with suitably adjusted $m \in \mathbb{N}$, see Lemma 23 in the appendix.

4.1.1 Inheritance of the Lyapunov condition

If the second and inverse second moment are uniformly bounded, $\|s\|_\infty < \infty$ as well as $\|i_{2,N}\|_\infty < \infty$, one can show that the Lyapunov condition of the MH transition kernel is inherited by the the MCwM algorithm. In the following corollary, we prove this inheritance and state the resulting error bound for MCwM.

Corollary 16. For a distribution m_0 on G let $m_n := m_0 M_a^n$ and $m_{n,N} := m_0 M_{a_N}^n$ be the respective distributions of the MH and MCwM algorithms after n steps. Let Assumption 12 be satisfied and for some $k \in \mathbb{N}$ let

$$D := 8L \|i_{2,k}\|_\infty \|s\|_\infty < \infty.$$

Further, define $\delta_N := \delta + D/\sqrt{N}$ and $\beta_n := n \max\{\delta_N, \alpha\}^{n-1}$. Then, for any

$$N > \max\left\{k, \frac{D^2}{(1-\delta)^2}\right\}$$

we have $\delta_N \in [0,1)$ and

$$\|m_n - m_{n,N}\|_{\text{tv}} \leq \frac{DC}{\sqrt{N}} \left[m_0(V)\beta_n + \frac{L}{(1-\delta_N)(1-\alpha)} \right].$$

Proof. Assumption 12 implies $\sup_{x\in G} \frac{M_a V(x)}{V(x)} \leq 2L$. By Lemma 14 and Proposition 13, with $B = G$, we obtain

$$\varepsilon_{V,V} = \sup_{x\in G} \frac{\|M_a(x,\cdot) - M_{a_N}(x,\cdot)\|_V}{V(x)} \leq \frac{D}{\sqrt{N}}.$$

Further, note that

$$M_{a_N}V(x) - M_a V(x) \leq \|M_a(x,\cdot) - M_{a_N}(x,\cdot)\|_V \leq \frac{D}{\sqrt{N}} V(x),$$

which implies, by Assumption 12, that for $N > D^2/(1-\delta)^2$ we have $\delta_N \in [0,1)$ and $M_{a_N}V(x) \leq \delta_N V(x) + L$. By Theorem 8 and Remark 9 we obtain for $r = 1$ the assertion. □

Observe that the estimate is bounded in $n \in \mathbb{N}$ so that the difference of the distributions converges uniformly in n to zero for $N \to \infty$. The constant δ_N decreases for increasing N, so that larger values of N improve the bound.

Log-normal example I. Let $G = \mathbb{R}$ and the target measure π be the standard normal distribution. We choose a Gaussian proposal kernel $Q(x,\cdot) = \mathcal{N}(x,\gamma^2)$ for some $\gamma^2 > 0$, where $\mathcal{N}(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. It is well known, see [JH00, Theorem 4.1, Theorem 4.3 and Theorem 4.6], that the MH transition kernel satisfies Assumption 12 for some numbers α, C, δ and L with $V(x) = \exp(x^2/4)$.

Let $g(y;\mu,\sigma^2)$ be the density of the log-normal distribution with parameters μ and σ, i.e., g is the density of $\exp(\mu + \sigma S)$ for a random variable $S \sim \mathcal{N}(0,1)$. Then, by the fact that $\int_0^\infty y\, g(y; -\sigma(x)^2/2, \sigma(x)^2)\mathrm{d}y = 1$ for all functions $\sigma\colon G \to (0,\infty)$, we can write the (unnormalized) standard normal density as

$$\pi_u(x) = \exp(-x^2/2) = \frac{\exp(-x^2/2)}{\int_0^\infty y\, g(y; -\sigma(x)^2/2, \sigma(x)^2)\mathrm{d}y}.$$

Hence π_u takes the form (18) with $\mathcal{Y} = [0,\infty)$, $\rho(x) = \exp(-x^2/2)$, $\bar{\rho}(x,y) = y$ and r_x being a log-normal distribution with parameters $-\sigma(x)^2/2$ and $\sigma(x)^2$. Independent draws from this log-normal distribution are used in the MCwM algorithm to approximate the integral. We have $\mathbb{E}[W_1(x)^p] = \exp(p(p-1)\sigma(x)^2/2)$ for all $x, p \in \mathbb{R}$ and, accordingly,

$$\begin{aligned} s(x) &= (\exp(\sigma(x)^2) - 1)^{1/2} \leq \exp(\sigma(x)^2/2) \\ i_{p,1}(x) &= \exp((p+1)\sigma(x)^2/2). \end{aligned}$$

By Lemma 23 we conclude that

$$i_{2,k}(x) \leq i_{2/k,1}(x) = \exp\left(\left(\frac{1}{2} + \frac{1}{k}\right)\sigma(x)^2\right).$$

Hence, $\|s\|_\infty$ as well as $\|i_{2,k}\|_\infty$ are bounded if for some constant $c > 0$ we have $\sigma(x)^2 \leq c$ for all $x \in G$. In that case Corollary 16 is applicable and provides estimates for the difference between the distributions of the MH and MCwM algorithms after n-steps. However, one might ask what happens if the function $\sigma(x)^2$ is not uniformly bounded, taking, for example, the form $\sigma(x)^2 = |x|^q$ for some $q > 0$. In Figure 1 we illustrate the difference of the distribution of the target measure to a kernel density estimator based on a MCwM algorithm sample for $\sigma(x)^2 = |x|^{1.8}$. Even though $s(x)$ and $i_{p,1}(x)$ grows super-exponentially in $|x|$, the MCwM still works reasonably well in this case.

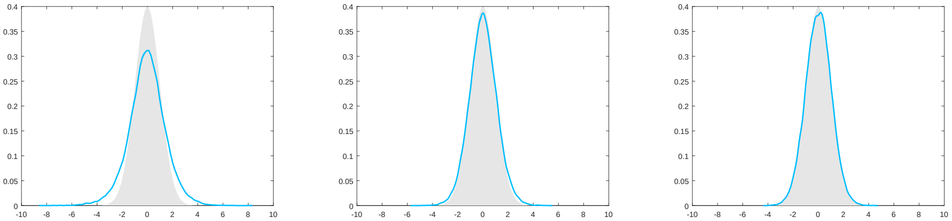

Figure 1: Here $\sigma(x)^2 := |x|^{1.8}$ for $x \in \mathbb{R}$. The target density (standard normal) is plotted in grey, a kernel density estimator based on 10^5 steps of the MCwM algorithm with $N = 10$ (left), $N = 10^2$ (middle) and $N = 10^3$ (right) is plotted in blue.

However, in Figure 2 we consider the case where $\sigma(x)^2 = |x|^{2.2}$ and the behavior changes dramatically. Here the MCwM algorithm does not seem to work at all. This motivates a modification of the MCwM algorithm in terms of restricting the state space to the "essential part" determined by the Lyapunov condition.

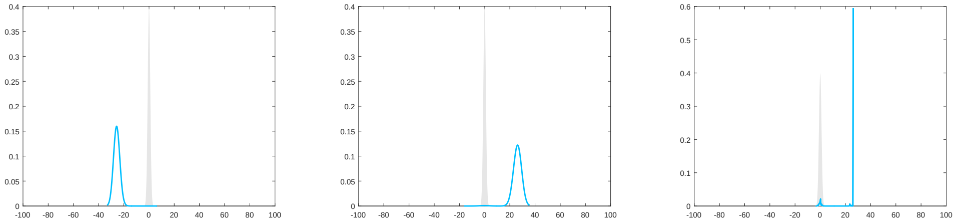

Figure 2: Here $\sigma(x)^2 := |x|^{2.2}$ for $x \in \mathbb{R}$. The target density (standard normal) is plotted in grey, a kernel density estimator based on 10^5 steps of the MCwM algorithm with $N = 10$ (left), $N = 10^2$ (middle) and $N = 10^3$ (right) is plotted in blue.

4.1.2 Restricted MCwM approximation

With the notation and definition from the previous section we consider the case where the functions $i_{2,k}(x)$ and $s(x)$ are not uniformly bounded. Under Assumption 12 there are two simultaneously used tools which help to control the difference of a transition of MH and MCwM:

1. The Lyapunov condition leads to a weight function and eventually to a weighted norm, see Proposition 13.

2. By restricting the MCwM to the "essential part" of the state space we prevent that the approximating Markov chain deteriorates. Namely, for some $R \geq 1$ we restrict the MCwM to B_R, see Section 3.2.

For $x, z \in G$ the acceptance probability given $W_N(x)$, $W_N(z)$ is now modified from $\widetilde{a}(x, z, W_N)$ to

$$\mathbf{1}_{B_R}(z) \cdot \widetilde{a}(x, z, W_N)$$

which leads to the *restricted MCwM algorithm*:

Algorithm 3. For given $R \geq 1$ and a proposal transition kernel Q a transition from x to y of the restricted MCwM algorithm works as follows.

1. Draw $U \sim \mathrm{Unif}[0, 1]$ and a proposal $Z \sim Q(x, \cdot)$ independently, call the result u and z, respectively.

2. Based on independent samples compute $W_N(x)$, $W_N(z)$ and then calculate $\widetilde{a}(x, z, W_N)$.

3. If $u < \mathbf{1}_{B_R}(z)\cdot\widetilde{a}(x,z,W_N)$, then accept the proposal, and return $y := z$, otherwise reject the proposal and return $y := x$.

Given the current state $x \in G$ and a proposed state $z \in G$ the overall acceptance probability is

$$a_N^{(R)}(x,z) := \mathbf{1}_{B_R}(z)\cdot\mathbb{E}[\widetilde{a}(x,z,W_N)] = \mathbf{1}_{B_R}(z)\cdot a_N(x,z),$$

which leads to the corresponding transition kernel of the form $M_{a_N^{(R)}}$, see (17). By using Theorem 10 and Proposition 13 we obtain the following estimate.

Corollary 17. Let Assumption 12 be satisfied, i.e., M_a is V-uniformly ergodic and the function V as well as the constants α, C, δ and L are determined. For $\beta \in (0,1)$ and $R \geq 1$ let

$$\begin{aligned}
&B_R := \{x \in G \mid V(x) \leq R\}, \\
&D_R := 12\cdot L \left\| i_{2,k}\cdot\mathbf{1}_{B_R}\right\|_{\infty,V^{1-\beta}} \left\| s\cdot\mathbf{1}_{B_R}\right\|_{\infty,V^{\beta}} < \infty.
\end{aligned}$$

Let m_0 be a distribution on B_R and $\kappa := \max\{m_0(V), L/(1-\delta)\}$. Then, for

$$N \geq \max\left\{k, 4\left(\frac{R\cdot D_R}{1-\delta}\right)^2\right\} \tag{23}$$

and $R \geq \exp(1)$ we have

$$\left\| m_n - m_{n,N}^{(R)} \right\|_{\mathrm{tv}} \leq \frac{33C(L+1)\kappa}{1-\alpha}\cdot\frac{\log R}{R},$$

where $m_{n,N}^{(R)} := m_0 M^n_{a_N^{(R)}}$ and $m_n := m_0 M_a^n$ are the distributions of the MH and restricted MCwM algorithm after n-steps.

Proof. We apply Theorem 10 with $P(x,\cdot) = M_a(x,\cdot)$ and

$$\widetilde{P}(x,\cdot) = \mathbf{1}_{B_R}(x)\, M_{a_N^{(R)}}(x,\cdot) + \mathbf{1}_{B_R^c}(x)\delta_{x_0}(\cdot), \quad x \in G,$$

for some $x_0 \in B_R$. Note that $\widetilde{P}(x, B_R) = 1$ for any $x \in G$. Further $\widetilde{P}$ and $M_{a_N^{(R)}}$ coincide on B_R, thus we also have $\widetilde{P}^n = M^n_{a_N^{(R)}}$ on B_R for $n \in \mathbb{N}$.

Observe also that the restriction of P to B_R, denoted by P_R, satisfies $P_R = M_{a^{(R)}}$ with $a^{(R)}(x,z) := \mathbf{1}_{B_R}(z)\, a(x,z)$. Hence

$$\Delta(R) = \sup_{x \in B_R} \frac{\left\| M_{a^{(R)}}(x,\cdot) - M_{a_N^{(R)}}(x,\cdot) \right\|_{\text{tv}}}{V(x)}.$$

Moreover, we have by Lemma 14 that

$$\begin{aligned} \left| a^{(R)}(x,z) - a_N^{(R)}(x,z) \right| =& \mathbf{1}_{B_R}(z)\, |a(x,z) - a_N(x,z)| \\ \leq& \mathbf{1}_{B_R}(z) \cdot a(x,z) \frac{1}{\sqrt{N}}\, i_{2,k}(z)(s(x)+s(z)) \\ =& a^{(R)}(x,z) \frac{1}{\sqrt{N}}\, i_{2,k}(z)(s(x)+s(z)). \end{aligned}$$

With Proposition 13 and

$$\sup_{x \in G} \frac{M_{a^{(R)}} V(x)}{V(x)} \leq \sup_{x \in G} \frac{M_a V(x)}{V(x)} + 1 \underset{\text{Ass. 12}}{\leq} 3L,$$

we have that $\Delta(R) \leq D_R/\sqrt{N}$. Then, by $N \geq 4(RD_R/(1-\delta))^2$ we obtain

$$R \cdot \Delta(R) \leq \frac{1-\delta}{2}$$

such that all conditions of Theorem 10 are verified and the stated estimate follows. □

Remark 18. The estimate depends crucially on the sample size N as well as on the parameter R. If the influence of R in D_R is explicitly known, then one can choose R depending on N in such away that the conditions of the corollary are satisfied and one eventually obtains an upper bound on the total variation distance of the difference between the distributions depending only on N and not on R anymore.

Log-normal example II. We continue with the log-normal example. In this setting we have

$$\begin{aligned} B_R &= \{x \in \mathbb{R} \mid |x| \leq 2\sqrt{\log R}\}, \\ \|i_{2,k} \cdot \mathbf{1}_{B_R}\|_{\infty, V^{1-\beta}} &\leq \sup_{|x| \leq 2\sqrt{\log R}} \exp\left(\left(\frac{1}{2} + \frac{1}{k} \right) \sigma(x)^2 - \frac{1-\beta}{4} x^2 \right), \\ \|s \cdot \mathbf{1}_{B_R}\|_{\infty, V^{\beta}} &\leq \sup_{|x| \leq 2\sqrt{\log R}} \exp\left(\sigma(x)^2/2 - \beta x^2/4 \right). \end{aligned}$$

As in the numerical experiments in Figure 1 and Figure 2 let us consider the cases $\sigma(x)^2 = |x|^{1.8}$ and $\sigma(x)^2 = |x|^{2.2}$. In Figure 3 we compare the normal target density with a kernel density estimator based on the restricted MCwM on $B_R = [-10, 10]$ and observe essentially the same reasonable behavior as in Figure 1. In Figure 4 we consider the same scenario and observe that the restriction indeed stabilizes. In contrast to Figure 2, convergence to the true target distribution is visible but, in line with the theory, slower than for $\sigma(x)^2 = |x|^{1.8}$.

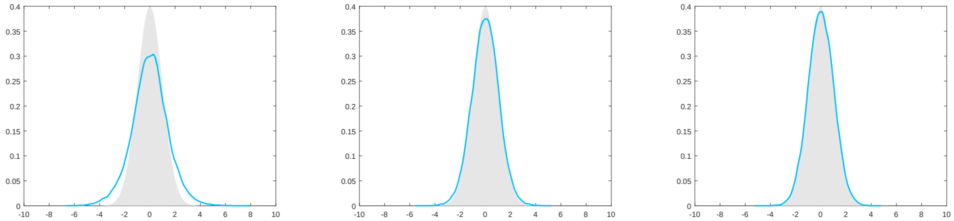

Figure 3: Here $\sigma(x)^2 := |x|^{1.8}$ for $x \in \mathbb{R}$ and $B_R = [-10, 10]$. The target density (standard normal) is plotted in grey, a kernel density estimator based on 10^5 steps of the MCwM algorithm with $N = 10$ (left), $N = 10^2$ (middle) and $N = 10^3$ (right) is plotted in blue.

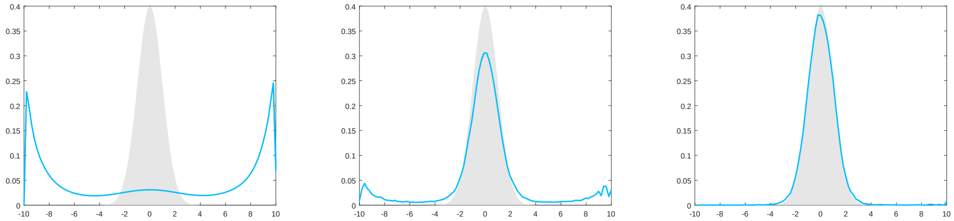

Figure 4: Here $\sigma(x)^2 := |x|^{2.2}$ for $x \in \mathbb{R}$ and $B_R = [-10, 10]$. The target density (standard normal) is plotted in grey, a kernel density estimator based on 10^5 steps of the MCwM algorithm with $N = 10$ (left), $N = 10^2$ (middle) and $N = 10^3$ (right) is plotted in blue.

Now we apply Corollary 17 in both cases:

1. Case $\sigma(x)^2 = |x|^{1.8}$. For $k = 100$ and $\beta = 1/2$ one can easily see that $\|i_{2,100} \cdot \mathbf{1}_{B_R}\|_{\infty,V^{1/2}}$ and $\|s \cdot \mathbf{1}_{B_R}\|_{\infty,V^{1/2}}$ is bounded by 6000, independent of R. Hence there is a constant $D \geq 1$ so that $D_R \leq D$. With this knowledge we choose $R = \frac{(1-\delta)}{\sqrt{2}D}\sqrt{N}$ such that for $N \geq \max\left\{100, \frac{2\exp(2)D^2}{(1-\delta)^2}\right\}$ condition (23) and $R \geq \exp(1)$ is satisfied. Then, Corollary 17 gives the existence of a constant $\widetilde{C} > 0$, so that

$$\left\| m_n - m_{n,N}^{(R)} \right\|_{\mathrm{tv}} \leq \widetilde{C}\, \frac{\log N}{\sqrt{N}}$$

for any initial distribution m_0 on B_R.

2. Case $\sigma(x)^2 = |x|^{2.2}$. For $k = 100$ and $\beta = 1/2$ we obtain

$$\begin{aligned} \|i_{2,100} \cdot \mathbf{1}_{B_R}\|_{\infty,V^{1/2}} &\leq \exp\left(2.5\,(\log R)^{11/10}\right), \\ \|s \cdot \mathbf{1}_{B_R}\|_{\infty,V^{1/2}} &\leq \exp\left(2.5\,(\log R)^{11/10}\right). \end{aligned}$$

Hence $D_R \leq 12L \exp\left(5\,(\log R)^{11/10}\right)$. Eventually, for

$$N \geq \max\left\{100, \frac{24^2 \exp(2 \cdot 6^{11/10}) L^2}{(1-\delta)^2}\right\}$$

we have with $R = \exp\left(\frac{1}{6}\left[\log\left(\frac{\sqrt{N}(1-\delta)}{24L}\right)\right]^{10/11}\right)$ that $R \geq \exp(1)$ and (23) is satisfied. Then, with $\widetilde{C}_1 := \frac{33C(L+1)\kappa}{1-\alpha}$, $\widetilde{C}_2 := \sqrt{\frac{1-\delta}{24L}}$ and Corollary 17 we have

$$\left\| m_n - m_{n,N}^{(R)} \right\|_{\mathrm{tv}} \leq \frac{\widetilde{C}_1 \cdot \frac{1}{6 \cdot 2^{10/11}} \left[\log\left(\widetilde{C}_2 N\right)\right]^{10/11}}{\exp\left(\frac{1}{6 \cdot 2^{10/11}} \left[\log\left(\widetilde{C}_2 N\right)\right]^{10/11}\right)} \leq \frac{\widetilde{C}_1 (q+1)!}{\left[\log\left(\widetilde{C}_2 N\right)\right]^{10q/11}},$$

for any initial distribution m_0 on B_R and all $q \in \mathbb{N}$. Here the last inequality follows by the fact that $\exp(x) \geq \frac{x^{q+1}}{(q+1)!}$ for any $x \geq 0$ and $q \in \mathbb{N}$.

To summarize, by suitably choosing N and R (possibly depending on N) sufficiently large the difference between the distributions of the restricted MCwM and the MH algorithms after n-steps can be made arbitrarily small.

4.2 Latent variables

In this section we consider π_u of the form (19). Here, as for doubly intractable distributions, the idea is to substitute $\pi_u(x)$ in the acceptance probability of the MH algorithm by a Monte Carlo estimate

$$\widehat{\rho}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \bar{\rho}(x, Y_i^{(x)})$$

where we assume that we have access to an iid sequence of random variables $(Y_i^{(x)})_{1\leq i\leq N}$ where each $Y_i^{(x)}$ has distribution r_x. Define a function $W_N\colon G \to \mathbb{R}$ by $W_N(x) := \widehat{\rho}_N(x)/\pi_u(x)$ and note that $\mathbb{E}[W_N(x)] = 1$. Then, the acceptance probability given $W_N(x)$, $W_N(z)$ modifies to

$$\widetilde{a}(x, z, W_N) := \min\left\{1, r(x,z)\cdot \frac{W_N(z)}{W_N(x)}\right\}$$

where $W_N(x)$, $W_N(z)$ are assumed to be independent random variables. Note that all the objects which depend on $\widetilde{a}$, such as a_N, $M_{a_N}, a_N^{(R)}, M_{a_N^{(R)}}$, are in this section defined just as in Section 4.1. The only difference is that the ratio within the minimum in $\widetilde{a}$ is reversed compared to (21). Thus, this leads to a MCwM algorithm as stated in Algorithm 2, where the transition kernel is given by M_{a_N}.

Also as in Section 4.1 we define $s(x) := \left(\mathbb{E}\left|W_1(x) - 1\right|^2\right)^{1/2}$ and $i_{p,N}(x) := (\mathbb{E}W_N(x)^{-p})^{1/p}$ for all $x \in G$ and $p > 0$. With those quantities we obtain the following estimate of the difference of the acceptance probabilities of M_a and M_{a_N} proved in Appendix A.2.

Lemma 19. Assume that there exists $k \in \mathbb{N}$ such that $i_{2,k}(x)$ and $s(x)$ are finite for all $x \in G$. Then, for all $x, z \in G$ and $N \geq k$ we have

$$|a(x,z) - a_N(x,z)| \leq a(x,z)\,\frac{1}{\sqrt{N}}\,i_{2,k}(x)(s(x) + s(z)). \tag{24}$$

If $\|s\|_\infty$ and $\|i_{2,k}\|_\infty$ are finite for some $k \in \mathbb{N}$, then the same statement as formulated in Corollary 16 holds. The proof works exactly as stated there. Examples which satisfy this condition are for instance presented in [MLR18]. However, there are cases where the functions s and $i_{2,k}$ are unbounded. In this setting, as in Section 4.1.2, we consider the restricted MCwM algorithm with transition kernel $M_{a_N^{(R)}}$. Here again the same statement and proof as formulated in Corollary 17 hold. We next provide an application of this corollary in the latent variable setting.

Normal-normal model. Let $G = \mathbb{R}$ and the function φ_{μ,σ^2} be the density of $\mathcal{N}(\mu, \sigma^2)$. For some $z \in \mathbb{R}$ and (precision) parameters $\gamma_Z, \gamma_Y > 0$ define

$$\pi_u(x) := \int_{\mathbb{R}} \varphi_{z,\gamma_Z^{-1}}(y)\, \varphi_{0,\gamma_Y^{-1}}(x-y)\mathrm{d}y,$$

that is, $\mathcal{Y} = \mathbb{R}$, $\bar{\rho}(x,y) = \varphi_{z,\gamma_Z^{-1}}(y)$ and $r_x = \mathcal{N}(x, \gamma_Y^{-1})$. By the convolution of two normals the target distribution π satisfies

$$\pi_u(x) = \varphi_{z,\gamma_{Z,Y}^{-1}}(x), \quad \text{with} \quad \gamma_{Z,Y}^{-1} := \gamma_Z^{-1} + \gamma_Y^{-1}. \tag{25}$$

Note that, for real-valued random variables Y, Z the probability measure π is the posterior distribution given an observation $Z = z$ within the model

$$Z|Y = y \sim \mathcal{N}\left(y, \gamma_Z^{-1}\right), \qquad Y|x \sim \mathcal{N}\left(x, \gamma_Y^{-1}\right),$$

with the improper Lebesgue prior imposed on x.

Pretending that we do not know $\pi_u(x)$ we compute

$$\widehat{\rho}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \varphi_{z,\gamma_Z^{-1}}(Y_i^{(x)}),$$

where $(Y_i^{(x)})_{1\le i\le N}$ is a sequence of iid random variables with $Y_1^{(x)} \sim \mathcal{N}(x, \gamma_Y^{-1})$. Hence

$$W_N(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{\varphi_{z,\gamma_Z^{-1}}(Y_i^{(x)})}{\varphi_{z,\gamma_{Z,Y}^{-1}}(x)} = \frac{1}{N}\left(\frac{\gamma_Z}{\gamma_{Z,Y}}\right)^{1/2} \sum_{i=1}^{N} \frac{\varphi_{0,1}(\sqrt{\gamma_Z}(z - Y_i^{(x)}))}{\varphi_{0,1}(\sqrt{\gamma_{Z,Y}}(z-x))}.$$

By using a random variable $\xi \sim \mathcal{N}(0,1)$ we have for $p > -\gamma_Y/\gamma_Z$ that

$$\begin{aligned}\mathbb{E}\left[W_1(x)^p\right] &= \left(\frac{\gamma_Z}{\gamma_{Z,Y}}\right)^{p/2} \mathbb{E}\left[\exp\left(\frac{p}{2}\gamma_{Z,Y}\,(z-x)^2 - \frac{p}{2}\frac{\gamma_Z}{\gamma_Y}(\gamma_Y^{1/2}(z-x) - \xi)^2\right)\right] \\ &\propto \exp\left(\frac{\gamma_Z\, \gamma_{Z,Y}\, p\,(p-1)}{2\,(\gamma_Y + p\gamma_Z)}\,(z-x)^2\right).\end{aligned} \tag{26}$$

Here $\propto$ means equal up to a constant independent of x. As a consequence, $\|s\|_\infty = \infty$ and therefore Corollary 16 (which is also true in the latent variable setting) cannot be applied. Nevertheless, we can obtain bounds for the restricted MCwM in this example using the statement of Corollary 17 by controlling s and $i_{2,k}$ using a Lyapunov function V. The following result, proved in Appendix A.2, verifies the necessary moment conditions under some additional restrictions on the model parameters.

Proposition 20. Assume that $\gamma_Y > \sqrt{2}\gamma_Z$, the unnormalized density π_u is given as in (25) and let the proposal transition kernel Q be a Gaussian random walk, that is, $Q(x, \cdot) = \mathcal{N}(x, \sigma^2)$ for some $\sigma > 0$. Then, there is a Lyapunov function $V\colon G \to [1, \infty)$ for M_a, such that M_a is V-uniformly ergodic, i.e., Assumption 12 is satisfied, and there are $\beta \in (0, 1)$ as well as $k \in \mathbb{N}$ such that

$$\|i_{2,k}\|_{\infty, V^{1-\beta}} < \infty \qquad \text{and} \qquad \|s\|_{\infty, V^\beta} < \infty.$$

The previous proposition implies that there is a constant $D < \infty$, such that D_R from Corollary 17 is bounded by D independent of R. Hence there are numbers $\widetilde{C}_1, \widetilde{C}_2 > 0$ such that with $R = \widetilde{C}_1\sqrt{N}$ and for N sufficiently large we have

$$\left\| m_n - m_{n,N}^{(R)} \right\|_{\mathrm{tv}} \leq \widetilde{C}_2 \, \frac{\log N}{\sqrt{N}}$$

for any initial distribution m_0 on B_R.

Acknowledgments. Daniel Rudolf gratefully acknowledges support of the Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences (Volkswagen Foundation), the Campus laboratory AIMS and the DFG within the project 389483880. Felipe Medina-Aguayo was supported by BBSRC grant BB/N00874X/1 and thanks Richard Everitt for useful discussions.

A Technical proofs

A.1 Proofs of Section 3

Before we come to the proofs of Section 3 let us recall a relation between geometric ergodicity and an ergodicity coefficient. Let $V\colon G \to [1, \infty]$ be a measurable, π-a.e. finite function, then, define the *ergodicity coefficient* $\tau_V(P)$ as

$$\tau_V(P) := \sup_{x,y\in G} \frac{\|P(x,\cdot) - P(y,\cdot)\|_V}{V(x)+V(y)}.$$

The next lemma provides a relation between the ergodicity coefficient and V-uniform ergodicity.

Lemma 21. If (7) is satisfied, then $\tau_V(P^n) \leq C\alpha^n$.

A proof of this fact is implicitly contained in [MZZ13] and can also be found in [RS18, Lemma 3.2]. Both references crucially use an observation of Hairer and Mattingly [HM11].

To summarize, if the transition kernel P is geometrically ergodic, then, by Theorem 1 there exist a function $V\colon G \to [1,\infty)$, $\alpha \in [0,1)$ and $C \in (0,\infty)$ such that, by Lemma 21, $\tau_V(P^n) \leq C\alpha^n$. The next proposition states two further useful properties (submultiplicativity and contractivity) of the ergodicity coefficient. For a proof of the corresponding inequalities see for example [MZZ13, Proposition 2.1].

Proposition 22. Assume P, Q are transition kernels and μ, ν are probability measures on G. Then

$$\begin{aligned} \tau_V(PQ) &\leq \tau_V(P)\,\tau_V(Q), && \text{(submultiplicativity)}\\ \|(\mu-\nu)P\|_V &\leq \tau_V(P)\,\|\mu-\nu\|_V\,. && \text{(contractivity)} \end{aligned}$$

Now we prove Lemma 4.

Proof of Lemma 4. As in the proof of [Mit05, Theorem 3.1] we use

$$\widetilde{p}_n - p_n = (\widetilde{p}_0 - p_0)P^n + \sum_{i=0}^{n-1} \widetilde{p}_i(\widetilde{P} - P)P^{n-i-1},$$

which can be shown by induction over $n \in \mathbb{N}_0$. Then

$$\|\widetilde{p}_n - p_n\|_{\text{tv}} \leq \|(\widetilde{p}_0 - p_0)P^n\|_{\text{tv}} + \sum_{i=0}^{n-1} \left\|\widetilde{p}_i(\widetilde{P} - P)P^{n-i-1}\right\|_{\text{tv}}. \tag{27}$$

With Proposition 22 and Lemma 21 we estimate the first term of the previous inequality by

$$\|(\widetilde{p}_0 - p_0)P^n\|_{\text{tv}} \leq \|(\widetilde{p}_0 - p_0)P^n\|_V \leq \tau_V(P^n)\,\|\widetilde{p}_0 - p_0\|_V \leq C\alpha^n\,\|\widetilde{p}_0 - p_0\|_V\,.$$

For the terms which appear in the sum of (27) we can use two types of estimates. Note that $\tau_1(P) \le 1$ (here the subscript indicates that $V = 1$) which leads by Proposition 22 to

$$\begin{aligned}
&\left\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\right\|_{\mathrm{tv}} \le \left\|\widetilde{p}_i(\widetilde{P}-P)\right\|_{\mathrm{tv}} \tau_1(P^{n-i-1}) \le \left\|\widetilde{p}_i(\widetilde{P}-P)\right\|_{\mathrm{tv}} \\
&= \sup_{|f|\le 1} \left| \int_G f(x)\, \widetilde{p}_i(\widetilde{P}-P)(\mathrm{d}x) \right| = \sup_{|f|\le 1} \left| \int_G (\widetilde{P}-P)f(x)\, \widetilde{p}_i(\mathrm{d}x) \right| \\
&\le \int_G \left\| \widetilde{P}(x,\cdot) - P(x,\cdot) \right\|_{\mathrm{tv}} \widetilde{p}_i(\mathrm{d}x) \le \varepsilon_{\mathrm{tv},W}\ \widetilde{p}_i(W).
\end{aligned}$$

On the other hand

$$\begin{aligned}
&\left\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\right\|_{\mathrm{tv}} \le \left\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\right\|_{V} \le \left\|\widetilde{p}_i(\widetilde{P}-P)\right\|_{V} \tau_V(P^{n-i-1}) \\
&\le C\alpha^{n-i-1} \left\|\widetilde{p}_i(\widetilde{P}-P)\right\|_{V} \le C\alpha^{n-i-1} \int_G \left\| \widetilde{P}(x,\cdot) - P(x,\cdot) \right\|_{V} \widetilde{p}_i(\mathrm{d}x) \\
&\le C\alpha^{n-i-1} \varepsilon_{V,W}\ \widetilde{p}_i(W).
\end{aligned}$$

Thus, for any $r \in (0,1]$ we obtain

$$\begin{aligned}
\left\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\right\|_{\mathrm{tv}} &\le \left\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\right\|_{\mathrm{tv}}^{1-r} \cdot \left\|\widetilde{p}_i(\widetilde{P}-P)P^{n-i-1}\right\|_{\mathrm{tv}}^{r} \\
&\le \varepsilon_{\mathrm{tv},W}^{1-r}\ \varepsilon_{V,W}^{r}\ C^r\ \widetilde{p}_i(W)\ \alpha^{(n-i-1)r},
\end{aligned}$$

which gives by (27) the final estimate. □

Next we prove Theorem 10.

Proof Theorem 10. Locally for $x \in B_R$ we have $P_R V(x) \le PV(x) \le \delta V(x) + L$, and, eventually,

$$\begin{aligned}
\widetilde{P}V(x) &\le P_R V(x) + \left| \widetilde{P}V(x) - P_R V(x) \right| \\
&\le \delta V(x) + R \left\| \widetilde{P}(x,\cdot) - P_R(x,\cdot) \right\|_{\mathrm{tv}} + L \\
&\le (\delta + R \cdot \Delta(R))V(x) + L. \qquad (28)
\end{aligned}$$

We write B_R^c for $G \setminus B_R$ and obtain for $x \in B_R^c$ that

$$\widetilde{P}V(x) = \int_{B_R} V(y)\widetilde{P}(x,\mathrm{d}y) \le V(x). \qquad (29)$$

Denote $\widetilde{\delta} := \delta + R \cdot \Delta(R) \leq 1/2 + \delta/2 < 1$. For $i \geq 2$ we obtain by (28), (29) and $(1-\widetilde{\delta}^i) \leq 2(1-\widetilde{\delta}^{i-1})$ that

$$\begin{aligned}\widetilde{p}_i(V) &\leq \widetilde{\delta}^i \int_{B_R} V(x) p_0(\mathrm{d}x) + (1-\widetilde{\delta}^i)\frac{L}{1-\widetilde{\delta}} \\ &\quad + \widetilde{\delta}^{i-1} \int_{B_R^c} \widetilde{P}V(x) p_0(\mathrm{d}x) + (1-\widetilde{\delta}^{i-1})\frac{L}{1-\widetilde{\delta}} \\ &\leq \widetilde{\delta}^{i-1} p_0(V) + (1-\widetilde{\delta}^{i-1})\frac{3L}{1-\widetilde{\delta}} \leq 6\kappa.\end{aligned}$$

Furthermore, $p_0(V) \leq \kappa$ and $\widetilde{p}_1(V) \leq 2\kappa$. Now it is easily seen that $\sum_{i=0}^{n-1} \widetilde{p}_i(V) \alpha^{(n-i-1)r} \leq \frac{6\kappa}{r(1-\alpha)}$.

For $\varepsilon_{\mathrm{tv},V}$ we have

$$\varepsilon_{\mathrm{tv},V} \leq \max\left\{ \sup_{x \in B_R} \frac{\left\| P(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_{\mathrm{tv}}}{V(x)}, \sup_{x \in B_R^c} \frac{\left\| P(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_{\mathrm{tv}}}{V(x)} \right\}.$$

The second term in the maximum is bounded by $2/R$. For $x \in B_R$ we have

$$\begin{aligned}\left\| P(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_{\mathrm{tv}} &\leq \| P(x,\cdot) - P_R(x,\cdot) \|_{\mathrm{tv}} + \left\| P_R(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_{\mathrm{tv}} \\ &\leq 2P(x, B_R^c) + \left\| P_R(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_{\mathrm{tv}}\end{aligned}$$

so that the first term in the maximum satisfies

$$\sup_{x \in B_R} \frac{\left\| P(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_{\mathrm{tv}}}{V(x)} \leq \Delta(R) + 2 \sup_{x \in B_R} \frac{P(x, B_R^c)}{V(x)}.$$

Consider a random variable X_1^x with distribution $P(x,\cdot)$, $x \in B_R$. Applying Markov's inequality to the random variable $V(X_1^x)$ leads to

$$PV(x) = \mathbb{E}[V(X_1^x)] \geq R \cdot \mathbb{P}(V(X_1^x) > R) = R \cdot P(x, B_R^c),$$

and thus

$$\sup_{x \in B_R} \frac{P(x, B_R^c)}{V(x)} \leq \sup_{x \in B_R} \frac{PV(x)}{R \cdot V(x)} \leq \frac{\delta + L}{R}.$$

Finally, $R \cdot \Delta(R) < 1 - \delta$ and $L \geq 1$ imply $\varepsilon_{\mathrm{tv},V} \leq \frac{2(L+1)}{R}$.

We obtain $\varepsilon_{V,V} \leq 2(L+1)$ by the use of

$$\left\| P(x,\cdot) - \widetilde{P}(x,\cdot) \right\|_V \leq PV(x) + \widetilde{P}V(x),$$

the fact that $\sup_{x\in G} \frac{PV(x)}{V(x)} \leq \delta + L$ and

$$\begin{aligned}\sup_{x\in G} \frac{\widetilde{P}V(x)}{V(x)} &\leq \max\left\{ \sup_{x\in B_R} \frac{\widetilde{P}V(x)}{V(x)}, \sup_{x\in B_R^{\mathrm{c}}} \frac{\widetilde{P}V(x)}{V(x)} \right\} \\ &\underset{(28),(29)}{\leq} \max\left\{ \widetilde{\delta} + L, 1 \right\} \leq L+1.\end{aligned}$$

Then, by Lemma 4 for $r \in (0,1]$,

$$\|p_n - \widetilde{p}_n\|_{\mathrm{tv}} \leq \frac{12C^r(L+1)\kappa}{r\cdot R^{1-r}(1-\alpha)} \leq \frac{12C(L+1)\kappa}{r\cdot R^{1-r}(1-\alpha)}.$$

By minimizing over r we obtain for $R \geq \exp(1)$ that

$$\|p_n - \widetilde{p}_n\|_{\mathrm{tv}} \leq \frac{12C(L+1)\kappa}{1-\alpha} \cdot \frac{R^{1/\log(R)}\log(R)}{R},$$

which yields the assertion. □

A.2 Proofs of Section 4

We start with the proof of Proposition 13.

Proof of Proposition 13. For any $f\colon G \to \mathbb{R}$ we have

$$\begin{aligned}M_b f(x) - M_c f(x) = &\int_G f(y)(b(x,y) - c(x,y))Q(x,\mathrm{d}y) \\ &+ f(x) \int_G (c(x,y) - b(x,y))Q(x,\mathrm{d}y).\end{aligned}$$

If $z = x$, we have for all $x \in B$ that

$$\begin{aligned}&\|M_b(x,\cdot) - M_c(x,\cdot)\|_{\mathrm{tv}} \leq 2\int_G |b(x,y) - c(x,y)|\, Q(x,\mathrm{d}y) \\ &\leq 2\int_B b(x,y)\xi(x)(\eta(x)+\eta(y))Q(x,\mathrm{d}y) \leq 2\xi(x)(\eta(x) + M_b(\eta\cdot \mathbf{1}_B)(x)) \\ &\leq 2\xi(x)(\eta(x) + M_b V^\beta(x)\, \|\eta\cdot\mathbf{1}_B\|_{\infty,V^\beta}) \\ &\leq 4T\, \|\xi\cdot\mathbf{1}_B\|_{\infty,V^{1-\beta}}\, \|\eta\cdot\mathbf{1}_B\|_{\infty,V^\beta}\, V(x),\end{aligned}$$

where we used that $\sup_{x\in G} \frac{M_b V(x)}{V(x)} \leq T$ implies $\sup_{x\in G} \frac{M_b V(x)^\beta}{V(x)^\beta} \leq T^\beta$ by Jensen's inequality. Moreover, for any $x \in B$ we obtain

$$\begin{aligned}
\|M_b(x,\cdot) - M_c(x,\cdot)\|_V &\leq \sup_{|f|\leq V} \Big| \int_G f(y)(b(x,y)-c(x,y))Q(x,\mathrm{d}y) \\
&\qquad + f(x)\left(\int_G (c(x,y)-b(x,y))Q(x,\mathrm{d}y)\right) \Big| \\
&\leq \int_G V(y)\,|b(x,y)-c(x,y)|\,Q(x,\mathrm{d}y) + V(x)\int_G |b(x,y)-c(x,y)|\,Q(x,\mathrm{d}y) \\
&\leq \int_B V(y)b(x,y)\xi(x)(\eta(x)+\eta(y))Q(x,\mathrm{d}y) \\
&\qquad + V(x)\int_B b(x,y)\xi(x)(\eta(x)+\eta(y))Q(x,\mathrm{d}y) \\
&\leq 2\,\|\eta\cdot\mathbf{1}_B\|_\infty\,\|\xi\cdot\mathbf{1}_B\|_\infty\,(M_bV(x)+V(x)),
\end{aligned}$$

which implies the assertion in that case. In the case where $z = y$, we have similarly for any $x \in B$ that

$$\begin{aligned}
&\|M_b(x,\cdot) - M_c(x,\cdot)\|_{\mathrm{tv}} \leq 2\eta(x)M_b(\xi\cdot\mathbf{1}_B) + 2M_b(\xi\cdot\eta\cdot\mathbf{1}_B) \\
&\leq 2\eta(x)\,\|\xi\cdot\mathbf{1}_B\|_{\infty,V^{1-\beta}}\,M_b(V^{1-\beta})(x) + 2\,\|\xi\cdot\eta\cdot\mathbf{1}_B\|_{\infty,V}\,M_bV(x) \\
&\leq 4L\,\|\eta\cdot\mathbf{1}_B\|_{\infty,V^\beta}\,\|\xi\cdot\mathbf{1}_B\|_{\infty,V^{1-\beta}}\,V(x)
\end{aligned}$$

and

$$\begin{aligned}
\|M_b(x,\cdot) - M_c(x,\cdot)\|_V &\leq \int_B V(y)b(x,y)\xi(y)(\eta(x)+\eta(y))Q(x,\mathrm{d}y) \\
&\qquad + V(x)\int_B b(x,y)\xi(y)(\eta(x)+\eta(y))Q(x,\mathrm{d}y) \\
&\leq 2\,\|\xi\cdot\mathbf{1}_B\|_\infty\,\|\eta\cdot\mathbf{1}_B\|_\infty\,(M_bV(x)+V(x)),
\end{aligned}$$

which finishes the proof. □

Before we come to further proofs of Section 4 we provide some properties of inverse moments of averages of non-negative real-valued iid random variables $(S_i)_{i\in\mathbb{N}}$. In this setting, the pth inverse moment, for $p > 0$, is defined by

$$j_{p,r} := \left(\mathbb{E}\left(\frac{1}{r}\sum_{i=1}^r S_i\right)^{-p}\right)^{1/p}.$$

Lemma 23. Assume that $j_{p,r} < \infty$ for some $r \in \mathbb{N}$ and $p > 0$. Then

i) $j_{p,s} \leq j_{p,r}$ for $s \in \mathbb{N}$ with $s \geq r$;

ii) $j_{q,r} \leq j_{p,r}$ for $0 < q < p$;

iii) $j_{k \cdot p, k \cdot r} \leq j_{p,r}$ for any $k \in \mathbb{N}$.

Proof. Properties i) and ii) follow as in [MLR16, Lemma 3.5]. For proving iii) we have to show that

$$\mathbb{E}\left[\left(\frac{1}{k \cdot r}\sum_{i=1}^{k \cdot r} S_i\right)^{-p \cdot k}\right] \leq \mathbb{E}\left[\left(\frac{1}{r}\sum_{i=1}^{r} S_i\right)^{-p}\right]^k .$$

To this end, observe first that we can write

$$\frac{1}{k \cdot r}\sum_{i=1}^{k \cdot r} S_i = \frac{1}{k}\sum_{i=1}^{k} V_i$$

where the "batch-means" $V_1, \ldots, V_k$ are non-negative, real-valued iid random variables which have the same distribution as $\frac{1}{r}\sum_{i=1}^{r} S_i$. With $Z_i = V_i^{-1}$ we obtain

$$\mathbb{E}\left[\left(\frac{1}{\frac{1}{k \cdot r}\sum_{i=1}^{k \cdot r} S_i}\right)^{p \cdot k}\right] = \mathbb{E}\left[\left(\frac{1}{\frac{1}{k}\sum_{i=1}^{k} \frac{1}{Z_i}}\right)^{p \cdot k}\right]$$

which is a moment of the harmonic mean of $Z_1, \ldots, Z_k$. Using the inequality between geometric and harmonic means as well as the independence we find that

$$\mathbb{E}\left[\left(\frac{1}{\frac{1}{k}\sum_{i=1}^{k} \frac{1}{Z_i}}\right)^{p \cdot k}\right] \leq \mathbb{E}\left[\prod_{i=1}^{k} Z_i^p\right] = \mathbb{E}\left[Z_1^p\right]^k = \mathbb{E}\left[\left(\frac{1}{\frac{1}{r}\sum_{i=1}^{r} S_i}\right)^{p}\right]^k . \quad \square$$

The previous lemma shows that when inverse moments of some positive order are finite, then so are inverse moments of all higher and lower orders if the sample size is adjusted accordingly.

Proof of Lemma 14. It is easily seen that

$$a(x,z)\mathbb{E}\left[\min\left\{1, \frac{W_N(x)}{W_N(z)}\right\}\right] \leq a_N(x,z)$$

for any $x, z \in G$. By virtue of Jensen's inequality and $\mathbb{E}[W_N(z)] = 1$ we have $\mathbb{E}[W_N(z)^{-1}] \geq 1$ as well as

$$a_N(x,z) \leq \min\left\{1, r(x,z) \cdot \mathbb{E}\left[\frac{W_N(x)}{W_N(z)}\right]\right\} \leq a(x,z)$$

where we also used the independence of $W_N(x)$ and $W_N(z)$ in the last inequality. (The previous arguments are similar to those in [MLR16, Lemma 3.3 and the proof of Lemma 3.2].) Note that $i_{2,N}(x) \leq i_{2,k}(x)$ for $N \geq k$ by Lemma 23. Hence, one can conclude that

$$\begin{aligned}
|a(x,z) - a_N(x,z)| &\leq a(x,z) \begin{cases} \mathbb{E}\left[\max\left\{0, 1 - \frac{W_N(x)}{W_N(z)}\right\}\right] & a(x,z) \geq a_N(x,z) \\ \mathbb{E}\left[\frac{W_N(x)}{W_N(z)} - 1\right] & a(x,z) < a_N(x,z) \end{cases} \\
&\leq a(x,z)\mathbb{E}\left|1 - \frac{W_N(x)}{W_N(z)}\right| \leq a(x,z)\, i_{2,N}(z) \left(\mathbb{E}\left|W_N(x) - W_N(z)\right|^2\right)^{1/2} \\
&\leq a(x,z) i_{2,N}(z) \left[\left(\mathbb{E}\left|W_N(x) - 1\right|^2\right)^{1/2} + \left(\mathbb{E}\left|W_N(z) - 1\right|^2\right)^{1/2}\right] \\
&\leq a(x,z) \frac{i_{2,k}(z)}{\sqrt{N}} (s(x) + s(z)).
\end{aligned}$$

□

Proof of Lemma 19. As in the previous proof or from [MLR16, Lemma 3.3 and the proof of Lemma 3.2] an immediate consequence is

$$a(x,z)\mathbb{E}\left[\min\left\{1, \frac{W_N(z)}{W_N(x)}\right\}\right] \leq a_N(x,z) \leq a(x,z)\, \mathbb{E}\left[\frac{W_N(z)}{W_N(x)}\right].$$

Note that $i_{2,N} \leq i_{2,k}$ for $N \geq k$, see Lemma 23. The rest of the lemma follows as in the previous proof, only the ratio $W_N(x)/W_N(z)$ is reversed. □

Proof of Proposition 20. For random-walk-based Metropolis chains (in particular for Q as assumed in the statement) by [JH00, Theorem 4.1 and the first sentence after the proof of the theorem, as well as, Theorem 4.3, Theorem 4.6] we have that M_a is V_t-uniformly ergodic with

$$V_t(x) \propto \pi_u(x)^{-t} \propto \exp\left(t\, \frac{\gamma_{Z,Y}}{2} (z - x)^2\right),$$

for any $t \in (0,1)$. Hence, Assumption 12 is satisfied and we need to find $t \in (0,1)$ as well as $\beta \in (0,1)$ such that $\|i_{2,k}\|_{\infty, V_t^{1-\beta}} < \infty$ and $\|s\|_{\infty, V_t^{\beta}} < \infty$

for some $k \in \mathbb{N}$. For showing $\|s\|_{\infty, V_t^\beta} < \infty$ we use (26) to see that

$$s\,(x) \leq \widetilde{C} \exp\left(\left(\frac{\gamma_Z}{\gamma_Y + 2\gamma_Z}\right) \frac{\gamma_{Z,Y}}{2} \,(z - x)^2\right),$$

for some $\widetilde{C} < \infty$. Hence

$$\frac{s(x)}{V_t(x)^\beta} \leq \widetilde{C} \exp\left(\left(\frac{\gamma_Z}{\gamma_Y + 2\gamma_Z} - t\beta\right) \frac{\gamma_{Z,Y}}{2}(z - x)^2\right),$$

and choosing $\beta \in (0, 1)$ such that

$$t\beta = \frac{\gamma_Z}{\gamma_Y + 2\gamma_Z} \tag{30}$$

leads to $\|s\|_{\infty, V_t^\beta} < \infty$. In order to show $\|i_{2,k}\|_{\infty, V_t^{1-\beta}} < \infty$, we first use Lemma 23 iii) and obtain for any $x \in G$ and any $k \in \mathbb{N}$

$$i_{2,k}(x) = \mathbb{E}[W_k(x)^{-2}]^{\frac{1}{2}} \leq \mathbb{E}\left[W_1(x)^{-\frac{2}{k}}\right]^{\frac{k}{2}}.$$

Then, for $k > 2\gamma_Z/\gamma_Y$ by (26) we have

$$\mathbb{E}\left[W_1(x)^{-\frac{2}{k}}\right]^{\frac{k}{2}} \propto \exp\left(\left(\frac{\gamma_Z\left(1 + \frac{2}{k}\right)}{\gamma_Y - \frac{2}{k}\gamma_Z}\right) \frac{\gamma_{Z,Y}}{2} \,(z - x)^2\right).$$

Therefore, there is a constant $\widetilde{C} < \infty$ such that

$$\frac{i_{2,k}(x)}{V_t(x)^{1-\beta}} \leq \widetilde{C} \exp\left(\left(\frac{\gamma_Z\left(1 + \frac{2}{k}\right)}{\gamma_Y - \frac{2}{k}\gamma_Z} - t(1 - \beta)\right) \frac{\gamma_{Z,Y}}{2}(z - x)^2\right).$$

We have $\|i_{2,k}\|_{\infty, V_t^{1-\beta}} < \infty$ if $\frac{\gamma_Z\left(1+\frac{2}{k}\right)}{\gamma_Y - \frac{2}{k}\gamma_Z} \leq t(1 - \beta)$. The latter condition holds whenever

$$k \geq \frac{2\gamma_Z\,(1 + t(1 - \beta))}{\gamma_Y t(1 - \beta) - \gamma_Z},$$

provided that $t(1 - \beta) > \gamma_Z/\gamma_Y$. This implies, by (30), that t should be chosen such that

$$t > \frac{\gamma_Z}{\gamma_Y} + \frac{\gamma_Z}{\gamma_Y + 2\gamma_Z}. \tag{31}$$

Choosing t such that it satisfies (31) is feasible whenever the right-hand side of (31) is smaller than 1. This is the case if $\gamma_Y > \sqrt{2}\gamma_Z$. □

References

[ADYC18] Ch. Andrieu, A. Doucet, S. Yıldırım, and N. Chopin, *On the utility of Metropolis-Hastings with asymmetric acceptance ratio*, arXiv preprint arXiv:1803.09527 (2018).

[AFEB16] P. Alquier, N. Friel, R. Everitt, and A. Boland, *Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels*, Statistics and Computing **26** (2016), no. 1, 29–47.

[AR09] C. Andrieu and G. Roberts, *The pseudo-marginal approach for efficient Monte Carlo computations*, Ann. Statist. **37** (2009), no. 2, 697–725.

[BDH14] R. Bardenet, A. Doucet, and C. Holmes, *Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach*, Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 405–413.

[BRR01] L. Breyer, G. Roberts, and J. Rosenthal, *A note on geometric ergodicity and floating-point roundoff error*, Statist. Probab. Lett. **53** (2001), no. 2, 123–127.

[EJREH17] R. G. Everitt, A. M. Johansen, E. Rowing, and M. Evdemon-Hogan, *Bayesian model comparison with un-normalised likelihoods*, Statistics and Computing **27** (2017), no. 2, 403–422.

[FHL13] D. Ferré, L. Hervé, and J. Ledoux, *Regular perturbation of V-geometrically ergodic Markov chains*, J. Appl. Prob. **50** (2013), no. 1, 184–194.

[HM11] M. Hairer and J. C. Mattingly, *Yet another look at Harris ergodic theorem for Markov chains*, Seminar on Stochastic Analysis, Random Fields and Applications VI, Springer, 2011, pp. 109–117.

[JH00] S. Jarner and E. Hansen, *Geometric ergodicity of Metropolis algorithms*, Stochastic Process. Appl. **85** (2000), no. 2, 341–361.

[JM17a] J. E. Johndrow and J. C. Mattingly, *Coupling and Decoupling to bound an approximating Markov Chain*, ArXiv preprint arXiv:1706.02040 (2017).

[JM17b] ———, *Error bounds for Approximations of Markov chains used in Bayesian Sampling*, ArXiv preprint arXiv:1711.05382 (2017).

[JMMD15] J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson, *Optimal approximating Markov chains for Bayesian inference*, ArXiv preprint arXiv:1508.03387 (2015).

[MGM06] I. Murray, Z. Ghahramani, and D. MacKay, *MCMC for doubly-intractable distributions*, Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence UAI06, 2006.

[Mit05] A. Mitrophanov, *Sensitivity and convergence of uniformly ergodic Markov chains*, J. Appl. Prob. **42** (2005), no. 4, 1003–1014.

[MLR16] F. J. Medina-Aguayo, A. Lee, and G. Roberts, *Stability of Noisy Metropolis-Hastings*, Stat. Comp. **26** (2016), no. 6, 1187–1211.

[MLR18] F. J. Medina-Aguayo, A. Lee, and G. O. Roberts, *Erratum to: Stability of noisy Metropolis–Hastings*, Stat. Comp. **28** (2018), no. 1, 239–239.

[MT09] S. Meyn and R. Tweedie, *Markov chains and stochastic stability*, second ed., Cambridge University Press, 2009.

[MZZ13] Y. Mao, M. Zhang, and Y. Zhang, *A generalization of Dobrushin coefficient*, Chinese J. Appl. Probab. Statist. **29** (2013), no. 5, 489–494.

[NR17] J. Negrea and J. S. Rosenthal, *Error Bounds for Approximations of Geometrically Ergodic Markov Chains*, ArXiv preprints arXiv:1702.07441 (2017).

[PS14] N. Pillai and A. Smith, *Ergodicity of approximate MCMC chains with applications to large data sets*, arXiv preprint arXiv:1405.0182 (2014).

[RR97] G. Roberts and J. Rosenthal, *Geometric ergodicity and hybrid Markov chains*, Electron. Comm. Probab. **2** (1997), no. 2, 13–25.

[RRS98] G. Roberts, J. Rosenthal, and P. Schwartz, *Convergence properties of perturbed Markov chains*, J. Appl. Probab. **35** (1998), no. 1, 1–11.

[RS18] D. Rudolf and N. Schweizer, *Perturbation theory for Markov chains via Wasserstein distance*, Bernoulli **24** (2018), no. 4A, 2610–2639.

[SS00] T. Shardlow and A. Stuart, *A perturbation theory for ergodic Markov chains and application to numerical approximations*, SIAM J. Numer. Analysis **37** (2000), 1120–1137.

[Tie98] L. Tierney, *A note on Metropolis-Hastings kernels for general state spaces*, Ann. Appl. Probab. **8** (1998), 1–9.