Check for updates

J. Appl. Probab. **62,** 1318–1339 (2025)
*doi:*10.1017/jpr.2025.4

WEAK CONVERGENCE OF ADAPTIVE MARKOV CHAIN MONTE CARLO

AUSTIN BROWN ,* ** AND
JEFFREY S. ROSENTHAL ,* *** *University of Toronto*

Abstract

We develop general conditions for weak convergence of adaptive Markov chain Monte Carlo processes and this is shown to imply a weak law of large numbers for bounded Lipschitz continuous functions. This allows an estimation theory for adaptive Markov chain Monte Carlo where previously developed theory in total variation may fail or be difficult to establish. Extensions of weak convergence to general Wasserstein distances are established, along with a weak law of large numbers for possibly unbounded Lipschitz functions. Applications are applied to autoregressive processes in various settings, unadjusted Langevin processes, and adaptive Metropolis–Hastings.

Keywords: Adaptive Markov chain Monte Carlo; Birkhoff ergodic theorem; weak convergence of adaptive MCMC

2010 Mathematics Subject Classification: Primary 60J05
Secondary 60J22

1. Introduction

Markov chain Monte Carlo (MCMC) provides a means to estimate integrals with respect to a target probability measure from the empirical average of a Markov chain. Many Markov chains require a delicate choice of tuning parameters to explore the state space properly, such as Metropolis–Hastings [31] and discretized Langevin diffusions [30, 35]. The optimal tuning parameter choice often depends on properties of the target probability measure, which may be challenging to compute precisely. At the same time, a poor tuning parameter choice may lead to unreliable estimation and diagnostics from the Markov chain. This motivates adaptive MCMC processes that automatically learn or adapt the tuning parameters of the Markov chain as the process progresses in time [18, 32].

The general theory for adaptive MCMC processes is accomplished through a convergence guarantee on the non-adapted Markov chain in the total variation distance combined with diminishing conditions on the adaptation of the tuning parameters [32]. Numerous adaptation strategies are possible such as stochastic approximation [29] or specifically designed strategies to rapidly decrease the adaptation as time progresses [9]. The existing general theory results in the ability to approximate arbitrary bounded functions through a weak law of large numbers. However, there has been increasing evidence that convergence in total variation is

Received 2 June 2024; accepted 9 January 2025.
* Postal address: Department of Statistical Sciences, University of Toronto, Toronto, Canada.
** Email address: ad.brown@utoronto.ca
*** Email address: jeff@math.toronto.edu

© The Author(s), 2025. Published by Cambridge University Press on behalf of Applied Probability Trust. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

inadequate for many high-dimensional target probability measures compared to convergence in Wasserstein distances from optimal transportation [15, 20, 27, 28]. The issues with analyzing convergence with total variation are not limited to high dimensions and may appear for certain diffusion processes in any dimension [20], and even for toy examples [8, 39].

Since the introduction of adaptive MCMC [18], many advancements have been made based upon convergence in total variation [2, 19, 33], but weak convergence appears less explored. For example, convergence theory for adaptive MCMC has been extended to handle augmented target distributions that may depend on the adaptation to target multi-modal distributions [26]. Under specific adaptation strategies based on stochastic approximation, convergence theory under stronger assumptions can lead to a central limit theorem [1]. However, each of these theoretical results and guarantees is based on convergence of the non-adapted Markov chain in total variation.

This article's main contribution is the weak convergence of adaptive MCMC processes under general conditions using Wasserstein distances that metrize the weak convergence of probability measures [17, 41]. Section 2 introduces the general adaptive MCMC regime, and Section 3 reviews the existing theory and some motivating examples that emphasize the inadequacy of the existing convergence theory. Section 5 extends the traditional convergence framework in total variation for adaptive MCMC [32] to a framework based on weak convergence. While the convergence result is weaker than total variation, it provides theoretical guarantees for approximations of bounded Lipschitz functions and arbitrary closed sets via Strassen's theorem [37]. Section 6 develops general conditions for a weak law of large numbers applied to bounded Lipschitz functions based on weak convergence.

Some examples and applications are explored in Section 5 with adapted autoregressive processes, adaptive unadjusted Langevin processes, adaptive Langevin diffusions, and adaptive Metropolis–Hastings. Beyond the examples studied here, the weak convergence theory for adaptive MCMC can be used to develop new adaptive algorithms for Bayesian inverse problems popular in physics that involve sampling posterior distributions on infinite-dimensional spaces where total variation can be problematic [10]. Another potentially useful application of the theory developed here is demonstrated in the adaptive Langevin diffusion example where using Wasserstein distances to show weak convergence can yield simpler proofs of the required conditions in comparison to proofs needed to show convergence in total variation.

Weak convergence and the law of large numbers are further extended to general Wasserstein distances under stronger conditions. The main application of this extension is a law of large numbers for unbounded Lipschitz functions in Section 6 and is of practical relevance in statistics. In particular, this extends the weak law of large numbers for Lipschitz functions for adaptive MCMC processes [32] and for Markov chains [36]. Recently, a law of large numbers for bounded Lipschitz functions has been developed under strong contraction conditions in Wasserstein distances combined with strong limitations on the adaptation [21]. The law of large numbers developed here holds under more general conditions and can apply to unbounded Lipschitz functions under suitable conditions. Section 9 discusses our theoretical results along with limitations and potential extensions of the newly developed theory.

2. Background: Adaptive Markov chain Monte Carlo processes

Let $\mathbb{Z}_+$ denote the positive integers and denote the minimum and maximum of $a, b \in \mathbb{R}$ by $a \wedge b$ and $a \vee b$ respectively. For a Borel measurable space S, let $\mathcal{B}(S)$ denote its Borel sigma field. The Euclidean norm is denoted by $\|\cdot\|$ and, for a measure μ, Lebesgue spaces are denoted by $L^p(\mu)$. For a real-valued function f, denote the optimal Lipschitz constant with respect to a metric d by $\|f\|_{\mathrm{Lip}(d)} = \sup_{x \neq y} |f(y) - f(x)|/d(x, y)$.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

We follow closely to the adaptive MCMC process framework of [32]. Let $(\Omega, \mathcal{B}(\Omega))$ be a Borel measurable space and let $\mathcal{X}$ and $\mathcal{Y}$ be complete separable metric spaces with respect to some metrics, with $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$ their respective Borel sigma fields. Let π be a target Borel probability measure on $\mathcal{X}$. For a discrete time index $t \in \mathbb{Z}_+$, the adaptive process updates a random tuning parameter $\Gamma_t \colon \Omega \mapsto \mathcal{Y}$ as the process progresses using the entire history to improve the distribution of $X_t \colon \Omega \mapsto \mathcal{X}$. The result is for the marginal distribution of X_t to approximate the target distribution π.

We define generalized Borel measurable probability transition kernels $(\mathcal{Q}_t)_{t\geq 0}$ with $\mathcal{Q}_t \colon (\mathcal{Y} \times \mathcal{X})^t \times \mathcal{B}(\mathcal{Y}) \mapsto [0, 1]$ and a family of Borel measurable Markov transition kernels $(\mathcal{P}_\gamma)_{\gamma\in\mathcal{Y}}$ with $\mathcal{P}_\gamma : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \mapsto [0, 1]$ to prescribe the adaptive process by the relations

$$\mathbb{P}(\Gamma_t \in \mathrm{d}\gamma \mid H_{t-1}) = \mathcal{Q}_t(H_{t-1}, \mathrm{d}\gamma), \qquad \mathbb{P}(X_t \in \mathrm{d}x \mid \Gamma_t, X_s, H_{t-1}) = \mathcal{P}_{\Gamma_t}(X_s, \mathrm{d}x),$$

where $H_t = (\Gamma_0, X_0, \ldots, \Gamma_t, X_t)$ denotes the history at time t. This prescribes the finite-dimensional distributions so that $(\Gamma_0, X_0, \ldots, \Gamma_t, X_t)$ for fixed Γ_0, X_0 has joint distribution

$$\mathcal{A}^{(0,\ldots,t)}((\gamma_0, x_0), \mathrm{d}\gamma_1, \mathrm{d}x_1, \ldots, \mathrm{d}\gamma_t, \mathrm{d}x_t) = \prod_{k=1}^{t} \mathcal{Q}_k(h_{k-1}, \mathrm{d}\gamma_k)\mathcal{P}_{\gamma_k}(x_{k-1}, \mathrm{d}x_k)$$

with history $h_k = (\gamma_0, x_0, \ldots, \gamma_k, x_k)$. This defines an *adaptive process* $(\Gamma_t, X_t)_{t\geq 0}$ adapted to the filtration $\mathcal{H}_t = \mathcal{B}(\Gamma_s, X_s \colon 0 \leq s \leq t)$ and initialized at any probability measure μ on $(\mathcal{Y} \times \mathcal{X}, \mathcal{B}(\mathcal{Y} \times \mathcal{X}))$ by the Ionescu–Tulcea extension theorem [38].

We will mostly be concerned with the marginal distribution X_t from fixed initialization points $\gamma, x \in \mathcal{Y} \times \mathcal{X}$ and general initializations μ on $(\mathcal{Y} \times \mathcal{X}, \mathcal{B}(\mathcal{Y} \times \mathcal{X}))$, defined by

$$X_t \mid \Gamma_0, X_0 = \gamma, x \sim \mathcal{A}^{(t)}((\gamma, x), \cdot), \qquad X_t \sim \mu\mathcal{A}^{(t)}(\,\cdot\,) = \int_{\mathcal{Y}\times\mathcal{X}} \mathcal{A}^{(t)}((\gamma, x), \cdot)\, \mu(\mathrm{d}\gamma, \mathrm{d}x). \quad (1)$$

3. Background: Wasserstein distances

Let $d \colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a lower semicontinuous metric. Define the Wasserstein distance or transportation distance of order $p \in \mathbb{Z}_+$ between two arbitrary Borel probability measures μ and ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ by

$$\mathcal{W}_{d,p}(\mu, \nu) = \left(\inf_{\xi\in\mathcal{C}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)^p\, \xi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p},$$

where $\mathcal{C}(\mu, \nu)$ is the set of all joint probability measures ξ such that $\xi(\,\cdot \times \mathcal{X}) = \mu(\,\cdot\,)$ and $\xi(\mathcal{X} \times \,\cdot\,) = \nu(\,\cdot\,)$. We generally suppress the 1 in the L^1 metric and write $\mathcal{W}_d(\mu, \nu) := \mathcal{W}_{d,1}(\mu, \nu)$.

The total variation distance denoted by $\mathcal{W}_{\mathrm{TV}}(\cdot, \cdot)$ between probability measures can be seen as a special case of a Wasserstein distance when the metric is defined by $I_{D^c}(\cdot, \cdot)$ with the off-diagonal set $D^c = \{x, y \in \mathcal{X} \times \mathcal{X} \colon x \neq y\}$. If $(\mathcal{X}, d)$ is a complete separable metric space, then the standard bounded metric $d(\cdot, \cdot) \wedge 1$ defines a Wasserstein distance $\mathcal{W}_{d\wedge 1}(\cdot, \cdot)$. This Wasserstein distance metrizes the weak convergence of probability measures through the bounded Lipschitz metric [13, Theorem 11.3.3] and is equivalent up to a constant to the bounded Lipschitz metric by Kantovorich–Rubinstein duality [40, Theorem 1.4].

Traditional theory of adaptive MCMC considers an adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ initialized at $\Gamma_0, X_0 = \gamma, x \in \mathcal{Y} \times \mathcal{X}$ satisfying a *(strong) diminishing adaptation* condition

$$\lim_{t\to\infty} \sup_{x\in\mathcal{X}} \mathcal{W}_{\mathrm{TV}}(\mathcal{P}_{\Gamma_{t+1}}(x, \cdot), \mathcal{P}_{\Gamma_t}(x, \cdot)) = 0 \quad (2)$$

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

in probability with the supremum assumed Borel measurable and a *(strong) containment* condition, which is to show that, for any $\varepsilon \in (0, 1)$, the sequence

$$M_\varepsilon(\Gamma_t, X_t) = \inf\left\{N \in \mathbb{Z}_+ \colon \mathcal{W}_{\mathrm{TV}}\left(\mathcal{P}^n_{\Gamma_t}(X_t, \cdot), \pi\right) \le \varepsilon \text{ for all } n \ge N\right\} \tag{3}$$

is bounded in probability, that is, $\lim_{N\to\infty} \sup_{t\ge 0} \mathbb{P}(M_\varepsilon(\Gamma_t, X_t) > N) = 0$. The (strong) diminishing adaptation restricts the adaptation plan for $(\Gamma_t)_t$ and (strong) containment is a uniform convergence requirement on the non-adapted Markov chain.

Under these two conditions, we have the guarantee [32] that for every fixed initialization $\gamma, x \in \mathcal{Y} \times \mathcal{X}$ and for every bounded Borel measurable function $\varphi \colon \mathcal{X} \to \mathbb{R}$,

$$\lim_{t\to\infty} \mathcal{W}_{\mathrm{TV}}\left(\mathcal{A}^{(t)}((\gamma, x), \cdot), \pi\right) = 0 \quad \text{and} \quad \frac{1}{t}\sum_{s=1}^{t} \varphi(X_s) \to \int_{\mathcal{X}} \varphi \, \mathrm{d}\pi \text{ in probability as } t \to \infty.$$

Both of these guarantees have many practical applications in the reliability of Monte Carlo simulations in Bayesian statistics. General conditions for (strong) containment (3) to hold have also been developed [3, 23].

The (strong) containment condition (3) is often established via simultaneous drift and minorization conditions [3, 32]. This requires a drift function $V \colon \mathcal{X} \to [0, \infty)$ and identification of a small set $S = \{x \in \mathcal{X} \colon V(x) \le R\}$ such that there are constants $\lambda, \alpha \in (0, 1)$ and $L \in (0, \infty)$ where $R > 2L/(1 - \lambda)$ is required and, for every $\gamma, x \in \mathcal{Y} \times \mathcal{X}$, there is a Borel probability measure ν_γ on $\mathcal{X}$ such that

$$\inf_{y\in S} \mathcal{P}_\gamma(y, \cdot) \ge \alpha \nu_\gamma(\,\cdot\,) \quad \text{and} \quad (\mathcal{P}_\gamma V)(x) \le \lambda V(x) + L. \tag{4}$$

These techniques yield (strong) containment (3) through a geometric rate $r \in (0, 1)$ and constant $M_0 > 0$ such that, for $t, n \in \mathbb{Z}_+$, $\mathcal{W}_{\mathrm{TV}}\left(\mathcal{P}^n_{\Gamma_t}(X_t, \cdot), \pi\right) \le M_0 r^n V(X_t)$ and $(V(X_t))_{t\ge 0}$ is bounded in probability [32, Theorem 18].

The identification of such a small set S and drift function V as in (4) often becomes problematic in large dimensions as probability measures often tend towards mutual singularity. Even in low dimensions, a small set may not exist as a non-adapted Markov kernel may fail to be irreducible, meaning that, for each $\gamma \in \mathcal{Y}$, there is no Borel probability measure φ_γ on $\mathcal{X}$ such that $\varphi_\gamma(\,\cdot\,) > 0$ implies $\mathcal{P}_\gamma(x, \cdot) > 0$ for all $x \in \mathcal{X}$. In this case, it is not possible to find such a small set regardless of the dimension [24, Theorem 5.2.2].

4. Motivating examples

The following running examples illustrate problematic points with analysis in total variation for adapting the tuning parameters of Markov chains compared to their alternative weak convergence properties. In particular, (strong) containment (3) may fail. Stark differences in the convergence characteristics may even appear when adapting a discrete Markov chain, as the following example illustrates.

Example 1. (*Discrete autoregressive process.*) Let $\gamma \in \mathbb{Z}_+$ with $\gamma \ge 2$ and $(\xi^\gamma_t)_{t\ge 0}$ be independent uniformly distributed discrete random variables on $\{0, 1/\gamma, 2/\gamma, \ldots, (\gamma - 1)/\gamma\}$. With $X_0 = x \in [0, 1)$, define the autoregressive process for $t \in \mathbb{Z}_+$ by $X^\gamma_t = (1/\gamma)X^\gamma_{t-1} + \xi^\gamma_t$. For each fixed $\gamma \ge 2$, this defines a Markov chain with Markov transition kernel denoted by $\mathcal{P}_\gamma$. It can be shown that the invariant probability measure $\pi \equiv \mathrm{Unif}(0, 1)$ is a Lebesgue measure on $[0, 1)$.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

For any adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ using these Markov kernels $(\mathcal{P}_\gamma)_\gamma$, traditional convergence theory in total variation [32, Theorem 13] is inadequate. Indeed, it can be shown that $\mathcal{W}_{\mathrm{TV}}\big(\mathcal{P}_\gamma^t(x,\cdot), \mathrm{Unif}(0,1)\big) = 1$ and (strong) containment (3) fails under any adaptive strategy. On the other hand, weak convergence is exponentially fast. For $t \in \mathbb{Z}_+$ and any fixed γ, starting from $X_0 = x \in [0,1)$ and $Y_0 = y \in [0,1)$, define $X_t^\gamma = (1/\gamma)X_{t-1}^\gamma + \xi_t^\gamma$ and $Y_t^\gamma = (1/\gamma)Y_{t-1}^\gamma + \xi_t^\gamma$ with shared discrete uniformly distributed random variable ξ_t^γ. These random variables (X_t^γ, Y_t^γ) define a coupling such that, for any $x \in [0,1)$ and $\gamma \geq 2$,

$$\mathcal{W}_{|\cdot|}\big(\mathcal{P}_\gamma^t(x,\cdot), \mathrm{Unif}(0,1)\big) \leq \int_{[0,1)} \mathbb{E}\big[|X_t^\gamma - Y_t^\gamma| X_0 = x, Y_0 = y\big]\,\mathrm{d}y \leq 2^{-t}.$$

In particular, we will show later that under a suitable adaptation strategy, the adaptive process converges weakly using this Wasserstein distance.

The next example shows how problems appear in infinite dimensions. Although the example is somewhat abstract, poor scaling properties in infinite dimensions can also appear in practical high-dimensional scenarios in statistics.

Example 2. (*Infinite-dimensional autoregressive process*.) Consider a Hilbert space H separable with infinite dimension and inner product $\langle\cdot,\cdot\rangle$. Let $\mathcal{N}(0, C)$ be a Gaussian Borel probability measure on H with mean $0 \in H$ and symmetric positive covariance operator $C\colon H \to H$ such that $\mathrm{tr}(C) = \sum_{k=1}^\infty \langle Cu_k, u_k\rangle < \infty$, where $(u_k)_k$ is any orthonormal basis of H. We will further assume that C is non-degenerate so that, for every $x, y \in H$, $Cx \equiv 0 \in H$ implies $x \equiv 0 \in H$. For some $\gamma^* < 1$, consider the family of Markov transition kernels $(\mathcal{P}_\gamma)_{\gamma\in(0,\gamma^*)}$ for the autoregressive process $(X_t^\gamma)_{t\geq 0}$ where the $(\xi_t)_t$ are independent with $\xi_t \sim \mathcal{N}(0, C)$ and

$$X_t^\gamma = \gamma X_{t-1}^\gamma + \sqrt{1-\gamma^2}\xi_t, \quad t \in \mathbb{Z}_+.$$

For any fixed $\gamma \in (0, \gamma^*)$, if $X_{t-1}^\gamma \sim \mathcal{N}(0, C)$, then $X_t^\gamma \sim \mathcal{N}(0, C)$ and the invariant probability measure is $\mathcal{N}(0, C)$.

For an adaptive autoregressive process $(\Gamma_t, X_t)_{t\geq 0}$ defined by (2), convergence theory in total variation [32, Theorem 13] fails to provide a convergence guarantee. For each $x \in H$ and $\gamma \in (0,1)$, $\mathcal{W}_{\mathrm{TV}}(\mathcal{P}_\gamma(x,\cdot), \mathcal{N}(0,C)) = 1$ due to the covariances differing and the Feldman–Hajeck theorem [12, Theorem 2.25]. It follows that (strong) containment (3) cannot hold under any adaptation strategy (2). However, convergence in L^2-Wasserstein distances is exponentially fast. Initialized with $X_0 = x \in H$ and $Y_0 = y \in H$, define $X_t^\gamma = \gamma X_{t-1}^\gamma + \sqrt{1-\gamma^2}\xi_t$ and $Y_t^\gamma = \gamma Y_{t-1}^\gamma + \sqrt{1-\gamma^2}\xi_t$ using the common random variable $\xi_t \sim N(0, C)$. This defines a coupling such that the L^2-Wasserstein distance is upper bounded with

$$\mathcal{W}_{\|\cdot\|,2}(\mathcal{P}_\gamma^t(x,\cdot), \mathcal{N}(0,C)) \leq \left[\int_H \mathcal{W}_{\|\cdot\|,2}(\mathcal{P}_\gamma^t(y,\cdot), \mathcal{P}_\gamma^t(x,\cdot))^2\,\pi(\mathrm{d}y)\right]^{1/2} \leq \gamma^{*t}[\,\|x\| + \sqrt{\mathrm{tr}(C)}].$$

5. Main results

This section extends previous results on convergence in total variation of adaptive MCMC processes to weak convergence and general Wasserstein distances [32, Theorems 5 and 13]. Let $\rho(\cdot,\cdot)$ be a lower semicontinuous metric on $\mathcal{X}$, so $\mathcal{W}_{\rho\wedge 1}(\cdot,\cdot)$ defines a Wasserstein distance. If (X, ρ), is a complete separable metric space, then $\mathcal{W}_{\rho\wedge 1}(\cdot,\cdot)$ metrizes weak convergence [13, Theorem 11.3.3]. A motivation for this convergence is Strassen's theorem, which gives

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

approximations to arbitrary closed sets [40, Corollary 1.28]. However, $\rho(\cdot, \cdot)$ need only satisfy the axioms of a metric and $\mathcal{W}_{\rho\wedge 1}(\cdot, \cdot)$ is defined more generally.

The first simple situation is to introduce a stopping time T such that the adaptation terminates and $\Gamma_T = \Gamma_t$ for all $t \geq T$. For any $T \geq 1$ determining a stopping point of adaptation, we can construct a finite adaptation process $(Y_t, \Gamma_t)_{t=0}^{\infty}$ adapted to the filtration $\mathcal{H}'_t = \mathcal{B}(Y_s, \Gamma_s \colon 0 \leq s \leq t)$ initialized at Γ_0, $Y_0, = \gamma, x$ such that $Y_t = X_t$ for $0 \leq t \leq T$ and is a Markov chain further out, that is, $Y_{t+1} \mid \mathcal{H}'_t, Y_t = y \sim \mathcal{P}_{\Gamma_T}(y, \cdot)$ for $t \geq T$. Denote the marginal distribution as $B^{(T,t-T)}((\gamma, x), \cdot) = \mathbb{P}(Y_t \in \cdot \mid \Gamma_0, Y_0 = \gamma, x)$.

Proposition 1. *Let $\rho(\cdot, \cdot)$ be a lower semicontinuous metric on $\mathcal{X}$ and let $(\Gamma_t, X_t)_{t\geq 0}$ be a finite adaptive process with initialization probability measure μ as in (1). If, for every initialization $x \in \mathcal{X}$ and every $\gamma \in \mathcal{Y}$, $\lim_{t\to\infty} \mathcal{W}_{\rho\wedge 1}\big(\mathcal{P}^t_\gamma(x, \cdot), \pi\big) = 0$, then $\lim_{t\to\infty} \mathcal{W}_{\rho\wedge 1}\big(\mu B^{(T,t-T)}, \pi\big) = 0$.*

Proof. Since the optimal coupling exists [41, Theorem 4.1] and is Borel measurable [41, Corollary 5.22], $\mathcal{W}_{\rho\wedge 1}\big(\mu B^{(T,t-T)}, \pi\big) \leq \mathbb{E}\big[\mathcal{W}_{\rho\wedge 1}(\mathcal{P}^{t-T}_{\Gamma_T}(X_T, \cdot), \pi)\big]$. The conclusion follows by dominated convergence. □

While finite adaptation may be a safe strategy, infinite adaptation where the process continually learns tuning parameters is often of greater interest in applications [32]. Consider now the following two weakened assumptions, both generalized from [32]. The first assumption is a weak restriction on the adaptation and the second is a weak containment condition on convergence of the non-adapted Markov chain.

Assumption 1. *Let $(\Gamma_t, X_t)_{t\geq 0}$ be an adaptive process with initialization probability measure μ as in (1). Let $\rho(\cdot, \cdot)$ and $\tilde{\rho}(\cdot, \cdot)$ be lower semicontinuous metrics on $\mathcal{X}$ such that $\rho(\cdot, \cdot) \wedge 1 \leq \tilde{\rho}(\cdot, \cdot) \wedge 1$. We make the following assumptions.*

(i) (Weak containment.) Suppose, for any $\varepsilon \in (0, 1)$, the sequence

$$M_{\varepsilon,\rho}(\Gamma_t, X_t) \colon = \inf\big\{N \geq 0 \colon \mathcal{W}_{\rho\wedge 1}\big[\mathcal{P}^n_{\Gamma_t}(X_t, \cdot), \pi\big] \leq \varepsilon \text{ for all } n \geq N\big\}$$

is bounded in probability, that is,

$$\lim_{T\to\infty} \sup_{t\geq 0} \mathbb{P}(M_{\varepsilon,\rho}(\Gamma_t, X_t) \geq T) = 0. \tag{5}$$

(ii) (Weak diminishing adaptation.) Suppose there is a conditional coupling $x, y, \gamma, \gamma' \mapsto \xi_{x,y,\gamma',\gamma} \in \mathcal{C}[\mathcal{P}_{\gamma'}(x, \cdot), \mathcal{P}_\gamma(y, \cdot)]$ and a non-negative real-valued sequence $(\delta_k)_{k\geq 0}$ with $\lim_{k\to\infty} \delta_k = 0$ such that

$$D_{\tilde{\rho}}(\Gamma_{t+1}, \Gamma_t) \colon = \lim_{k\to\infty} \sup_{\{x,y\in\mathcal{X}\times\mathcal{X} \colon \tilde{\rho}(x,y)\leq\delta_k\}} \int_{\mathcal{X}\times\mathcal{X}} \tilde{\rho}(x', y') \wedge 1 \, \xi_{x,y,\Gamma_{t+1},\Gamma_t}(\mathrm{d}x', \mathrm{d}y') \tag{6}$$

is $\mathcal{H}_{t+1}$-measurable and $D_{\tilde{\rho}}(\Gamma_{t+1}, \Gamma_t) \to 0$ in probability as $t \to \infty$.

There are existing results to bound the convergence rate of non-adapted Markov chains that can be modified to satisfy the weak containment condition using drift and coupling techniques [15, 20]. Note that $\rho(\cdot, \cdot) \wedge 1 \leq I_{\{x\neq y\}}(\cdot, \cdot)$ and (strong) diminishing adaptation [32] implies weak diminishing adaptation (6). We then immediately have Proposition 2. In certain cases

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

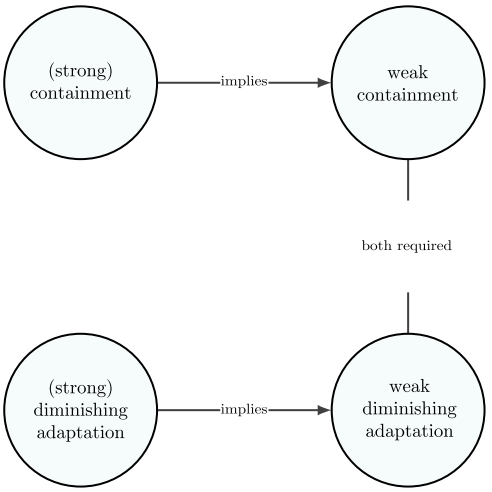


FIGURE 1. Illustration of comparison of strong/weak containment and strong/weak diminishing adaptation conditions required to obtain weak convergence of adaptive MCMC.

it may be simpler to show (strong) diminishing adaptation where only weak containment (5) holds; the implications of Proposition 2 are visualized in Figure 1.

Proposition 2. *Let* $(\Gamma_t, X_t)_{t\geq 0}$ *be an adaptive process with initialization probability measure* μ *as in (1). If the process satisfies (strong) containment (3), then weak containment (5) is satisfied. If the process satisfies (strong) diminishing adaptation (2), then weak diminishing adaptation (6) is satisfied.*

The following result shows weak convergence of the adaptive MCMC process.

Theorem 1. *Let* $(\Gamma_t, X_t)_{t\geq 0}$ *be an adaptive process with initialization probability measure* μ *as in (1). If weak containment (5) holds and weak diminishing adaptation (6) holds, then* $\lim_{t\to\infty} \mathcal{W}_{\rho\wedge 1}(\mu\mathcal{A}^{(t)}, \pi) = 0$.

We will prove Theorem 1 through the subsequent lemmas by comparing the adaptive process to an adaptive process where adaptation stops at a finite time. The first result shows that weak containment ensures the convergence of the finite adaptation process to the target measure uniformly in the finite adaptation stopping time.

Lemma 1. *If weak containment holds (5), then, for any* $\gamma, x \in \mathcal{Y}\times\mathcal{X}$,

$$\lim_{n\to\infty} \sup_{T\geq n} \mathcal{W}_{\rho\wedge 1}\big[B^{(T,n-T)}((\gamma, x), \cdot), \pi\big] = 0.$$

Proof. Fix $\varepsilon \in (0, 1)$. For any $\gamma, x \in \mathcal{Y}\times\mathcal{X}$ and each $n \in \mathbb{Z}_+$, the infimum is attained at an optimal coupling $\xi^{(n)}_{x,\gamma} \in \mathcal{C}[\mathcal{P}^n_\gamma(x, \cdot), \pi]$ [41, Theorem 4.1] so that

$$\mathcal{W}_{\rho\wedge 1}\big[\mathcal{P}^n_\gamma(x, \cdot), \pi\big] = \int_{\mathcal{X}^2} \rho(x', y') \wedge 1 \, \xi^{(n)}_{\gamma,x}(\mathrm{d}x', \mathrm{d}y').$$

The coupling is Borel measurable due to $\rho(\cdot, \cdot) \wedge 1$ being lower semicontinuous, and can be approximated by a non-decreasing sequence of bounded Lipschitz functions so we can

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

choose a measurable selection [41, Corollary 5.22] such that the limit is Borel measurable using the approximation techniques in [40, Theorem 1.3]. Define the set $A_\varepsilon = \{\gamma, x \in \mathcal{Y} \times \mathcal{X} : M_{\varepsilon,\rho}(\gamma, x) \le N\}$. For all $\gamma, x \in A_\varepsilon$ and for all $n \ge N$,

$$\mathcal{W}_{\rho\wedge 1}\big[\mathcal{P}_\gamma^n(x, \cdot), \pi\big] \le \varepsilon. \tag{7}$$

Let $\nu_{\gamma,x}^{(T)}$ denote the probability measure for (X_T, Γ_T) given $\Gamma_0, X_0 = \gamma, x$. Then

$$\hat{\xi}_{\gamma,x}^{(T+n)}(\mathrm{d}x_{T+n}, \mathrm{d}y) = \int_{\mathcal{Y}\times\mathcal{X}} \xi_{\gamma_T,y_T}^{(n)}(\mathrm{d}x', \mathrm{d}y')\nu_{\gamma,x}^{(T)}(\gamma_T, y_T)$$

defines a coupling for the finite adaptation process $Y_{T+n} \sim B^{(T,n)}((\gamma, x), \cdot)$ and $Y \sim \pi$ [41, Theorem 4.8]. By the weak containment assumption (5), there is an N depending on ε such that, uniformly in $T \ge 0$, $\nu_{\gamma,x}^{(T)}(A_\varepsilon^{\mathrm{c}}) = \mathbb{P}(M_{\varepsilon,\rho}(\Gamma_T, X_T) > N) \le \varepsilon$. Using (7), uniformly in $T \ge n$,

$$\begin{aligned}\mathcal{W}_{\rho\wedge 1}\big[B^{(T,n-T)}((\gamma, x), \cdot), \pi\big] &\le \int_{\mathcal{X}^2} \rho(x', y') \wedge 1 \, \hat{\xi}_{\gamma,x}^{(T+n)}(\mathrm{d}x', \mathrm{d}y') \\ &\le \int_{\mathcal{X}^2}\int_{A_\varepsilon} \rho(y_{T+n}, y) \wedge 1 \, \xi_{\gamma_T,y_T}^{(n)}(\mathrm{d}x', \mathrm{d}y')\nu_{\gamma,x}^{(T)}(\gamma_T, y_T) \\ &\quad + \sup_{T'\ge 0} \nu_{\gamma,x}^{(T')}(A_\varepsilon^{\mathrm{c}}) \le 2\varepsilon.\end{aligned}$$

□

The weak diminishing adaptation condition will ensure our next goal, which is to have the adaptive MCMC process converge to the finite adaptation process.

Lemma 2. *If weak diminishing adaptation (6) holds, then, for any $\gamma, x \in \mathcal{Y} \times \mathcal{X}$ and any $N \ge 0$, $\lim_{T\to\infty} \mathcal{W}_{\tilde{\rho}\wedge 1}\big(\mathcal{A}^{(T+N)}((\gamma, x), \cdot), B^{(T,N)}((\gamma, x), \cdot)\big) = 0$.*

Proof. It will suffice to assume that $\tilde{\rho} = \rho$ and the optimal coupling in the weak diminishing adaptation assumption (6). Fix $N \ge 1$ and $\varepsilon \in (0, 1)$. For each γ, γ' and each x, y, there exists a Borel measurable optimal coupling $\xi_{x,y,\gamma',\gamma}^*$ such that

$$\mathcal{W}_{\rho\wedge 1}[\mathcal{P}_{\gamma'}(x, \cdot), \mathcal{P}_\gamma(y, \cdot)] = \int_{\mathcal{X}^2} \rho(x', y') \wedge 1 \, \xi_{x,y,\gamma',\gamma}^*(\mathrm{d}x', \mathrm{d}y').$$

Using these conditional couplings, we define a joint probability measure ζ_{γ_0,x_0} by

$$\begin{aligned}&\zeta_{\gamma_0,x_0}(\mathrm{d}x_1, \mathrm{d}\gamma_1, \mathrm{d}y_1, \ldots, \mathrm{d}x_{T+N}, \mathrm{d}\gamma_{T+N}, \mathrm{d}y_{T+N}) \\ &\quad = \prod_{s=1}^{T} \mathcal{P}_{\gamma_s}(x_{s-1}, \mathrm{d}x_s)\mathcal{Q}_s(h_{s-1}, \mathrm{d}\gamma_s)\delta_{x_1,\ldots,x_s}(\mathrm{d}y_1, \ldots, \mathrm{d}y_s) \\ &\qquad \times \prod_{s=T+1}^{T+N} \xi_{x_{s-1},y_{s-1},\gamma_s,\gamma_T}^*(\mathrm{d}x_s, \mathrm{d}y_s)\mathcal{Q}_s(h_{s-1}, \mathrm{d}\gamma_s),\end{aligned}$$

where, for $0 \le s \le t$, the history $h_s = (\gamma_0, x_0, \ldots, \gamma_s, x_s)$. The marginal is a coupling $\zeta_{\gamma_0,x_0}(\mathrm{d}x_t, \mathrm{d}y_t)$ for the adaptive process $X_t \mid \Gamma_0, X_0 = \gamma, x$ and the finite adaptation process $Y_t \mid \Gamma_0, Y_0 = \gamma, x$ initialized so that they are identical up to time T and use conditional couplings thereafter. □

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

For $\gamma', \gamma \in \mathcal{Y}$ and $\delta \in (0, 1)$, define $D_{\rho,\delta}(\gamma', \gamma) = \sup_{\{x,y:\rho(x,y)\le\delta\}} \mathcal{W}_{\rho\wedge 1}(\mathcal{P}_{\gamma'}(x, \cdot), \mathcal{P}_{\gamma}(y, \cdot))$. For any $\varepsilon', \delta' \in (0, 1)$ and $k \in \mathbb{Z}_+$, define the set

$$E^{(T,N)}_{\varepsilon',\delta'} = \{\gamma_{T+1}, \dots, \gamma_{T+N} \colon D_{\rho,\delta'}(\gamma_{t+1}, \gamma_t) \le \varepsilon'/N^2, T+1 \le t \le T+N-1\}.$$

Starting with $\delta_N = r \in (0, 1)$, for each $1 \le k \le N$, given $\delta_k \in (0, 1)$, by weak diminishing adaptation (6) we can choose T large enough depending on ε, N, δ_k and δ_{k-1} sufficiently small such that, for all $t \ge T$, $D_{\rho,\delta}(\Gamma_{t+1}, \Gamma_t)$ is $\mathcal{H}_{t+1}$-measurable and

$$\mathbb{P}(D_{\rho,\delta_{k-1}}(\Gamma_{t+1}, \Gamma_t) > \delta_k\varepsilon/N^2) \le \varepsilon/N^2.$$

This constructs $\delta_0, \dots, \delta_N = r$ such that, using a union probability bound, we can choose T sufficiently large depending on $\varepsilon, N, \delta_1, \dots, \delta_N$ that, for each $1 \le k \le N$, we have $\mathbb{P}\big(E^{(T,N)}_{\varepsilon\delta_k,\delta_{k-1}}\big) \ge 1 - \varepsilon/N$. Define $E = \bigcap_{k=1}^N E^{(T,N)}_{\delta_k\varepsilon,\delta_{k-1}}$; a union probability bound then implies that

$$\mathbb{P}(E) = \mathbb{P}\left(\bigcap_{k=1}^N E^{(T,N)}_{\delta_k\varepsilon,\delta_{k-1}}\right) \ge 1 - \varepsilon.$$

The triangle inequality for the Wasserstein distance [40, Lemma 7.6] holds for every $1 \le k \le N$ with

$$\begin{aligned}\mathcal{W}_{\rho\wedge 1}[\mathcal{P}_{\gamma_{T+k}}(x, \cdot), \mathcal{P}_{\gamma_T}(y, \cdot)] &\le \mathcal{W}_{\rho\wedge 1}[\mathcal{P}_{\gamma_{T+1}}(x, \cdot), \mathcal{P}_{\gamma_T}(y, \cdot)] \\ &\quad + \sum_{s=1}^{N-1} \mathcal{W}_{\rho\wedge 1}[\mathcal{P}_{\gamma_{T+s+1}}(x, \cdot), \mathcal{P}_{\gamma_{T+s}}(x, \cdot)].\end{aligned}$$

For each k, if $\gamma_{T+1}, \dots, \gamma_{T+k} \in E^{(T,N)}_{\varepsilon\delta_k,\delta_{k-1}}$, then by the previous inequality and Markov's inequality,

$$\inf_{\{\rho(x,y)\le\delta_{k-1}\}} \mathcal{W}_{\{\rho(x',y')\le\delta_k\}}[\mathcal{P}_{\gamma_{T+k}}(x, \cdot), \mathcal{P}_{\gamma_T}(y, \cdot)] \ge 1 - \frac{\varepsilon}{N}.$$

By the construction of the distribution ζ, $\rho(x_T, y_T) \le \delta_0$ regardless of how small δ_0 is, and we have the lower bound

$$\zeta\left(\bigcap_{k=1}^N \{\rho(x_{T+k}, y_{T+k}) \le \delta_k\} \mid E\right) \ge \left(1 - \frac{\varepsilon}{N}\right)^N \ge 1 - \varepsilon.$$

Combining these results, we have

$$\begin{aligned}\mathcal{W}_{I_{\{x',y':\rho(x',y')>r\}}}\big(\mathcal{A}^{(T+N)}((\gamma, x), \cdot), B^{(T,N)}((\gamma, x), \cdot)\big) &\le \zeta(\{\rho(x_{T+N}, y_{T+N}) > r\}) \\ &\le \zeta(\{\rho(x_{T+N}, y_{T+N}) > \delta_N\} \mid E) + \mathbb{P}(E) \\ &\le 2\varepsilon.\end{aligned}$$

Since this holds for any r, ε, $\mathcal{W}_{\rho\wedge 1}\big(\mathcal{A}^{(T+N)}((\gamma, x), \cdot), B^{(T,N)}((\gamma, x), \cdot)\big) \le r + 2\varepsilon$ and the conclusion follows. □

Combining these lemmas, we may now prove Theorem 1.

Proof of Theorem 1. Fix $\varepsilon \in (0, 1)$. From Lemma 1, we can choose N_ε sufficiently large that, for all $n \ge N_\varepsilon$ with a particular adaptation stopping time $T_n = n - N_\varepsilon \ge 0$,

$$\mathcal{W}_{\rho\wedge 1}\big[B^{(T_n,N_\varepsilon)}((\gamma, x), \cdot), \pi\big] \le \varepsilon/2.$$

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

Given this N_ε and using Lemma 2, we can choose n_ε sufficiently large that, for all $n \geq n_\varepsilon$, $\mathcal{W}_{\rho\wedge 1}\big[\mathcal{A}^{(T_n+N_\varepsilon)}((\gamma, x), \cdot), B^{(T_n,N_\varepsilon)}((\gamma, x), \cdot)\big] \leq \varepsilon/2$. The triangle inequality holds by [40, Lemma 7.6], so that

$$\begin{aligned}\mathcal{W}_{\rho\wedge 1}\big[\mathcal{A}^{(n)}((\gamma, x), \cdot), \pi\big] &\leq \mathcal{W}_{\rho\wedge 1}\big[\mathcal{A}^{(T_n+N_\varepsilon)}((\gamma, x), \cdot), B^{(T_n,N_\varepsilon)}((\gamma, x), \cdot)\big] \\ &\quad + \mathcal{W}_{\rho\wedge 1}\big[B^{(T_n,N_\varepsilon)}((\gamma, x), \cdot), \pi\big] \leq \varepsilon.\end{aligned}$$

Since the conditional optimal coupling is attained and is Borel measurable [41, Theorem 4.8], we have, by dominated convergence,

$$\lim_{t\to\infty} \mathcal{W}_{\rho\wedge 1}[\mu\mathcal{A}^{(n)}, \pi] \leq \lim_{t\to\infty} \int_{\mathcal{Y}\times\mathcal{X}} \mathcal{W}_{\rho\wedge 1}[\mathcal{A}^{(n)}((\gamma, x), \cdot), \pi]\, \mu(\mathrm{d}\gamma, \mathrm{d}x) = 0. \qquad \square$$

Interestingly, we do not assume that π is invariant. Denote the distance to a closed set $C \subseteq \mathcal{X}$ by $\rho(x, C) = \inf_{y\in C} \rho(x, y)$ and the ε-inflation of the set by $C^\varepsilon = \{x \in X\colon \rho(x, C) \leq \varepsilon\}$. Theorem 1 and Strassen's theorem [37] ensures, uniformly for any closed Borel measurable set $C \subseteq \mathcal{X}$, that $\mu\mathcal{A}^{(t)}(C) \to \pi(C^\varepsilon)$. Theorem 1 also ensures that, for every bounded ρ-Lipschitz Borel measurable function $\varphi\colon \mathcal{X} \to \mathbb{R}$, $\int_{\mathcal{X}} \varphi\, \mathrm{d}\mu\mathcal{A}^{(t)} \to \int_{\mathcal{X}} \varphi\, \mathrm{d}\pi$.

The following extends Theorem 1 to L^p-Wasserstein distances with unbounded metrics.

Proposition 3. *Suppose an adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ with initialization probability measure μ (1) satisfies weak containment (5) and weak diminishing adaptation (6). Suppose further that, for some $x_0 \in \mathcal{X}$ and $p \in \mathbb{Z}_+$, $\int \rho(x, x_0)^p\, \pi(\mathrm{d}x) < \infty$. Then the following are equivalent:*

(i) Convergence in the L^p-Wasserstein distance holds: $\lim_{t\to\infty} \mathcal{W}_{\rho,p}(\mu\mathcal{A}^{(t)}, \pi) = 0$.

(ii) The sequence $(\rho(X_t, x_0)^p)_{t\geq 0}$ is uniformly integrable:

$$\lim_{R\to\infty} \sup_{t\geq 0} \int_{\rho(x,x_0)>R} \rho(x, x_0)^p \mu\mathcal{A}^{(t)}(\mathrm{d}x) = 0.$$

(iii) If $(\mathcal{X}, \rho)$ is a complete separable metric space, then the following are also equivalent to (i):

(iv) $\lim_{t\to\infty} \int_{\mathcal{X}} \rho(x, x_0)^p \mu\mathcal{A}^{(t)}(\mathrm{d}x) = \int_{\mathcal{X}} \rho(x, x_0)^p\, \pi(\mathrm{d}x)$.

(v) $\limsup_{t\to\infty} \int_{\mathcal{X}} \rho(x, x_0)^p \mu\mathcal{A}^{(t)}(\mathrm{d}x) \leq \int_{\mathcal{X}} \rho(x, x_0)^p\, \pi(\mathrm{d}x)$.

Proof. By Theorem 1 and Markov's inequality, $\inf_{\xi\in\mathcal{C}(\mu\mathcal{A}^{(t)},\pi)} \xi(\{\rho(u, v) > \varepsilon\}) \to 0$ for any $\varepsilon > 0$. Let $\xi^{(t)}$ be the attained optimal coupling for each t [41, Theorem 4.1]. $\square$

Assume (i) holds. For any $\varepsilon \in (0, 1)$, $\lim_{t\to\infty} \xi^{(t)}(|\rho(x, x_0) - \rho(y, x_0)| \geq \varepsilon) = 0$. By Young's inequality, for any $\varepsilon \in (0, 1)$ there is a constant $C_{\varepsilon,p}$ depending on p, ε such that

$$\rho(x, x_0)^p \leq (1 + \varepsilon)\rho(y, x_0)^p + C_{\varepsilon,p}\rho(x, y)^p.$$

Integrating with the coupling implies that

$$\limsup_{t\to\infty} \int_{\mathcal{X}} \rho(x, x_0)^p \mu\mathcal{A}^{(t)}(\mathrm{d}x) \leq \int_{\mathcal{X}} \rho(y, x_0)^p\, \pi(\mathrm{d}y).$$

By [22, Theorem 5.11], (ii) holds.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

Now assume (ii) holds. By convexity,

$$\limsup_t \int_{\rho(x,y)>R} \rho(x,y)^p \, \xi^{(t)}(\mathrm{d}x, \mathrm{d}y) \le 2^{p-1} \limsup_t \int_{\rho(x,y)>R} \rho(x,x_0)^p \, \mu\mathcal{A}^{(t)}(\mathrm{d}x) + 2^{p-1} \limsup_t \int_{\rho(x,y)>R} \rho(y,x_0)^p \, \xi^{(t)}(\mathrm{d}x, \mathrm{d}y).$$

By the characterization of uniform integrability [22, Theorem 5.11] and dominated convergence, this bound tends to 0 as $R \to 0$. This implies that

$$\lim_{t\to\infty} \mathcal{W}^p_{\rho,p}\big[\mu\mathcal{A}^{(t)}, \pi\big] = \lim_{t\to\infty} \int_{\mathcal{X}^2} \rho(x_t, y)^p \, \xi^{(t)}(\mathrm{d}x', \mathrm{d}y') = 0.$$

The remaining equivalences follow from [22, Lemma 5.11]. □

Proposition 3 has many interesting applications to extend weak convergence of adaptive MCMC, and also extending convergence in total variation of adaptive MCMC. For example, if $\mathcal{X} = \mathbb{R}^d$, then weak convergence from Theorem 1 can be used to extend the convergence to the L^2 Wasserstein distance $\mathcal{W}_{\|\cdot\|,2}$. Another possibility is to extend traditional convergence of adaptive MCMC to stronger convergence [32, Theorem 13] in the case when (strong) containment and (strong) diminishing adaptation hold. The following corollary extends convergence in total variation to a stronger Wasserstein distance under similar conditions [32, Theorem 18].

Corollary 1. *Suppose an adaptive process $(\Gamma_t, X_t)_{t\ge0}$ with initialization probability measure μ as in (1) satisfies (strong) containment (3) and (strong) diminishing adaptation (2). Suppose there is a lower semicontinuous function $V\colon \mathcal{X} \to [0,\infty)$ and constants $\lambda \in (0,1)$ and $L \in (0,\infty)$ such that, for all γ, $x \in \mathcal{Y}\times\mathcal{X}$, $(\mathcal{P}_\gamma V)(x) \le \lambda V(x) + L$. If $\int_{\mathcal{X}} V \, \mathrm{d}\mu < \infty$, then $\lim_{t\to\infty} \mathcal{W}_{\bar\rho}\big(\mu\mathcal{A}^{(t)}, \pi\big) = 0$, where $\bar\rho(x,y) = [(1 + V(x) + V(y))]^{1/2}$ if $x \ne y$ and 0 otherwise.*

Proof. The drift condition and assumption on μ imply that $(\sqrt{V(X_t)})_{t\ge0}$ is uniformly integrable, and Proposition 3 implies the conclusion. □

Remark 1. An alternative way to extend weak convergence to a stronger convergence in total variation convergence is through addition of an independent random variable [6]. Consider $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{Z}_+$, and an adaptive process $(\Gamma_t, X_t)_{t\ge0}$ with initialization probability measure μ (1) that satisfies weak containment (5) and weak diminishing adaptation (6), both with metric $\rho(\cdot,\cdot) = \|\cdot - \cdot\|$. Let $h \in (0,1)$ and let σ_h be a Gaussian distribution on $\mathbb{R}^d$, $N(0, h\,I)$. Then $\lim_{t\to\infty} \mathcal{W}_{\mathrm{TV}}\big(\mu\mathcal{A}^{(t)} * \sigma_h, \pi * \sigma_h\big) = 0$, where $*$ denotes convolution.

The following is a useful coupling technique to show weak containment (5).

Lemma 3. *Let $(\Gamma_t, X_t)_{t\ge0}$ be an adaptive process with initialization probability measure μ as in (1). Suppose $\pi\mathcal{P}_\gamma = \pi$ for every $\gamma \in \mathcal{Y}$. Assume, for some $x_0 \in \mathcal{X}$ and some $p \in \mathbb{Z}_+$, that $L = \int_{\mathcal{X}} \rho(x,x_0)^p \, \pi(\mathrm{d}x) < \infty$ and, for every γ, $x \in \mathcal{Y}\times\mathcal{X}$, $\int_{\mathcal{X}} \rho(y,x_0)^p \, \mathcal{P}_\gamma(x, \mathrm{d}y) < \infty$. Suppose there is an $\alpha \in (0,1)$ such that, for all $x, y \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$,*

$$\mathcal{W}_{\rho,p}(\mathcal{P}_\gamma(x,\cdot), \mathcal{P}_\gamma(y,\cdot)) \le (1-\alpha)\rho(x,y).$$

Then, for every $t \in \mathbb{Z}_+$ and $x \in \mathcal{X}$, $\mathcal{W}_{\rho,p}\big(\mathcal{P}^t_\gamma(x,\cdot), \pi\big) \le (1-\alpha)^t[\rho(x,x_0) + L]$. Further, $\sup_{t\ge0} \mathbb{E}[\rho(X_t,x_0)^p] < \infty$ and $(\rho(X_t,x_0))_{t\ge0}$ is bounded in probability.

Proof. For each $t \in Z_+$ and each $x, y \in \mathcal{X}$, $\mathcal{W}_{\rho,p}\big(\mathcal{P}^t_\gamma(x,\cdot), \mathcal{P}^t_\gamma(y,\cdot)\big)^p \le (1-\alpha)^{tp}\rho(x,y)^p$. The optimal coupling is attained [41, Theorem 4.1] at some conditional coupling $\xi_{x,y}$ and is

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

Borel measurable [41, Corollary 5.22]. Since $\int_{\mathcal{X}} \xi_{x,y}(\cdot,\cdot)\,\pi(\mathrm{d}y) \in \mathcal{C}[\mathcal{P}_\gamma^t, \pi]$,

$$\mathcal{W}_{\rho,p}(\mathcal{P}_\gamma^t(x,\cdot),\pi)^p \le \int_{\mathcal{X}}\int_{\mathcal{X}^2} \rho(x',y')^p\,\xi_{x,y}(\mathrm{d}x',\mathrm{d}y')\,\pi(\mathrm{d}y) \le (1-\alpha)^{tp}\int_{\mathcal{X}} \rho(x,y)^p\,\pi(\mathrm{d}y).$$

By Young's inequality, for any $\varepsilon > 0$ there is a constant $C_\varepsilon > 0$ such that, for any $a, b \ge 0$, $(a+b)^p \le (1+\varepsilon)a^p + C_\varepsilon b^p$. For any $x_0 \in \mathcal{X}$, we can choose ε sufficiently small that $(1+\varepsilon)(1-\alpha)^p < 1$, and a constant $C_{\varepsilon,p}$ such that

$$\int_{\mathcal{X}} \rho(x',x_0)^p\,\mathcal{P}_\gamma(x,\mathrm{d}x') \le [(1-\alpha)\rho(x,x_0) + 2L^{1/p}]^p \le (1+\varepsilon)(1-\alpha)^p\rho(x,x_0)^p + C_{\varepsilon,p}L.$$

By [32, Lemma 15], this simultaneous geometric drift condition implies that $(\rho(X_t,x_0))_t$ is bounded in probability. □

6. A weak law of large numbers

The point of this section is to develop convergence in probability of the empirical average of the adaptive MCMC process or weak law of large numbers. The convergence theory developed so far in Wasserstein distances provides estimation accuracy for the marginal distribution of X_t but this generally has a large variability. Estimation from the entire adaptive process $X_s \sim \mu\mathcal{A}^{(s)}$ for $s \le t$ requires theory for the empirical average. It is then of interest for reliable estimation to develop conditions such that, for bounded ρ-Lipschitz functions, $(1/t)\sum_{s=1}^t \varphi(X_s) \to \int_{\mathcal{X}} \varphi\,\mathrm{d}\pi$ in probability.

The law of large numbers for non-adapted Markov chains is well studied under convergence in total variation. On the other hand, convergence in Wasserstein distances and its connection to the law of large numbers is less understood [36, Theorem 1.2]. The first result is general and relies on the convergence of the adaptive process, but may even apply the law of large numbers to unbounded functions if the conditions are satisfied.

Theorem 2. *Let $(\Gamma_t, X_t)_{t\ge0}$ be an adaptive process with initialization probability measure μ such that $X_t \sim \mu\mathcal{A}^{(t)}$ (1). Let $d(\cdot,\cdot)$ be a lower semicontinuous metric and suppose, for some $x_0 \in \mathcal{X}$ and for each $t \in \mathbb{Z}_+$, $\int d(x,x_0)^2\mu\mathcal{A}^{(t)}(\mathrm{d}x) < \infty$ and $\int d(x,x_0)^2\,\pi(\mathrm{d}x) < \infty$. If*

$$\lim_{t\to\infty} \mathcal{W}_{d,2}(\mu\mathcal{A}^{(t)},\pi) = 0,$$

then for every Borel measurable $\varphi\colon \mathcal{X} \to \mathbb{R}$ with $\|\varphi\|_{\mathrm{Lip}(d)} < \infty$,

$$\lim_{t\to\infty} \mathbb{E}\left[\left(\frac{1}{t}\sum_{s=1}^t \varphi(X_s) - \int_{\mathcal{X}} \varphi\,\mathrm{d}\pi\right)^2\right] = 0. \tag{8}$$

In particular, the weak law of large numbers holds, that is, $(1/t)\sum_{s=1}^t \varphi(X_s) \to \int_{\mathcal{X}} \varphi\,\mathrm{d}\pi$ in probability.

Proof. We can assume that $\|\varphi\|_{\mathrm{Lip}(d)} \le 1$ since we may normalize φ. We may also assume that $\int \varphi\,\mathrm{d}\pi = 0$ since $\psi = \varphi - \int \varphi\,\mathrm{d}\pi$ is also d-Lipschitz. We can assume there is an $x_0 \in \mathcal{X}$ such that $\varphi(x_0) = 0$. Let Γ be a coupling of $X_t \sim \mu\mathcal{A}^{(t)}$ and $Y \sim \pi$. By disintegration [22, Theorem 3.4], there is a Borel measurable conditional probability measure $\Gamma_{h_s}(\mathrm{d}x_t,\mathrm{d}y)$ with $h_s = (\gamma_1, x_1, \ldots, \gamma_s, x_s)$ such that $\Gamma(\mathrm{d}x_t,\mathrm{d}y) = \int_{\mathcal{X}^s} \Gamma_{h_s}(\mathrm{d}x_t,\mathrm{d}y)\mu\mathcal{A}^{(1,\ldots,s)}(\mathrm{d}h_s)$. With the

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

history $H_s = (\Gamma_k, X_k)_{k=1}^s$ and since φ is d-Lipschitz, for $t \geq s$,

$$\left| \mathbb{E}[\varphi(X_t) \mid \mathcal{H}_s] - \int_{\mathcal{X}} \varphi \, \mathrm{d}\pi \right| \leq \int_{\mathcal{X}^2} d(x_t, y) \, \Gamma_{H_s}(\mathrm{d}x_t, \mathrm{d}y).$$

For $T \in \mathbb{Z}_+$, we have the upper bound

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{t=1}^T \varphi(X_t) \right)^2 \right] &= \sum_{t=1}^T \sum_{s=1}^T \mathbb{E}[\varphi(X_t)\varphi(X_s)] \\
&= \sum_{t=1}^T \mathbb{E}[\varphi(X_t)^2] + 2 \sum_{t=2}^T \sum_{s=1}^{t-1} \mathbb{E}[\mathbb{E}[\varphi(X_t) \mid \mathcal{H}_s]\varphi(X_s)] \\
&\leq T \sup_{t \geq 0} \int_{\mathcal{X}} d(x, x_0)^2 \mu \mathcal{A}^{(t)}(\mathrm{d}x) \\
&\quad + 2 \sum_{t=2}^T \sum_{s=1}^{t-1} \mathbb{E}\left[\int_{\mathcal{X}^2} d(x_t, y) \, \Gamma_{H_s}(\mathrm{d}x_t, \mathrm{d}y) d(X_s, x_0) \right].
\end{aligned}$$

Using Cauchy–Schwarz and Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathcal{X}^2} d(x_t, y) \, \Gamma_{H_s}(\mathrm{d}x_t, \mathrm{d}y) d(X_s, x_0) \right] \\
&\quad \leq \sqrt{\mathbb{E}\left[\left(\int_{\mathcal{X}^2} d(x_t, y) \, \Gamma_{H_s}(\mathrm{d}x_t, \mathrm{d}y) \right)^2 \right]} \sqrt{\mathbb{E}[d(X_s, x_0)^2]} \\
&\quad \leq \sqrt{\int_{\mathcal{X} \times \mathcal{X}} d(x_t, y)^2 \, \Gamma(\mathrm{d}x_t, \mathrm{d}y)} \sqrt{\sup_{t \geq 0} \int_{\mathcal{X}} d(x, x_0)^2 \mu \mathcal{A}^{(t)}(\mathrm{d}x)}.
\end{aligned}$$

By assumption, we can choose a T_ε depending on ε such that, for all $t \geq T_\varepsilon$, $\mathcal{W}_{d,2}(\mu \mathcal{A}^{(t)}, \pi) \leq \varepsilon$. By assumption, $\max_{0 \leq t \leq T_\varepsilon} \mathbb{E}[d(X_t, x_0)^2] < \infty$ and it follows by the triangle inequality [40, Lemma 7.6] that there is an $R \in (0, \infty)$ such that

$$\sup_{t \geq 0} \mathcal{W}_{d,2}(\mu \mathcal{A}^{(t)}, \pi) \leq \sqrt{\sup_{t \geq 0} \mathbb{E}(d(X_t, x_0)^2)} + \sqrt{\int d(x, x_0)^2 \, \pi(\mathrm{d}x)} \leq R.$$

Since the coupling Γ is arbitrary, we have the upper bound for every $T \geq T_\varepsilon + 1$:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{T} \sum_{t=1}^T \varphi(X_t) \right)^2 \right] &\leq \frac{R^2}{T} + \frac{2R}{T^2} \sum_{t=2}^T (t-1) \mathcal{W}_{d,2}(\mu \mathcal{A}^{(t)}, \pi) \\
&\leq \frac{R^2}{T} + \frac{2R^2}{T^2} \sum_{t=2}^{T_\varepsilon} (t-1) + \frac{2R}{T^2} \sum_{t=T_\varepsilon+1}^T (t-1)\varepsilon \\
&\leq \frac{R^2}{T} + \frac{R^2 T_\varepsilon (T_\varepsilon - 1)}{T^2} + \frac{R\varepsilon (T - T_\varepsilon)(T + T_\varepsilon - 1)}{T^2}.
\end{aligned}$$

The conclusion follows since we can choose T sufficiently large and ε sufficiently small. □

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

Theorem 2 also provides general conditions for weakly converging Markov chains [36, Theorem 1.2]. We may now show that the conditions of Theorem 1 are sufficient for a weak law of large numbers for bounded $\rho-$Lipschitz functions.

Corollary 2. *Suppose an adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ with initialization probability measure μ (1) satisfies weak containment (5) and weak diminishing adaptation (6). Then, for every bounded Borel measurable $\varphi\colon \mathcal{X} \to \mathbb{R}$ with $\|\varphi\|_{\mathrm{Lip}(\rho)} < \infty$ and any $p \in \mathbb{Z}_+$,*

$$\lim_{t\to\infty} \mathbb{E}\left[\left|\frac{1}{t}\sum_{s=1}^{t} \varphi(X_s) - \int_{\mathcal{X}} \varphi \,\mathrm{d}\pi\right|^p\right] = 0. \tag{9}$$

If, in addition, for some $x_0 \in \mathcal{X}$, $(\rho(X_t, x_0)^2)_{t\geq 0}$ is uniformly integrable and also $\int \rho(x, x_0)\,\pi(\mathrm{d}x) < \infty$, then (9) holds with $p = 2$ and for all $\|\varphi\|_{\mathrm{Lip}(\rho)} < \infty$.

Proof. For Borel measurable $\varphi\colon \mathcal{X} \to \mathbb{R}$ with $\|\varphi\|_{\mathrm{Lip}(\rho)} < \infty$, apply Theorems 1 and 2 and it follows that (8) holds. Since φ is bounded and (8) holds, L^p convergence follows for $p \in \mathbb{Z}_+$. To remove the bounded assumption on φ, since it is assumed that $(\rho(X_t, x_0)^2)_{t\geq 0}$ is uniformly integrable, $\mathbb{E}(\rho(X_t, x_0)^2) < \infty$ for each $t \in \mathbb{Z}_+$ and $\int \rho(x, x_0)\,\pi(\mathrm{d}x) < \infty$ then we can apply Proposition 3. □

7. Examples and applications

Let us now revisit constructing adaptive processes for the running examples of Markov chains (Example 1) and Example 2, where (strong) containment (3) fails to hold.

7.1. Example: Discrete adaptive autoregressive process

Consider an adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ using the Markov kernels $(\mathcal{P}_\gamma)_\gamma$ for the discrete autoregressive process defined in Example 1, which fails to satisfy (strong) containment (3). Assume the adaptation satisfies $|\Gamma_{t+1} - \Gamma_t| \to 0$ in probability as $t \to \infty$. For any $\gamma \in \mathbb{Z}_+$ with $\gamma \geq 2$ and $x \in [0, 1)$, we showed previously in Example 1 that, for any $t \in \mathbb{Z}_+$,

$$\mathcal{W}_{|\cdot|}\big(\mathcal{P}_\gamma^t(x, \cdot), \mathrm{Unif}(0, 1)\big) \leq 2^{-t},$$

where Unif([0, 1) is Lebesgue measure on [0, 1) and so weak containment holds (5). For every $x, y \in [0, 1)$ define $X_1^{\Gamma_t} = x/\,\Gamma_t + \xi_1^{\Gamma_t}$ and $Y_1^{\Gamma_t} = y/\,\Gamma_t + \xi_1^{\Gamma_t}$ with common discrete random variable $\xi_1^{\Gamma_t}$ defined previously in Example 1. The random variables $\big(X_1^{\Gamma_{t+1}}, Y_1^{\Gamma_t}\big)$ define a coupling and since, for sufficiently large t, $\Gamma_{t+1} = \Gamma_t$ with high probability, then

$$\sup_{|x-y|\leq\delta} \mathcal{W}_{|\cdot|}(\mathcal{P}_{\Gamma_{t+1}}(x, \cdot), \mathcal{P}_{\Gamma_t}(y, \cdot)) \leq \sup_{|x-y|\leq\delta} \mathbb{E}\big[\big|X_1^{\Gamma_{t+1}} - Y_1^{\Gamma_t}\big| \mid X_0 = x, Y_0 = y\big] \leq \frac{\delta}{2}$$

holds with high probability. Then this bound tends to 0 as $\delta \to 0$ and we conclude that weak diminishing adaptation (6) holds. By Theorem 1, for every $\gamma \geq 2$ and $x \in [0, 1)$,

$$\lim_{t\to\infty} \mathcal{W}_{|\cdot|}[\mathcal{A}^{(t)}((\gamma, x), \cdot), \mathrm{Unif}(0, 1)] = 0$$

and this discrete autoregressive adaptive process converges weakly. Corollary 2 then implies a weak law of large numbers for all bounded Lipschitz continuous functions.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

7.2. Example: Infinite-dimensional adaptive autoregressive process

Consider an adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ using Markov kernels $(\mathcal{P}_\gamma)_\gamma$ for the infinite-dimensional autoregressive process (2), which cannot satisfy (strong) containment (3). Assume the adaptation is restricted to a bounded set, that is, for some $R \in (0, \infty)$, if $\|X_t\| > R$ then $\Gamma_{t+1} = \Gamma_t$ and $|\Gamma_t - \Gamma_{t+1}| \to 0$ in probability as $t \to \infty$.

We showed previously (2) that for any $\gamma \in (0, \gamma^*)$ and any $x, y \in H$,

$$\mathcal{W}_{\|\cdot\|,2}(\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(y, \cdot)) \leq \gamma^* \|x - y\| ;$$

combined with Lemma 3. this implies that weak containment (5) holds. For $x, y \in H$, let $Y_t = \Gamma_{t+1}x + \sqrt{1 - {\Gamma_{t+1}}^2}\xi_t$ and $Y'_t = \Gamma_t y + \sqrt{1 - {\Gamma_t}^2}\xi_t$ with common independent random variable $\xi_t \sim \mathcal{N}(0, C)$. We have the upper bound, for any $t \in \mathbb{Z}_+$,

$$\begin{aligned}\mathcal{W}_{\|\cdot\|,2}(\mathcal{P}_{\Gamma_{t+1}}(x, \cdot), \mathcal{P}_{\Gamma_t}(y, \cdot)) &\leq [\mathbb{E}(\left\|Y_t - \Gamma_t x + \Gamma_t x - Y'_t\right\|^2)]^{1/2} \\ &\leq |\Gamma_{t+1} - \Gamma_t| \, \|x\| + \gamma^* \|x - y\| + \left|\sqrt{1-\Gamma_{t+1}^2} - \sqrt{1-\Gamma_t^2}\right| \sqrt{\mathrm{tr}(C)}.\end{aligned}$$

If $\|x\| > R$ and $\|x - y\| \leq \delta$, then $\mathcal{W}_{\|\cdot\|,2}(\mathcal{P}_{\Gamma_{t+1}}(x, \cdot), \mathcal{P}_{\Gamma_t}(y, \cdot)) \leq \gamma^*\delta$. Otherwise, if $\|x\| \leq R$ and $\|x - y\| \leq \delta$,

$$\mathcal{W}_{\|\cdot\|,2}(\mathcal{P}_{\Gamma_{t+1}}(x, \cdot), \mathcal{P}_{\Gamma_t}(y, \cdot)) \leq |\Gamma_{t+1} - \Gamma_t|R + \gamma^*\delta + \left|\sqrt{1 - \Gamma_{t+1}^2} - \sqrt{1 - \Gamma_t^2}\right| \sqrt{\mathrm{tr}(C)}.$$

In either case, weak diminishing adaptation (6) holds.

For any $p \in \mathbb{Z}_+$, Young's inequality and Fernique's theorem [5, Theorem 2.8.5] imply we can choose $\varepsilon > 0$ such that $(1 + \varepsilon)\gamma^{*p} < 1$ and a constant $C_\varepsilon > 0$ depending on ε such that, for every $\gamma, x \in \mathcal{Y} \times \mathcal{X}$, $(\mathcal{P}_\gamma)(\|x\|^p) \leq (1 + \varepsilon)\gamma^{*p} \|x\|^p + C_\varepsilon$. This implies that $(\|X_t\|^p)_{t\geq 0}$ is uniformly integrable. From Theorem 1 and Proposition 3, for every $\gamma, x \in (0, \gamma^*] \times H$ and every $p \in \mathbb{Z}_+$, $\lim_{t\to\infty} \mathcal{W}_{\|\cdot\|,p}[\mathcal{A}^{(t)}((\gamma, x), \cdot), \mathcal{N}(0, C)] = 0$ and Corollary 2 implies a weak law of large numbers for all Lipschitz continuous functions.

7.3. Example: Adaptive random-walk Metropolis–Hastings

We look at a discrete version of the adaptive random-walk Metropolis–Hastings algorithm [18], which adapts the covariance of the proposal towards the covariance of the target probability measure. This concrete example illustrates an issue in practical applications since current computers only produce floating-point approximations to real numbers. As a result, convergence theory in total variation corresponding to an adaptive Markov chain Monte Carlo simulation targeting a continuous target probability measure is infeasible and has other issues that have been studied previously through perturbation theory [7, 34]. This is not necessarily the case in alternative distances which metrize weak convergence.

Let π be a target Borel probability measure on $\mathbb{R}^d$ with $d \in \mathbb{Z}_+$ and Lebesgue density f. Let $D = (x_k)_{k=1}^\infty \subset \mathbb{R}^d$ be a dense subset in $\mathbb{R}^d$, Σ be a symmetric, positive-definite matrix, $\mu \in \mathbb{R}^d$, and let $\mathcal{N}_D(\mu, \Sigma)$ denote the discrete Gaussian with probability mass function

$$g_\Sigma(\mu, x) = \frac{\exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)}{\sum_{j=1}^\infty \exp\left(-\frac{1}{2}(x_j - \mu)^\top \Sigma^{-1}(x_j - \mu)\right)}.$$

When $\mu \in D$, $g_\Sigma(\mu, x) = g_\Sigma(x, \mu)$ and it is symmetric. Let $(M(\gamma))_{\gamma\in\mathcal{Y}}$ be a family of symmetric, positive-definite matrices on $\mathbb{R}^d$. We define a discrete Markov chain $\left(X_t^\gamma\right)_{t\geq 0}$ using a

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

discrete random-walk Metropolis–Hastings kernel $\mathcal{P}_\Sigma$ with discrete Gaussian proposal, where the proposal X' given the previous state x is $X' \sim \mathcal{N}_D(x, M(\gamma))$. The Markov kernel is defined for each $x_l, x_k \in D$ by

$$\mathcal{P}_\gamma(x_l, x_k) = \left[1 \wedge \frac{f(x_k)}{f(x_l)}\right] g_{M(\gamma)}(x_l x_k) + \delta_{x_l}(\{x_k\}) \left(1 - \sum_{j=1}^{\infty} \left[1 \wedge \frac{f(x_j)}{f(x_l)}\right] g_{M(\gamma)}(x_l, x_j)\right).$$

We will assume that f is continuous with compact support $K \subset \mathbb{R}^d$. Further, we will assume that the set of Σ is compact and so the eigenvalues of Σ are uniformly bounded, so there are constants $\lambda_*, \lambda^* \in (0, \infty)$ such that $\lambda_* \le \lambda_i(\Sigma) \le \lambda^*$ for all $i = 1, \ldots, d$. It follows by a minorization argument over K that there is an $\alpha \in (0, 1)$ such that, for any $\gamma \in \mathcal{Y}$, any $x_i, x_j \in D$, and any bounded Lipschitz continuous function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$, $|\mathcal{P}_\gamma^t \varphi(x_i) - \mathcal{P}_\gamma^t \varphi(x_j)| \le (1-\alpha)^t$. By the density of D and continuity of $\mathcal{P}\varphi(\cdot)$, $\left|\mathcal{P}_\gamma^t \varphi(x_i) - \int_K \varphi \, d\pi\right| \le (1-\alpha)^t$. Weak containment (5) holds since it follows by Kantorovich–Rubinstein duality [40, Theorem 1.14] that $\mathcal{W}_{\|\cdot\| \wedge 1}\left(\mathcal{P}_\gamma^t(x_i, \cdot), \pi\right) \le (1-\alpha)^t$.

Now let $\gamma, x \in \mathcal{Y} \times D$ and let $(\Gamma_t, X_t)_{t \ge 0}$ where $X_t \sim \mathcal{A}^{(t)}((\gamma, x), \cdot)$ be the adaptive process using these Metropolis–Hastings kernels. Under any valid weak diminishing adaptation strategy satisfying (6), we have $\lim_{t\to\infty} \mathcal{W}_{\|\cdot\| \wedge 1}(\mathcal{A}^{(t)}((\gamma, x), \cdot), \pi) = 0$. Corollary 2 then implies a weak law of large numbers for bounded Lipschitz continuous functions. On the other hand, it can be shown that $\mathcal{W}_{\mathrm{TV}}(\mathcal{A}^{(t)}((\gamma, x), \cdot), \pi) = 1$ and fails to converge in total variation under any adaptation plan.

7.4. Example: Adaptive unadjusted Langevin process

In certain cases, it has been observed [41, p. 21] that Wasserstein distances can be simpler to prove convergence results than total variation; the following example provides a concrete illustration. Consider the Euclidean space $\mathbb{R}^d$ where $d \in \mathbb{Z}_+$ with Euclidean norm $\|\cdot\|$. Let the potential $V\colon \mathbb{R}^d \to \mathbb{R}$ have gradient $\nabla V(\cdot)$ with constants $\alpha, \beta > 0$ such that, for every $x, y \in \mathbb{R}^d$,

$$\|\nabla V(y) - \nabla V(x)\| \le \beta \|y - x\|, \tag{10}$$

$$V(y) \ge V(x) + \langle \nabla V(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2. \tag{11}$$

Let $(\mathcal{Y}, d)$ be a complete separable metric space and, for each $\gamma \in \mathcal{Y}$, let $M(\gamma)$ define a symmetric positive-definite matrix. Let $h \in (0, 1)$ be a fixed discretization size and consider the unadjusted Langevin process

$$X_{t+1}^{\gamma,h} = X_t^{\gamma,h} - hM(\gamma)\nabla V\left(M(\gamma)X_t^{\gamma,h}\right) + \sqrt{2h} Z_{t+1},$$

where $(Z_t)_{t \ge 0}$ are independent and identically distributed $\mathcal{N}(0, I_d)$. We can define a family of Markov kernels $(\mathcal{P}_{\gamma,h})_{\gamma,h}$ prescribing the conditional distributions $X_{t+1}^{\gamma,h} \mid X_t^{\gamma,h} = x \sim \mathcal{P}_{\gamma,h}(x, \cdot)$. For an adaptive strategy $(\Gamma_t, h_t)_{t \ge 0}$, we define the adaptive process $X_t := M(\Gamma_t) X_t^{\Gamma_t, h_t}$ for $t \ge 0$. For example, $M(\Gamma_t)$ can estimate the inverse covariance matrix using the entire history of the process as in adaptive Metropolis–Hastings [18] and adaptive piecewise-deterministic Markov processes [4]. We make the following assumption on the adaptation of the matrix $M(\gamma)$.

Assumption 2. *Assume $M(\cdot)$ is continuous and, for each $\gamma \in \mathcal{Y}$ and $x \in \mathbb{R}^d$, there is a constant $\lambda_* \in (0, \infty)$ such that, for every $v \in \mathbb{R}^d \setminus \{0\}$ with $\|v\| = 1$, $\lambda_* \le \langle vM(\gamma), v \rangle \le 1$. Assume $d(\Gamma_{t+1}, \Gamma_t) \to 0$ in probability and $\Gamma_{t+1} = \Gamma_t$ if $\|X_t\| > R$ for some $R > 0$.*

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

We have the following convergence result.

Proposition 4. *Assume the adaptation plan satisfies Assumption 2 and, for some $h_* \in (0, 1)$, let $H = [h_*, 1/(\alpha+\beta)]$ and assume $(h_t)_{t\geq 0}$ is a deterministic sequence with $h_t \in H$. Further, assume there is a limit $h^* \in H$ such that $\lim_{t\to\infty} |h_t - h^*| = 0$. Then the adaptive unadjusted Langevin process converges in the L^1-Wasserstein distance for every $\gamma, h, x \in \mathcal{Y} \times H \times \mathbb{R}^d$ to some probability measure π_{h^*}, that is, $\lim_{t\to\infty} \mathcal{W}_{\|\cdot\|}[\mathcal{A}^{(t)}((\gamma, h, x), \cdot), \pi_{h^*}] = 0$, and a weak law of large numbers holds for all bounded Lipschitz continuous functions.*

Proof. It can be shown that $M(\gamma)\nabla V(M(\gamma)\cdot)$ is β Lipschitz and α strongly convex. For $x, y \in \mathbb{R}^d$ and $\gamma, \gamma' \in \mathcal{Y}$, let

$$X_1^{\gamma} = x - hM(\gamma)\nabla V(M(\gamma)x) + \sqrt{2h}Z_1, \qquad Y_1^{\gamma'} = y - hM(\gamma')\nabla V(M(\gamma')y) + \sqrt{2h}Z_1$$

with shared Gaussian random variable $Z_1 \sim N(0, I)$. By [25, Theorem 2.1.12],

$$\begin{aligned}\mathbb{E}\left[\left\|X_1^{\gamma,h} - Y_1^{\gamma,h}\right\|^2\right] &\leq \left(1 - \frac{2h\alpha\beta}{\alpha+\beta}\right)\|x-y\|^2 + h\left(h - \frac{1}{\alpha+\beta}\right)\|\nabla V(x) - \nabla V(y)\|^2 \\ &\leq \left(1 - \frac{2h_*\alpha\beta}{\alpha+\beta}\right)\|x-y\|^2 .\end{aligned}$$

For each adapted discretization size h, h', $\left(\mathbb{E}\left\|M(\gamma)X_1^{\gamma,h'} - M(\gamma)X_1^{\gamma,h}\right\|^2\right)^{1/2} \leq |h'-h|\beta\,\|x\|$. Along with the assumed adaptation strategy, these imply that there exists an invariant measure π_{h^*} with finite second moment. By Lemma 3, weak containment (5) holds.

For each adapted discretization size h, h' and each adaptation parameter $\gamma, \gamma' \in \mathcal{Y}$,

$$\begin{aligned}&\left(\mathbb{E}\left\|M(\gamma')X_1^{\gamma',h'} - M(\gamma)Y_1^{\gamma,h}\right\|^2\right)^{1/2} \\ &\leq \left(\mathbb{E}\left\|M(\gamma')X_1^{\gamma',h'} - M(\gamma')X_1^{\gamma',h}\right\|^2\right)^{1/2} + \left(\mathbb{E}\left\|M(\gamma')X_1^{\gamma',h} - M(\gamma)X_1^{\gamma,h}\right\|^2\right)^{1/2} \\ &\quad + \left(\mathbb{E}\left\|X_1^{\gamma,h} - Y_1^{\gamma,h}\right\|^2\right)^{1/2} \\ &\leq \beta|h'-h|\,\|x\| + \left\|M(\gamma')\nabla V(M(\gamma')x) - M(\gamma)\nabla V(M(\gamma)x)\right\| + \|\nabla V(M(\gamma)x) - \nabla V(M(\gamma)y)\| \\ &\leq \beta|h'-h|\,\|x\| + 2\beta\left\|M(\gamma') - M(\gamma)\right\|\|x\| + \left(1 - \frac{2h_*\alpha\beta}{\alpha+\beta}\right)\|x-y\| .\end{aligned}$$

This upper bound implies weak diminishing adaptation (6) under this adaptation strategy. Therefore, this adaptive process converges weakly by Theorem 1. Lemma 3 implies $(\|X_t\|)_{n\geq 0}$ is uniformly integrable and then, by Proposition 3, the Wasserstein convergence follows. Corollary 2 implies a weak law of large numbers for all bounded Lipschitz functions. □

7.5. Example: Adaptive diffusion process

Let the potential $V\colon \mathbb{R}^d \to \mathbb{R}$ satisfy (10). Let $(M(\gamma))_{\gamma\in\mathcal{Y}}$ be defined as in Section 7.4 and consider adapting a stochastic differential equation with $M(\gamma)$ defined for $t \in [0, 1]$ by

$$\mathrm{d}X_t^{\gamma} = -M(\gamma)\nabla V(M(\gamma)X_t^{\gamma})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

where $(W_t)_{t\geq 0}$ is standard Brownian motion in $\mathbb{R}^d$. Then, for any $\gamma \in \mathcal{Y}$ and $x \in \mathbb{R}^d$, there exists a strong solution $(X_t^{\gamma})_{t\geq 0}$ that is a Markov process with kernel $\tilde{\mathcal{P}}_{\gamma}^t$. Using the solution at $t = 1$, we can define a new Markov chain $X_n^{\gamma} \mid X_0 = x$ with Markov transition kernel $\tilde{P}_{\gamma}^1$. We

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

can define an adaptive process $(\Gamma_n, X_n)_{n\geq 0}$ with $X_n = M(\Gamma_n)X_n^{\Gamma_n}$ so that $X_n \mid \Gamma_n = \gamma, X_{n-1} = x$ has Markov transition $\mathcal{P}_\gamma(x, \cdot)$ with invariant measure $\pi(\mathrm{d}x) = Z^{-1}\exp(-V(x))\,\mathrm{d}x$, where $Z = \int \exp(-V(x))\,\mathrm{d}x$. This type of adaptive scheme using matrices has been successful in piecewise-deterministic Markov processes [4].

Proposition 5. *Assume the adaptation plan satisfies Assumption 2 and assume the potential $V\colon \mathbb{R}^d \to \mathbb{R}$ satisfies (10) and (11). Then, for every $\gamma, x \in \mathcal{Y}\times\mathbb{R}^d$, the adaptive diffusion process converges in the L^1-Wasserstein distance $\lim_{t\to\infty}\mathcal{W}_{\|\cdot\|}[\mathcal{A}^{(t)}((\gamma, x), \cdot), \pi] = 0$ and a weak law of large numbers holds for all bounded Lipschitz continuous functions.*

Proof. For $\gamma, \gamma' \in \mathcal{Y}$ and $x, y \in \mathbb{R}^d$, let $X_t^\gamma \mid X_0 = x$ and $Y_t^{\gamma'} \mid Y_0 = y, Y_0 = y$, share the same Brownian motion so that these random variables define a coupling. By strong convexity, for every $x, y \in \mathbb{R}^d$, $\langle \nabla V(x) - \nabla V(y), y - x\rangle \geq \alpha \|x - y\|^2$. Define $f_\gamma(t) = \left\|X_t^\gamma - Y_t^\gamma\right\|^2$; it follows that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}f_\gamma(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\|x - y - M(\gamma)\int_0^t \left[\nabla V(M(\gamma)X_s^\gamma) - \nabla V(M(\gamma)Y_s^\gamma)\right]\mathrm{d}s\right\|^2 \\ &= -2\left\langle \nabla V(M(\gamma)X_t^\gamma) - \nabla V(M(\gamma)Y_t^\gamma), M(\gamma)(X_t^\gamma - Y_t^\gamma)\right\rangle \leq -2\alpha f_\gamma(t).\end{aligned}$$

By Gronwall's inequality, $\sqrt{\mathbb{E}\left\|X_1^\gamma - Y_1^\gamma\right\|^2} \leq \exp(-\alpha)\|x - y\|$. We have the upper bound

$$\mathcal{W}_{\|\cdot\|,2}\left[\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(y, \cdot)\right] \leq \sqrt{\mathbb{E}\left\|M(\gamma)X_1^\gamma - M(\gamma)Y_1^\gamma\right\|^2} \leq \exp(-\alpha)\|x - y\|.$$

By [16, Proposition 1], $\int \|y\|^2\,\pi(\mathrm{d}y)$ is finite, and by Lemma 3, weak containment (5) holds. Lemma 3 then implies weak containment (5) and $(\|X_n\|)_{n\geq 0}$ is uniformly integrable. A similar argument as in Example 7.4 shows weak diminishing adaptation (6). By Proposition 3, the convergence in Wasserstein follows and Corollary 2 implies a weak law of large numbers for all bounded Lipschitz functions. □

8. Connections to geometric drift and coupling conditions

A general approach to satisfy the containment condition (5) is through a simultaneous version of the weak Harris theorem [20, Theorem 4.8]. Other similar convergence bounds for non-adapted Markov chains could be modified to simultaneous versions as well [15, 28].

Theorem 3. *(Simultaneous weak Harris theorem). Let $(\mathcal{P}_\gamma)_{\gamma\in\mathcal{Y}}$ be a family of Markov kernels on $\mathcal{X}$ with invariant probability measure π.*

- *(Simultaneous geometric drift.) Suppose there is a Borel drift function $V\colon \mathcal{Y}\times\mathcal{X} \to [0, \infty)$ and constants $\lambda \in (0, 1)$, $K \in (0, \infty)$ such that, for every $\gamma, x \in \mathcal{Y}\times\mathcal{X}$,*

$$\int_{\mathcal{X}} V(\gamma, x')\mathcal{P}_\gamma(x, \mathrm{d}x') \leq \lambda V(\gamma, x) + K.$$

- *(ρ-contracting.) Suppose there is a $\kappa \in (0, 1)$ such that, for every $x, y \in \mathcal{X}$ with $\rho(x, y) < 1$ and every $\gamma \in \mathcal{Y}$, $\mathcal{W}_{\rho\wedge 1}(\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(y, \cdot)) \leq (1 - \kappa)\rho(x, y)$.*

- *(ρ-small.) Suppose, for some constants $\alpha, \delta \in (0, 1)$ and every $\gamma \in \mathcal{Y}$,*

$$\sup_{x,y\in C_\gamma} \mathcal{W}_{\rho\wedge 1}(\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(y, \cdot)) \leq 1 - \alpha,$$

where $C_\gamma = \{x \in \mathcal{X}\colon V(\gamma, x) \leq (1 + \delta)2K/(1 - \lambda)\}$.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

Then there is an explicit $\alpha^* \in (0, 1)$ *depending on* α, κ, λ *such that, for every* $t \in \mathbb{Z}_+$ *and every* $\gamma, x \in \mathcal{Y} \times \mathcal{X}$, $\alpha^* \in (0, 1)$,

$$\mathcal{W}_{\rho_\gamma}(\mathcal{P}_\gamma^t(x, \cdot), \pi) \leq (1 - \alpha^*)^t \sqrt{(1 + \beta^* V(\gamma, x) + (\alpha \wedge \kappa)/[4(1 - \lambda))]},$$

where $\rho_\gamma(u, v) = \sqrt{(\rho(u, v) \wedge 1)[1 + \beta^* V(\gamma, u) + \beta^* V(\gamma, v)]}$ *and* $\beta^* = (\alpha \wedge \kappa)/(4K)$.

Proof. The argument is inspired by [20, Theorem 4.8]. Fix $\gamma \in \mathcal{Y}$. For $\beta > 0$, define $\rho_{\beta,\gamma}(x, y) = \sqrt{(\rho(x, y) \wedge 1)(1 + \beta(V(\gamma, x) + V(\gamma, y))}$. First, assume for $x, y \in \mathcal{X}$ that $\rho(x, y) \geq 1$ and $V(\gamma, x) + V(\gamma, y) > R$. Now, for $\delta > 0$, choose $R = (1 + \delta)2K/(1 - \lambda)$. Then, using the simultaneous drift condition,

$$\begin{aligned}
\mathcal{W}_{\rho_{\beta,\gamma}}(\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(x, \cdot))^2 &\leq 1 + \beta \mathcal{P}_\gamma V(\gamma, x) + \beta \mathcal{P}_\gamma V(\gamma, y) \\
&\leq 1 - \lambda + \lambda(1 + \beta V(\gamma, x) + \beta V(\gamma, y)) + \beta 2K \\
&\leq \left[(1 - \lambda)\frac{1 + \beta 2K/(1 - \lambda)}{1 + \beta R} + \lambda\right] \rho_{\beta,\gamma}(x, y)^2.
\end{aligned}$$

Now assume for $x, y \in \mathcal{X}$ that $\rho(x, y) \geq 1$ and $V(\gamma, x) + V(\gamma, y) \leq R$. Then, using that C_γ is ρ-small and the simultaneous drift condition, and choosing $\beta \leq \alpha/(4K)$,

$$\begin{aligned}
\mathcal{W}_{\rho_{\beta,\gamma}}(\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(x, \cdot))^2 &\leq (1 - \alpha)[1 + \beta \mathcal{P}_\gamma V(\gamma, x) + \beta \mathcal{P}_\gamma V(\gamma, y)] \\
&\leq (1 - \alpha)\rho(x, y) \wedge 1[1 + \lambda\beta(V(x) + V(y)) + \beta 2K] \\
&\leq (1 - \alpha/2)\rho_{\beta,\gamma}(x, y)^2.
\end{aligned}$$

Next, assume for $x, y \in \mathcal{X}$ that $\rho(x, y) < 1$. Then, using ρ-contracting and the simultaneous drift condition with $\beta \leq \kappa/(4K)$,

$$\begin{aligned}
\mathcal{W}_{\rho_{\beta,\gamma}}(\mathcal{P}_\gamma(x, \cdot), \mathcal{P}_\gamma(x, \cdot))^2 &\leq (1 - \kappa)\rho(x, y)[1 + \beta \mathcal{P}_\gamma V(\gamma, x) + \beta \mathcal{P}_\gamma V(\gamma, y)] \\
&\leq \rho(x, y) \wedge 1[1 - \kappa + \beta 2K + (1 - \kappa)\lambda\beta(V(\gamma, x) + V(\gamma, y))] \\
&\leq (1 - \kappa/2)\rho_{\beta,\gamma}(x, y)^2.
\end{aligned}$$

□

Theorem 3 can be seen as an extension of the simultaneous geometric drift and minorization conditions [32, Theorem 18]. We allow the drift function to depend on the adapted tuning parameter and the metric ρ is not restricted to the Hamming metric. For drift functions V constant in $\gamma \in \mathcal{Y}$ so that we can write $V\colon \mathcal{X} \to [0, \infty)$, we have the following result for the adaptive process if the conditions of Theorem 3 hold.

Theorem 4. *Suppose the adaptive process* $(\Gamma_t, X_t)_{t \geq 0}$ *with initialization probability measure* μ *as in (1) constructed with Markov kernels* $(\mathcal{P}_\gamma)_{\gamma \in \mathcal{Y}}$ *satisfies weak diminishing adaptation (6). Suppose the conditions of Theorem 3 are satisfied for* $(\mathcal{P}_\gamma)_{\gamma \in \mathcal{Y}}$ *with a drift function* V *constant in* $\gamma \in \mathcal{Y}$. *Then* $\lim_{t \to \infty} \mathcal{W}_{\rho \wedge 1}(\mu \mathcal{A}^{(t)}, \pi) = 0$.

Proof. Since $\rho \wedge 1$ is bounded, it will suffice to assume that the adaptive process $(\Gamma_t, X_t)_t$ is initialized at $\gamma_0, x_0 \in \mathcal{Y} \times \mathcal{X}$. The conclusion follows from Theorem 1 if we verify weak containment (5). The conditions of Theorem 3 imply that, in order to satisfy weak containment, it suffices to show that $(V(X_t))_t$ is bounded in probability. The geometric drift condition implies $(V(X_t))_{t \geq 0}$ is bounded in probability by [32, Lemma 15] since it is constant in $\gamma \in \mathcal{Y}$. □

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

To satisfy weak containment (5), Theorem 4 can be weakened to subgeometric rates of convergence. If there is a constant $M_0 > 0$, a Borel measurable function $V\colon \mathcal{X} \to [0, \infty)$ such that $(V(X_t))_{t\geq 0}$ is bounded in probability, and a rate function $r\colon \mathbb{Z}_+ \to [0, 1]$ with $\lim_{n\to\infty} r(n) = 0$ such that, for all $n, t \in \mathbb{Z}_+$, $\mathcal{W}_{\rho\wedge 1}\big(\mathcal{P}^n_{\Gamma_t}(X_t, \cdot), \pi\big) \leq M_0 r(n) V(X_t)$, then weak containment (5) holds. For example, a polynomial drift condition is sufficient for $(V(X_t))_{t\geq 0}$ to be bounded in probability [3] and existing subgeometric rates of convergence for non-adapted Markov chains in Wasserstein distances can be modified to simultaneous versions [8, 14].

In certain cases, it can be difficult to find a drift function that does not change with $\gamma \in \mathcal{Y}$ and alternative techniques can be used to show (strong) containment. One successful strategy here is to only apply adaptation on a compact or bounded set of the state space [11, Theorem 21]. We will say adaptation is restricted to a Borel set $S \subset \mathcal{X}$ for the adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ if, for all $t \in \mathbb{Z}_+$, $X_{t-1} \notin S$, $\Gamma_t = \Gamma_{t-1}$.

Our next result shows the benefit of Theorem 3 with drift functions depending on the adapted tuning parameter if adaptation is restricted to a set.

Proposition 6. *Suppose the adaptive process $(\Gamma_t, X_t)_{t\geq 0}$ with initialization probability measure μ as in (1) constructed with Markov kernels $(\mathcal{P}_\gamma)_{\gamma\in\mathcal{Y}}$ satisfies weak diminishing adaptation (6). Suppose the conditions of Theorem 3 are satisfied with Markov kernels $(\mathcal{P}_\gamma)_{\gamma\in\mathcal{Y}}$ and drift function $V(\cdot, \cdot)$. Additionally, assume $(\Gamma_t, X_t)_{t\geq 0}$ has adaptation restricted to the Borel set $S \subset \mathcal{X}$ and $\sup_{x\in S} \sup_{t\in\mathbb{Z}_+} \mathbb{E}[V(\Gamma_{t+1}, X_{t+1}) \mid \mathcal{H}_t, X_t = x] < \infty$. Then $\lim_{t\to\infty} \mathcal{W}_{\rho\wedge 1}(\mu\mathcal{A}^{(t)}, \pi) = 0$.*

Proof. It will suffice to assume the adaptive process $(\Gamma_t, X_t)_t$ is initialized at $\gamma_0, x_0 \in \mathcal{Y} \times \mathcal{X}$, and by Theorem 1 we may show that $(V(\Gamma_t, X_t))_t$ is bounded in probability. By the assumptions, we can assume that $\mathbb{E}[V(\Gamma_t, X_t) I_S(X_{t-1})] \leq K$. We have the upper bound

$$\begin{aligned}\mathbb{E}[V(\Gamma_t, X_t)] &\leq \mathbb{E}[V(\Gamma_t, X_t) I_S(X_{t-1})] + \mathbb{E}[V(\Gamma_t, X_t) I_{S^c}(X_{t-1})] \\ &\leq K + \mathbb{E}[V(\Gamma_{t-1}, X_t) I_{S^c}(X_{t-1})] \\ &\leq 2K + \lambda\mathbb{E}[V(\Gamma_{t-1}, X_{t-1})] \leq 2K/(1-\lambda) + V(\gamma_0, x_0).\end{aligned}$$

Therefore, $(V(\Gamma_t, X_t))_t$ is bounded in probability by Markov's inequality. □

9. Concluding remarks

This article developed weak convergence of adaptive MCMC processes under general conditions that is suited to situations where convergence in total variation is inadequate. One motivation is adapting the tuning parameters of reducible Markov chains where the traditional theory of adaptive MCMC may not be applied. Another application of the developed theory can be used to analyze adaptive MCMC processes where (strong) containment is difficult to show but weak containment may be more tractable. The weak law of large numbers developed here can be seen as an extension of [32, Theorem 23] and appears of practical interest for the reliability and stability of adaptive MCMC simulations widely used in statistics and machine learning.

There are many future research directions worthy of pursuit. While the developed theory for weak convergence allows a Markov process in continuous time to be adapted at discrete times, the proof techniques do not appear limited to discrete-time adaptive processes. In particular, a precise formulation of an adaptive MCMC process in continuous time with similar convergence results appears feasible. It also appears that some techniques for the convergence results developed here might be extended to develop quantitative convergence rates in Wasserstein

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

distances or mixing times for adaptive MCMC, but would require stronger assumptions on the adaptation plan. Another interesting direction is to try to generalize Theorem 1 to hold for general, possibly unbounded, Wassserstein distances by adapting weak containment and weak diminishing adaptation to hold using general Wasserstein distances.

Acknowledgements

The authors would like to thank the Associate Editor and the two anonymous referees for their insightful comments in helping to improve this article.

Funding information

This work was partially funded by NSERC Discovery Grant RGPIN-2019-04142, and by a postdoctoral fellowship at the University of Toronto.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] ANDRIEU, C. AND MOULINES, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Prob.* **16**, 1462–1505.
[2] ATCHADÉ, Y. F. AND ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**, 815–828.
[3] BAI, Y., ROBERTS, G. O. AND ROSENTHAL, J. S. (2009). On the containment condition for adaptive Markov chain Monte Carlo algorithms. Technical report, Centre for Research in Statistical Methodology, University of Warwick.
[4] BERTAZZI, A. AND BIERKENS, J. (2022). Adaptive schemes for piecewise deterministic Monte Carlo algorithms. *Bernoulli* **28**, 2404–2430.
[5] BOGACHEV, V. I. (1998). *Gaussian Measures*. American Mathematical Society, Providence, RI.
[6] BOGACHEV, V. I. (2018). *Weak Convergence of Measures*. American Mathematical Society, Providence, RI.
[7] BREYER, L., ROBERTS, G. O. AND ROSENTHAL, J. S. (2001). A note on geometric ergodicity and floating-point roundoff error. *Statist. Prob. Lett.* **53**, 123–127.
[8] BUTKOVSKY, O. (2014). Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *Ann. Appl. Prob.* **24**, 526–552.
[9] CHIMISOV, C., LATUSZYNSKI, K. AND ROBERTS, G. (2018). Air Markov chain Monte Carlo. Preprint, arXiv:1801.09309.
[10] COTTER, S. L., ROBERTS, G. O., STUART, A. M. AND WHITE, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statist. Sci.* **28**, 424–446.
[11] CRAIU, R. V., GRAY, L., LATUSZYŃSKI, K., MADRAS, N., ROBERTS, G. O. AND ROSENTHAL, J. S. (2015). Stability of adversarial Markov chains, with an application to adaptive MCMC algorithms. *Ann. Appl. Prob.* **25**, 3592–3623.
[12] DA PRATO, G. AND ZABCZYK, J. (2014). *Stochastic Equations in Infinite Dimensions*, 2nd edn (Encycl. Math. Appl. 152). Cambridge University Press.
[13] DUDLEY, R. M. (2018). *Real Analysis and Probability*. Chapman and Hall/CRC, Boca Raton, FL.
[14] DURMUS, A., FORT, G. AND MOULINES, É. (2016). Subgeometric rates of convergence in Wasserstein distance for Markov chains. *Ann. Inst. H. Poincaré Prob. Statist.* **52**, 1799–1822.
[15] DURMUS, A. AND MOULINES, É. (2015). Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm. *Statist. Comput.* **25**, 5–19.
[16] DURMUS, A. AND MOULINES, É. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25**, 2854–2882.
[17] 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.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press

[18] HAARIO, H., SAKSMAN, E. AND TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.
[19] HAARIO, H., SAKSMAN, E. AND TAMMINEN, J. (2005). Componentwise adaptation for high-dimensional MCMC. *Comput. Statist.* **20**, 265–273.
[20] HAIRER, M., MATTINGLY, J. C. AND SCHEUTZOW, M. (2011). Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Prob. Theory Relat. Fields* **149**, 223–259.
[21] HOFSTADLER, J., LATUSZYNSKI, K., ROBERTS, G. O. AND RUDOLF, D. (2024). Almost sure convergence rates of adaptive increasingly rare Markov chain Monte Carlo. Preprint, arXiv:2402.12122.
[22] KALLENBERG, O. (2021). *Foundations of Modern Probability*, 3rd edn. Springer, Cham.
[23] LATUSZYNSKI, K. AND ROSENTHAL, J. S. (2014). The containment condition and AdapFail algorithms. *J. Appl. Prob.* **51**, 1189–1195.
[24] MEYN, S. P. AND TWEEDIE, R. L. (2012). *Markov Chains and Stochastic Stability*. Springer, New York.
[25] NESTEROV, Y. (2018). *Lectures on Convex Optimization*, 2nd edn. Springer, New York.
[26] POMPE, E., HOLMES, C. AND LATUSZYŃSKI, K. (2020). A framework for adaptive MCMC targeting multimodal distributions. *Ann. Statist.* **48**, 2930–2952.
[27] QIN, Q. AND HOBERT, J. P. (2021). On the limitations of single-step drift and minorization in Markov chain convergence analysis. *Ann. Appl. Prob.* **31**, 1633–1659.
[28] QIN, Q. AND HOBERT, J. P. (2022). Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *Ann. Appl. Prob.* **32**, 124–166.
[29] ROBBINS, H. AND MONRO, S. (1951). A stochastic approximation method. *Ann. Statist.* **22**, 400–407.
[30] ROBERTS, G. O. AND ROSENTHAL, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc. B* **60**, 255–268.
[31] ROBERTS, G. O. AND ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16**, 351–367.
[32] ROBERTS, G. O. AND ROSENTHAL, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.* **44**, 458–475.
[33] ROBERTS, G. O. AND ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18**, 349–367.
[34] ROBERTS, G. O., ROSENTHAL, J. S. AND SCHWARTZ, P. O. (1998). Convergence properties of perturbed Markov chains. *J. Appl. Prob.* **35**, 1–11.
[35] ROBERTS, G. O. AND TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363.
[36] SANDRIĆ, N. (2017). A note on the Birkhoff ergodic theorem. *Results Math.* **72**, 715–730.
[37] STRASSEN, V. (1965). The existence of probability measures with given marginals. *Ann. Statist.* **36**, 423–439.
[38] TULCEA, C. I. (1949). Mesures dans les espaces produits. *Atti Accad. Naz. Lincei Rend.* **7**, 208–211.
[39] TWEEDIE, R. L. (1977). Modes of convergence of Markov chain transition probabilities. *J. Math. Anal. Appl.* **60**, 280–291.
[40] VILLANI, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI.
[41] VILLANI, C. (2009). *Optimal Transport: Old and New*. Springer, Berlin.

https://doi.org/10.1017/jpr.2025.4 Published online by Cambridge University Press