Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions

Qian Qin and James P. Hobert
Department of Statistics
University of Florida
February 2019

Abstract

Let $\{X_n\}_{n=0}^{\infty}$ denote an ergodic Markov chain on a general state space that has stationary distribution $\pi$. This article concerns upper bounds on the $L_1$-Wasserstein distance between the distribution of $X_n$ and $\pi$ in the case where the underlying metric is potentially unbounded. In particular, an explicit geometric bound on the distance to stationarity is derived using generalized drift and contraction conditions whose parameters vary across the state space. A corollary of the main result provides an extension of the results in Butkovsky (2014) and Durmus and Moulines (2015), which are applicable only when the metric is bounded. The main conclusion is that the generalized versions of drift and contraction, which comprise the main technical innovation in the article, can yield sharper convergence bounds than the standard versions, whose parameters are constant. Application of the results is illustrated in the context of a simple non-linear autoregressive process.

1 Introduction

Study of the convergence to stationarity of Markov chains commonly requires the specification of a metric on an appropriate space of probability distributions. The standard has long been total variation (TV) distance, but, over the last decade or so, Wasserstein distance has received a good deal of attention as well. One obvious reason for studying convergence in Wasserstein distance is that there exist Markov chains that do not actually converge in TV distance, but do converge in Wasserstein distance (see, e.g. Butkovsky 2014). Another, perhaps more important, reason stems from the current focus on so-called big data problems, which

Key words and phrases. Convergence analysis, Exponential convergence, Kantorovich-Rubinstein distance, Lyapunov drift function, Polish space, Quantitative bound
leads to the study of Markov chains on high-dimensional state spaces. Indeed, it is becoming clear that the techniques used for developing Wasserstein bounds are more robust to increasing dimension than are those used to construct TV bounds (see, e.g., Hairer et al. [2011, 2014], Durmus and Moulines [2015], Qin and Hobert [2018]). In this paper, we study convergence rates of Markov chains with respect to $L_1$-Wasserstein distances induced by potentially unbounded metrics. In particular, we develop explicit geometric (exponential) convergence bounds using generalized versions of the usual drift and contraction conditions.

Previous work devoted to convergence analysis of Markov chains with respect to Wasserstein distances includes Jarner and Tweedie (2001), Hairer et al. (2011), Butkovsky (2014), and Durmus and Moulines (2015). A recurring theme in these papers is the combination of drift and contraction. To be specific, the basic program is to first establish a strong contraction condition on a coupling set (which is a subset of $X \times X$, where $X$ is the state space), and then to use a Lyapunov drift function to drive a coupled version of the Markov chain towards that subset. Our results improve upon this work in two different ways. First, the aforementioned papers deal only with Wasserstein distances induced by bounded metrics. We provide a direct extension that allows for unbounded metrics. All we require of the metric is that it be controlled by the Lyapunov drift function. This extension is important because it is often the case in applications (especially in Markov chain Monte Carlo) that the natural metric is unbounded. While the literature does contain other convergence bounds for Wasserstein distances based on unbounded metrics, they are all substantially different from our bounds in one way or another. Steinsaltz (1999) provides a bound based on the local contractive behavior of the chain. However, this result is difficult to apply because it requires the construction of a special type of drift function that is rather unwieldy relative to the usual Lyapunov drift (see, e.g., Madras and Sezer, 2010). In Qin and Hobert (2018), geometric convergence bounds are constructed using a Lyapunov drift condition and a point-wise contraction condition that is considerably stronger than the average contraction condition that we consider. Finally, Eberle and Majka (2018) provide a method for constructing new (potentially unbounded) metrics that yield geometric convergence in the corresponding Wasserstein distances.

Our second, and more important, contribution is to demonstrate that geometric convergence bounds can be constructed using generalized versions of drift and contraction, which we now describe. When developing drift and contraction conditions for specific problems, the parameters in these inequalities are often initially non-constant. The varying parameters may encode rich information about the dynamics of the chain. The usual drift and contraction conditions (with constant parameters) are typically obtained by taking the supremum of the parameters over the coupling set, and again over the compliment of the coupling set. Naturally, this process can result in a substantial loss of information. The bounds that we provide can be constructed directly from the drift and contraction conditions with non-constant parameters - the generalized drift and
contraction conditions. This procedure does not require selecting a coupling set, and can potentially lead to sharper bounds based on weaker assumptions, compared to previous results. For example, when the metric is bounded (and hence Durmus and Moulines’s (2015) result is applicable), our upper bound on the geometric convergence rate is always better (i.e. smaller) than the corresponding bound of Durmus and Moulines (2015). Our work draws inspiration from the “small function” version of the minorization condition (see, e.g., Nummelin, 1984, Section 2.3), which can be considered a generalized version of the usual minorization condition (with constant parameter). The local contractive behavior of a Markov chain considered in Steinsaltz (1999) is also related to the generalized conditions that we use.

The rest of the article is organized as follows. In Section 2, after setting notation, we state two corollaries of our main result that directly extend results in Butkovsky (2014) and Durmus and Moulines (2015). The statement and proof of our main result are given in Section 3. In Section 4, we use a simple autoregressive process to demonstrate how our results can be applied. This application provides a concrete example of the extent to which bounds constructed via generalized drift and contraction improve upon bounds based on traditional drift and contraction. Finally, Appendix A contains a rigorous comparison of our geometric convergence bounds with that of Durmus and Moulines (2015), and Appendix B contains some technical details supporting the analysis in Section 4.

2 Geometric Convergence with Respect to Unbounded Metrics

Let \((X, \mathcal{B})\) be a countably generated measurable space such that each singleton in \(X\) is measurable. Let \(\psi : X \times X \to [0, \infty)\) be a measurable metric. When we assume that \((X, \psi, \mathcal{B})\) is a Polish metric space, we mean that \((X, \psi)\) is a complete separable metric space, and that \(\mathcal{B}\) is the associated Borel algebra. When Polish-ness is not assumed, we make no explicit assumption about the relationship between \(\psi\) and \(\mathcal{B}\). Let \(\mathcal{P}(X)\) denote the set of probability measures on \((X, \mathcal{B})\), and let \(\delta_x\) denote the point mass (or Dirac measure) at \(x\). For \(\mu, \nu \in \mathcal{P}(X)\), let

\[
C(\mu, \nu) = \{ \nu \in \mathcal{P}(X \times X) : \nu(A_1 \times X) = \mu(A_1), \nu(X \times A_2) = \nu(A_2) \text{ for all } A_1, A_2 \in \mathcal{B} \}.
\]

This is the set of couplings of \(\mu\) and \(\nu\). The \(L_1\)-Wasserstein divergence between \(\mu\) and \(\nu\) is defined to be

\[
W_{\psi}(\mu, \nu) = \inf_{\nu \in C(\mu, \nu)} \int_{X \times X} \psi(x, y) \nu(dx, dy).
\]

3
Define
\[ P_\psi(X) = \left\{ \mu \in P(X) : \int_X \psi(x, y) \mu(dy) < \infty \text{ for some } x \in X \right\}. \]

If \((X, \psi, \mathcal{B})\) is a Polish metric space, then \(W_\psi\) satisfies the triangle inequality and the identity of indiscernibles, \(W_\psi\) is finite on \(P_\psi(X)\), and \((P_\psi(X), W_\psi)\) is itself a complete separable metric space (see, e.g., [Vilani 2009] Definition 6.1 and Theorem 6.18). In this case, \(W_\psi\) is called the \(L_1\)-Wasserstein (or Kantorovich-Rubinstein) distance induced by \(\psi\).

Let \(P : X \times \mathcal{B} \to [0, 1]\) be a Markov transition kernel (Mtk), and for \(n \in \mathbb{N}\), let \(P^n : X \times \mathcal{B} \to [0, 1]\) be the corresponding \(n\)-step Mtk. (As usual, we write \(P\) instead of \(P^1\).) Let \(P^0\) be the identity kernel, i.e., \(P^0(x, C) = 1_{x \in C}\). For \(\mu \in P(X)\) and a measurable function \(f : X \to \mathbb{R}\), let \(\mu f = \int_X f(x) \mu(dx)\), \(\mu P^n(\cdot) = \int_X P^n(x, \cdot) \mu(dx)\), and \(P^n f(\cdot) = \int_X f(x) P^n(\cdot, dx)\) (assuming the integrals are well-defined).

Our goal is to establish conditions under which the Markov chain defined by \(P\) converges in \(W_\psi\) to a limiting distribution \(\pi \in P(X)\) at a geometric rate. More specifically, we will construct convergence bounds of the following form:
\[ W_\psi(\mu P^n, \pi) \leq c_\mu \rho^n, \]
where \(\mu \in P(X)\), \(n \in \{0, 1, 2, \ldots \} =: \mathbb{Z}_+\), \(c_\mu < \infty\), and \(\rho < 1\). Moreover, we will provide explicit formulas for \(c_\mu\) and \(\rho\). Our main result, which is stated and proven in Section 3 requires a good deal of build-up. In the remainder of this section, we present two corollaries of our main result that extend geometric convergence bounds in [Butkovsky 2014] and [Durmus and Moulines 2015].

**Corollary 2.1.** Let \(P\) be an Mtk on a Polish metric space \((X, \psi, \mathcal{B})\). Suppose that the following three conditions hold:

(A1) There exists a measurable function \(V : X \to [0, \infty)\) and \(a \in (0, \infty)\) such that
\[ a^{-1}\psi(x, y) \leq V(x) + V(y) + 1, \quad (x, y) \in X \times X, \]
and there exist \(\eta \in [0, 1)\) and \(L \in [0, \infty)\) such that
\[ PV(x) \leq \eta V(x) + L, \quad x \in X. \]
There exist $\gamma < 1$ and $K < \infty$ such that for each $(x, y) \in X \times X$,

$$W^\psi(\delta_x P, \delta_y P) \leq \begin{cases} 
\gamma \psi(x, y), & (x, y) \in C, \\
K \psi(x, y), & (x, y) \notin C,
\end{cases}$$

where the coupling set $C$ is defined to be $\{(x, y) \in X \times X : V(x) + V(y) < d\}$, with $d > 2L/(1 - \eta)$.

Either $K \leq 1$ or

$$\log K \log(2L + 1) < \log \gamma \log \frac{\eta d + 2L + 1}{d + 1}.$$  \hfill (3)

Then $P$ admits a unique stationary distribution $\pi$. Moreover, $\pi \in \mathcal{P}(X)$, and for each $\mu \in \mathcal{P}(X)$ such that $\mu V < \infty$ and every $n \in \mathbb{Z}_+$,

$$W^\psi(\mu P^n, \pi) \leq a\left(\frac{(\eta + 1)\mu V + L + 1}{1 - \rho_r}\right)^{\rho_r^n},$$  \hfill (4)

where

$$\frac{\log(2L + 1)}{\log(2L + 1) - \log \gamma} < r < \frac{-\log[(\eta d + 2L + 1)/(d + 1)]}{\log(K \vee 1) - \log[(\eta d + 2L + 1)/(d + 1)]},$$  \hfill (5)

and

$$\rho_r = \left[\gamma^r(2L + 1)^{1-r}\right] \vee \left[K^r \left(\frac{\eta d + 2L + 1}{d + 1}\right)^{1-r}\right] < 1.$$

Remark 2.2. The fact that $d > 2L/(1 - \eta)$ implies that $(\eta d + 2L + 1)/(d + 1) < 1$, and it follows that

$$0 \leq \frac{\log(2L + 1)}{\log(2L + 1) - \log \gamma} < \frac{-\log[(\eta d + 2L + 1)/(d + 1)]}{\log(K \vee 1) - \log[(\eta d + 2L + 1)/(d + 1)]} \leq 1$$

whenever $K \leq 1$ or (3) is satisfied. Thus, (5) makes sense, and $r \in (0, 1)$.

Remark 2.3. For a given $r$, $\rho_r$ is an upper bound on the chain’s (true) geometric convergence rate. In practice, it makes sense to optimize the values of $r$ and $d$ so that $\rho_r$ is as small as possible.

Butkovsky’s (2014) Theorem 2.1 and Durmus and Moulines’s (2015) Theorem 1 also provide geometric convergence bounds on the $L_1$-Wasserstein distance to stationarity, but those results are applicable only when the metric $\psi$ is bounded and $K \leq 1$. Note that when these two conditions are satisfied, (A3) and (I) in Corollary 2.1 trivially hold, and the conditions that remain are essentially the same as those of Butkovsky (2014) and Durmus and Moulines (2015). In Appendix A we prove that, in cases where the metric is bounded, our upper bound on the geometric convergence rate is better (i.e., smaller) than that of Durmus and
The convergence bounds in Butkovsky (2014) are not fully computable, so no analytical comparison is attempted. Finally, it is important to note that, while our Corollary 2.1 and Durmus and Moulines (2015) Theorem 1 apply only to chains with geometric convergence rates, Butkovsky’s (2014) result can also be applied to Markov chains with subgeometric convergence rates.

The inequality (2) is commonly referred to as a Lyapunov drift condition, and the condition in (A2) will be called a contraction condition. These are “traditional” in the sense that their parameters, e.g., $\gamma$ and $K$, are constants. In our main result (Theorem 3.5), we will go beyond this setting and consider generalized drift and contraction conditions where the parameters may vary with $(x, y)$. We will demonstrate that this added flexibility can lead to significantly improved convergence bounds.

We end this section with a continuous-time analog of Corollary 2.1.

**Corollary 2.4.** Let $\{P_t\}_{t \geq 0}$ be a Markov semigroup on a Polish metric space $(X, \psi, \mathcal{B})$. Suppose that there exists $t_* > 0$ such that $P_{t_*}$ (in place of $P$) satisfies all the assumptions of Corollary 2.1. Suppose further that there exists $b < \infty$ such that for every $(x, y) \in X \times X$ and $t \in [0, t_*)$,

\[ W_\psi(\delta_x P^t, \delta_y P^t) \leq b \psi(x, y). \]  (7)

Then $\{P_t\}$ has a unique stationary distribution $\pi$. Moreover, $\pi \in \mathcal{P}_\psi(X)$, and for any $\mu \in \mathcal{P}(X)$ such that $\mu V < \infty$ and each $t \geq 0$,

\[ W_\psi(\mu P^t, \pi) \leq ab \left( \frac{(\eta + 1)\mu V + L + 1}{1 - \rho_r} \right)^{\lfloor t/t_* \rfloor} \rho_r^{\lfloor t/t_* \rfloor}, \]  

where $\lfloor \cdot \rfloor$ returns the largest integer that does not exceed its argument, $r$ satisfies (5), and $\rho_r \in [0, 1)$ is defined as in (6).

### 3 Generalized Drift and Contraction and the Main Result

In this section, we will study the convergence properties of the generic Markov chain from Section 2 under the following conditions, which are generalized versions of (A1) and (A2).

(B1) There exists a measurable function $V : X \to [0, \infty)$ and $a \in (0, \infty)$ such that

\[ a^{-1} \psi(x, y) \leq V(x) + V(y) + 1, \quad (x, y) \in X \times X, \]  (1)

and $PV(x) < \infty$ for each $x \in X$. 

[6]
(B2) There exists a measurable function $\Gamma : X \times X \to [0, \infty)$ such that for each $(x, y) \in X \times X$,

$$W_\psi(\delta_x P; \delta_y P) \leq \Gamma(x, y) \psi(x, y).$$

We call these generalized drift and contraction conditions. Note that by (B1), $W_\psi(\delta_x P; \delta_y P) < \infty$ for each $(x, y) \in X \times X$, and thus, (B2) holds trivially with

$$\Gamma(x, y) = \frac{W_\psi(\delta_x P; \delta_y P)}{\psi(x, y)} 1_{x \neq y}.$$

Of course, in practice, we would like to find a $\Gamma$ that yields a sharp contraction inequality, but is also simple and well-behaved. For concrete examples of (B1) and (B2), see Section 4. Needless to say, there is no hope of getting anywhere with (B1) and (B2) alone. In order to construct a convergence bound, we will also need an analog of condition (A3) that regulates the relationship between $\Gamma(x, y)$ and $(PV(x), PV(y))$. This analog is given by (B3) later in this section.

We now introduce the notion of Markovian coupling kernels. Suppose that $P_1$ and $P_2$ are Mtks on $(X, B)$. We say that $\tilde{P} : (X \times X) \times (B \times B) \to [0, 1]$ is a (Markovian) coupling kernel of $P_1$ and $P_2$ if $\tilde{P}$ is an Mtk such that for each $(x, y) \in X \times X$, $\delta(x, y) \tilde{P}$ is in $C(\delta_x P_1, \delta_y P_2)$. In the special case that $P_1 = P_2 = P$, we simply say that $\tilde{P}$ is a coupling kernel of $P$. It’s obvious that (B2) holds if there exists a coupling kernel of $P$, denoted by $\tilde{P}$, such that

$$\tilde{P}\psi(x, y) \leq \Gamma(x, y) \psi(x, y)$$

for each $(x, y) \in X \times X$. The following lemma, which is a direct corollary of Theorem 1.1 in [Zhang 2000], shows that these two conditions are, under reasonable assumptions, equivalent.

**Lemma 3.1.** ([Zhang 2000]) Suppose that $P_1$ and $P_2$ are Mtks on a Polish metric space $(X, \psi, B)$. Then there exists a coupling kernel of $P_1$ and $P_2$, denoted by $\tilde{P}$, such that for each $(x, y) \in X \times X$,

$$W_\psi(\delta_x P_1; \delta_y P_2) = \tilde{P}\psi(x, y).$$

It is well-known that in a Polish metric space $(X, \psi, B)$, for any $\mu, \nu \in \mathcal{P}(X)$, there exists $\nu \in C(\mu, \nu)$ such that $W_\psi(\mu, \nu) = \nu \psi$ (see, e.g., [Villani 2009] Theorem 4.1). However, taking $\mu = \delta_x P_1$ and $\nu = \delta_y P_2$ does not trivially yield Lemma 3.1. Indeed, the key feature of Lemma 3.1 is that $\tilde{P}$ is a *bona fide* Mtk. This is important in our analysis as it protects us from potential measurability problems. An important consequence of Lemma 3.1 is the convexity of $W_\psi$. 

7
Lemma 3.2. Suppose that $P_1$ and $P_2$ are Mtk s on a Polish metric space $(X, \psi, \mathcal{B})$. Let $\mu, \nu \in \mathcal{P}(X)$. Then for any $\nu \in \mathcal{C}(\mu, \nu)$,

$$W_\psi(\mu P_1, \nu P_2) \leq \int_{X \times X} W_\psi(\delta_x P_1, \delta_y P_2) \nu(dx, dy). \quad (9)$$

Moreover,

$$W_\psi(\mu P_1, \nu) \leq \int_X W_\psi(\delta_x P_1, \nu) \mu(dx).$$

Proof. Establishing (9) using Lemma 3.1 is standard. See, e.g., Villani (2009), Lemma 4.8.

Let $P_2$ in (9) be such that $P_2(x, \cdot) = \nu(\cdot)$ for each $x \in X$, and let $\nu(dx, dy) = \mu(dx)\nu(dy)$. Then

$$W_\psi(\mu P_1, \nu) = W_\psi(\mu P_1, \nu P_2) \leq \int_{X \times X} W_\psi(\delta_x P_1, \nu P_2) \mu(dx)\nu(dy) = \int_X W_\psi(\delta_x P_1, \nu) \mu(dx).$$

The following lemma describes a way of constructing a potential contraction condition based on (8) and a “bivariate” drift condition.

Lemma 3.3. Suppose that $P$ is an Mtk on $(X, \mathcal{B})$ that admits a coupling kernel $\tilde{P}$. Suppose further that there exist measurable functions $h : X \times X \to [0, \infty)$, $\Lambda : X \times X \to [0, \infty)$, and $\Gamma : X \times X \to [0, \infty)$ such that for each $(x, y) \in X \times X$,

$$\tilde{P}h(x, y) \leq \Lambda(x, y)h(x, y) \quad \text{and} \quad \tilde{P}\psi(x, y) \leq \Gamma(x, y)\psi(x, y).$$

For each $r \in (0, 1)$, define $\psi_r : X \times X \to [0, \infty)$ by $\psi_r(x, y) = \psi(x, y)^r h(x, y)^{1-r}$, and set

$$\rho_r = \sup_{(x,y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r}.$$ 

Then for every $(x, y) \in X \times X$ and $r \in (0, 1)$,

$$\tilde{P}\psi_r(x, y) \leq \rho_r \psi_r(x, y).$$

Proof. By Hölder’s inequality, for each $r \in (0, 1)$ and $(x, y) \in X \times X$,

$$\tilde{P}\psi_r(x, y) \leq [\tilde{P}\psi(x, y)]^r [\tilde{P}h(x, y)]^{1-r} \leq \Gamma(x, y)^r \Lambda(x, y)^{1-r} \psi_r(x, y) \leq \rho_r \psi_r(x, y).$$
Remark 3.4. The fact that \( \psi \) is a metric was not used in the proof of Lemma 3.3. Thus, the result actually holds for any \( \psi \) that is a non-negative measurable function.

In the proof of Lemma 3.3, we use Hölder’s inequality to establish a new contraction condition. Butkovsky (2014) and Hairer et al. (2011) make similar use of the inequality.

We are now ready to state and prove the main result.

**Theorem 3.5.** Suppose that \( P \) is an \( Mtk \) on a Polish metric space \( (X, \psi, B) \). Assume that \( (B1) \) and \( (B2) \) hold. Let \( \Lambda : X \times X \to [0, \infty) \) be such that

\[
\Lambda(x, y) \geq \frac{P\psi(x) + P\psi(y) + 1}{V(x) + V(y) + 1}.
\]

Assume further that the following condition holds:

\( (B3) \) There exists \( r \in (0, 1) \) such that

\[
\sup_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} < 1. \tag{10}
\]

Then there exists \( \pi \in \mathcal{P}_\psi(X) \) such that for every \( r \in (0, 1) \) that satisfies (10) and each \( x \in X \),

\[
W_\psi(\delta_x P^n, \pi) \leq a \left( \frac{P\psi(x) + V(x) + 1}{1 - \rho_r} \right)^{\rho^n}, \quad n \in \mathbb{Z}_+ \tag{11},
\]

where \( \rho = \sup_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} < 1 \). Moreover, \( P \) has at most one stationary distribution.

**Remark 3.6.** It can be shown that \( (B3) \) holds whenever

\[
\sup_{(x, y) \in X \times X} \left[ \Gamma(x, y) \wedge \Lambda(x, y) \right] < 1,
\]

and

\[
\left( \sup_{(x, y) : \Lambda(x, y) > \Gamma(x, y)} \log \Lambda(x, y) \right)^{\wedge} 0 < \left( \inf_{(x, y) : \Lambda(x, y) < \Gamma(x, y)} \log \Lambda(x, y) \right)^{\wedge} 1.
\]

In this case, \( \rho_r < 1 \) whenever

\[
\left( \sup_{(x, y) : \Lambda(x, y) > \Gamma(x, y)} \log \Lambda(x, y) \right)^{\wedge} 0 < \left( \inf_{(x, y) : \Lambda(x, y) < \Gamma(x, y)} \log \Lambda(x, y) \right)^{\wedge} 1.
\]
Proof of Theorem 3.5. By (B2) and Lemma 3.1, there exists a coupling kernel of $P$, denoted by $	ilde{P}$, such that for each $(x, y) \in X \times X$,

$$\tilde{P}\psi(x, y) \leq \Gamma(x, y)\psi(x, y).$$

Define $h : X \times X \rightarrow [1, \infty)$ as $h(x, y) = V(x) + V(y) + 1$. For each $(x, y) \in X \times X$,

$$\tilde{P}h(x, y) \leq \Lambda(x, y)h(x, y).$$

As in Lemma 3.3, for $r \in (0, 1)$, let $\psi_r : X \times X \rightarrow [0, \infty)$ be such that $\psi_r(x, y) = \psi(x, y)^rh(x, y)^{1-r}$. Let $r$ satisfy (10) so that $\rho_r < 1$. It follows from (1) and Lemma 3.3 that for each $(x, y) \in X \times X$ and $n \in \mathbb{Z}_+$,

$$W_\psi(\delta_xP^n, \delta_yP^n) \leq \tilde{P}^n\psi_r(x, y) \leq a^{1-r}\tilde{P}^n\psi_r(x, y) \leq a^{1-r}\psi_r(x, y)\rho_r^n \leq ah(x, y)\rho_r^n. \quad (12)$$

Now, fix $x \in X$, and let $\upsilon_x \in C(\delta_x, \delta_xP)$. Then by Lemma 3.2 and (12), for each $n \in \mathbb{Z}_+$,

$$W_\psi(\delta_xP^n, \delta_xP^{n+1}) \leq \int_{X \times X} W_\psi(\delta_xP^n, \delta_yP^n) \upsilon_x(dx', dy') \leq a(\upsilon_xh)\rho_r^n = a(PV(x) + V(x) + 1)\rho_r^n. \quad (13)$$

By (13) and the triangle inequality, for each positive integer $n$,

$$\int_X \psi(x, y)P^n(x, dy) = W_\psi(\delta_x, \delta_xP^n) \leq \sum_{k=0}^{n-1} W_\psi(\delta_xP^k, \delta_xP^{k+1}) < \infty.$$

Thus, $\delta_xP^n \in \mathcal{P}_\psi(X)$ for each $n \in \mathbb{Z}_+$. Moreover, (13) shows that

$$\sum_{k=n}^{\infty} W_\psi(\delta_xP^k, \delta_xP^{k+1}) \leq a \left(\frac{PV(x) + V(x) + 1}{1 - \rho_r}\right)\rho_r^n < \infty. \quad (14)$$

This means that $\{\delta_xP^n\}_{n \in \mathbb{Z}_+}$ is Cauchy in the metric space $(\mathcal{P}_\psi(X), W_\psi)$. Recall that $(\mathcal{P}_\psi(X), W_\psi)$ is Polish, and thus, complete. Hence, there exists $\pi_x \in \mathcal{P}_\psi(X)$ such that

$$\lim_{n \rightarrow \infty} W_\psi(\delta_xP^n, \pi_x) = 0.$$

Note that $\pi_x$ does not depend on $x$. To see this, let $y \in X$. Then by (12),

$$\lim_{n \rightarrow \infty} W_\psi(\delta_yP^n, \delta_xP^n) = 0.$$
Thus, \( \pi_x = \pi_y \) for each \((x, y) \in X \times X \). We will simply denote \( \pi_x \) by \( \pi \). By (14) and the triangle inequality,

\[
W_\psi(\delta_x P^n, \pi) \leq \sum_{k=n}^{\infty} W_\psi(\delta_x P^k, \delta_x P^{k+1}) \leq a \left( \frac{PV(x) + V(x) + 1}{1 - \rho_r} \right) \rho_r^n,
\]

which establishes (11).

To show that \( P \) has at most one stationary distribution, we consider the truncated metric \( \psi_* = \psi \wedge 1 \).

Note that \( \psi_* \) is topologically equivalent to \( \psi \). Hence, \((X, \psi_*)\) is Polish, and \((\mathcal{P}(X), W_{\psi_*})\) is a metric space.

It follows from (11) that for any \( x \in X \) and \( n \in \mathbb{Z}_+ \),

\[
W_{\psi_*}(\delta_x P^n, \pi) \leq \left[ a \left( \frac{PV(x) + V(x) + 1}{1 - \rho_r} \right) \rho_r^n \right] \wedge 1.
\]

By Lemma 3.2 and the dominated convergence theorem, for any \( \mu \in \mathcal{P}(X) \),

\[
\lim_{n \to \infty} W_{\psi_*}(\mu P^n, \pi) \leq \lim_{n \to \infty} \int_X W_{\psi_*}(\delta_x P^n, \pi) \mu(dx) = 0.
\]

Now, if \( \mu \) is stationary, then the above inequality shows that \( W_{\psi_*}(\mu, \pi) = 0 \), and the proof is complete.

Let \((X, \psi, \mathcal{B})\) be a Polish metric space, and let \( C_b(X) \) be the set of bounded, continuous real-valued functions on \( X \). We say that a sequence \( \{\mu_n\} \subset \mathcal{P}(X) \) converges weakly to \( \mu \in \mathcal{P}(X) \) if \( \lim_{n \to \infty} \mu_n f = \mu f \) for every \( f \in C_b(X) \). Let \( P \) be an Mtk on \((X, \psi, \mathcal{B})\). We say that \( P \) is weak Feller if \( Pf \in C_b(X) \) for each \( f \in C_b(X) \). The following result complements Theorem 3.5.

**Proposition 3.7.** Let \( P \) be an Mtk on a Polish metric space \((X, \psi, \mathcal{B})\) that satisfies all the assumptions in Theorem 3.5. If either \( \Gamma : X \times X \to [0, \infty) \) in (B2) is bounded, or \( P \) is weak Feller, then \( \pi \) (as in Theorem 3.5) is the unique stationary distribution of \( P \).

**Proof.** It suffices to show that \( \pi \) is stationary. Assume first that \( \Gamma \) is bounded above by a constant \( K \). Then for each \((x, y) \in X \times X \),

\[
W_\psi(\delta_x P, \delta_y P) \leq K \psi(x, y).
\]  

(15)

Now, fix \( x_0 \in X \). For \( n \in \mathbb{Z}_+ \), let \( v_n \in \mathcal{C}(\delta_{x_0} P^n, \pi) \) be such that \( v_n \psi = W_\psi(\delta_{x_0} P^n, \pi) \). For each \( n \in \mathbb{Z}_+ \),
by the triangle inequality, Lemma 3.2 and (15),
\[
W_\psi(\pi, \pi P) \leq W_\psi(\pi, \delta_{x_0}P^{n+1}) + W_\psi(\delta_{x_0}P^{n+1}, \pi P)
\]
\[
\leq W_\psi(\pi, \delta_{x_0}P^{n+1}) + \int_X W_\psi(\delta_x P, \delta_y P) v_n(dx, dy)
\]
\[
\leq W_\psi(\pi, \delta_{x_0}P^{n+1}) + K v_n \psi
\]
\[
= W_\psi(\pi, \delta_{x_0}P^{n+1}) + K W_\psi(\delta_{x_0}P^n, \pi).
\]

By Theorem 3.5 letting \( n \to \infty \) yields \( \pi = \pi P \).

Suppose alternatively that \( P \) is weak Feller. By Theorem 3.5, for any \( x \in X \), \( \{\delta_x P^n\}_n \) converges to \( \pi \) in \( W_\psi \), which implies that \( \{\delta_x P^n\}_n \) converges weakly to \( \pi \) (see, e.g., Villani 2009, Theorem 6.9). Let \( f \in C_b(X) \) be arbitrary. Since \( P \) is weak Feller, \( Pf \in C_b(X) \), and \( \lim_{n \to \infty} \delta_x P^n(Pf) = \pi(Pf) \). It follows that \( \{\delta_x P^n\} \) converges weakly to \( \pi P \) as well. This is enough to ensure that \( \pi = \pi P \) (see, e.g., Billingsley 1999, Theorem 1.2).

We now prove Corollaries 2.1 and 2.4.

**Proof of Corollary 2.1.** By (A1), condition (B1) holds with the same \( V : X \to [0, \infty) \). Recall that \( d > 2L/(1-\eta) \), and \( C = \{(x, y) \in X \times X : V(x) + V(y) < d\} \). By (A2), condition (B2) is satisfied with
\[
\Gamma(x, y) = \gamma 1_{(x, y) \in C} + K 1_{(x, y) \notin C}, \quad (x, y) \in X \times X.
\]
Again by (A1), for each \( (x, y) \in X \times X \),
\[
\frac{PV(x) + PV(y) + 1}{V(x) + V(y) + 1} \leq \eta + \frac{2L - \eta + 1}{V(x) + V(y) + 1}
\]
\[
\leq \begin{cases} 
2L + 1, & (x, y) \in C, \\
\lambda, & (x, y) \notin C,
\end{cases}
\]
where \( \lambda = (\eta d + 2L + 1)/(d + 1) < 1 \). Let
\[
\Lambda(x, y) = (2L + 1) 1_{(x, y) \in C} + \lambda 1_{(x, y) \notin C}, \quad (x, y) \in X \times X.
\]
Then
\[
\Lambda(x, y) \geq \frac{PV(x) + V(x) + 1}{V(x) + V(y) + 1}
\]
for each \((x, y) \in X \times X\), as in Theorem 3.5.

We now establish \((B3)\) using \((A3)\). Let \(r\) satisfy \((5)\), that is,

\[
\frac{\log(2L + 1)}{\log(2L + 1) - \log \gamma} < r < \frac{-\log \lambda}{\log K - \log \lambda} \quad 1_{K > 1} + 1_{K \leq 1}.
\]

Note that

\[
\sup_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} = \left[ \gamma^r (2L + 1)^{1-r} \right] \vee \left( K^r \lambda^{1-r} \right).
\]

Suppose that \(K \leq 1\), then

\[
\left[ \gamma^r (2L + 1)^{1-r} \right] \vee \left( K^r \lambda^{1-r} \right) \leq \left[ \frac{\gamma}{2L + 1} \right]^r (2L + 1) \vee \lambda^{1-r} < 1.
\]

If, on the other hand, \(K > 1\), then

\[
\left[ \gamma^r (2L + 1)^{1-r} \right] \vee \left( K^r \lambda^{1-r} \right) = \left[ \left( \frac{\gamma}{2L + 1} \right)^r (2L + 1) \right] \vee \left[ \left( \frac{K}{\lambda} \right)^r \lambda \right] < 1.
\]

Thus, \((B3)\) holds, and for each \(r\) satisfying \((5)\), \(\sup_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} < 1\).

By Theorem 3.5 and \((A1)\), there exists \(\pi \in \mathcal{P}_\psi(X)\) such that for each \(x \in X\) and \(n \in \mathbb{Z}_+\),

\[
W_\psi(\delta_x P^n, \pi) \leq a \left( \frac{(\eta + 1)V(x) + L + 1}{1 - \rho_r} \right) \rho_r^n,
\]

where \(r\) satisfies \((5)\), and

\[
\rho_r = \sup_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} = \left[ \gamma^r (2L + 1)^{1-r} \right] \vee \left( K^r \lambda^{1-r} \right).
\]

Let \(\mu \in \mathcal{P}(X)\) be such that \(\mu V < \infty\). Applying Lemma 3.2 yields the geometric convergence bound \((4)\).

Finally, since \(\Gamma\) is bounded, by Proposition 3.7, \(\pi\) is the unique stationary distribution.

\[\square\]

Proof of Corollary 2.4. By Corollary 2.1 there exists a unique stationary distribution \(\pi\) for \(P^{t*}\). Moreover, for any \(\mu \in \mathcal{P}(X)\) such that \(\mu V < \infty\), \(r\) satisfying \((5)\), and \(n \in \mathbb{Z}_+\),

\[
W_\psi(\mu P^{nt*}, \pi) \leq a \left( \frac{(\eta + 1)\mu V + L + 1}{1 - \rho_r} \right) \rho_r^n. \tag{16}
\]

To show that \(\pi\) is stationary for \(\{P^t\}\), we use an argument from Butkovsky (2014). Note that for any \(s \geq 0\), \(\pi P^s\) is also stationary for \(P^{t*}\), since \(\pi P^s P^{t*} = \pi P^{t*} P^s = \pi P^s\). Because \(\pi\) is the unique stationary
distribution for $P^{t_s}, \pi P^s = \pi$ for any $s \geq 0$. Hence, $\pi$ is the unique stationary distribution for $\{P^t\}$.

Now, let $t \geq 0$, and let $s = t - \lfloor t/t_s \rfloor t_s \in [0, t_s)$. Let $\mu \in \mathcal{P}(X)$ be such that $\mu V < \infty$. Let $\nu \in \mathcal{C}(\mu P^{t/t_s} t_s, \pi)$ be such that $\nu \psi = W_\psi(\mu P^{t/t_s} t_s, \pi)$. Applying Lemma 3.2 and (1) shows that

$$W_\psi(\mu P^t, \pi) = W_\psi(\mu P^{t/t_s} t_s, \pi P^s, \pi P^s) \leq \int_{X \times X} W_\psi(\delta_x P^s, \delta_y P^s) \nu(dx, dy) \leq bW_\psi(\mu P^{t/t_s} t_s, \pi).$$

The result then follows immediately from (16). \hfill \square

From the proof of Corollary 2.1, we can see that the said result is essentially an application of Theorem 3.5 when $\Lambda$ and $\Gamma$ are constant over a coupling set, $C$, as well as over its complement, $(X \times X) \setminus C$. To make a comparison between the two results, consider the following scenario. Let $P$ be an Mtk on a Polish metric space $(X, \psi, B)$ that satisfies (A1) with a drift function $V : X \to [0, \infty)$, $\alpha > 0$, $\eta < 1$, and $L \in [0, \infty)$. Then $P$ satisfies (B1) with the same $V$ and $\alpha$. Suppose that $P$ also satisfies (B2) with $\Gamma : X \times X \to [0, \infty)$. Let

$$\Lambda(x, y) = \frac{\eta V(x) + \eta V(y) + 2L + 1}{V(x) + V(y) + 1} \geq \frac{PV(x) + PV(y) + 1}{V(x) + V(y) + 1}$$

for $(x, y) \in X \times X$, and assume that (B3) holds. Then by Theorem 3.5, the chain admits a limiting distribution $\pi$ to which it converges geometrically in $W_\psi$. For $r \in (0, 1)$, let $\rho^B_r = \sup_{(x,y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r}$. By Theorem 3.5 whenever $\rho^B_r < 1$, this is a nontrivial upper bound on the chain’s geometric convergence rate. Here, we use the superscript “$B$” to indicate that $\rho_r$ is constructed based on (B1) − (B3).

Suppose now that we are ignorant of Theorem 3.5, and wish to find a convergence bound using Corollary 2.1. We need to convert (B2) to (A2) by letting $\gamma = \sup_{(x,y) \in C} \Gamma(x, y)$, and $K = \sup_{(x,y) \notin C} \Gamma(x, y)$, where $C = \{(x, y) : V(x) + V(y) < d\}$ is a coupling set, and $d > 2L/(1 - \eta)$. Of course, we need to further assume that $\gamma < 1$ and $K < \infty$, so that (A2) is satisfied. For $r \in (0, 1)$,

$$\rho^A_r := \left[\gamma^r(2L + 1)^{1-r} \right] \lor \left[K^r \left(\frac{\eta d + 2L + 1}{d + 1}\right)^{1-r}\right] \geq \left\{ \sup_{(x,y) \in C} \Gamma(x, y)^r \right\} \lor \left\{ \sup_{(x,y) \notin C} \Gamma(x, y)^r \right\}.$$
Corollary 2.1. On the other hand, note that
\[
\rho^B_r = \left[ \sup_{(x, y) \in C} \Gamma(x, y)^r \Lambda(x, y)^{1-r} \right] \vee \left[ \sup_{(x, y) \not\in C} \Gamma(x, y)^r \Lambda(x, y)^{1-r} \right].
\]
It’s clear that \( \rho^B_r \leq \rho^A_r \) for each \( r \in (0, 1) \). So \( \rho^A_r < 1 \) only if \( \rho^B_r < 1 \). Moreover, the inequality between \( \rho^A_r \) and \( \rho^B_r \) will be strict unless \( \Gamma \) and \( \Lambda \) are related in a very specific manner that seems unlikely to hold in practice. Thus, Theorem 3.5 provides a sharper convergence rate bound than Corollary 2.1.

4 Example: A Nonlinear Autoregressive Process

Let \((X, \psi)\) be \(p\)-dimensional Euclidean space equipped with the Euclidean distance. Consider a Markov chain, \( \{X_n\}_{n=0}^\infty \), defined as follows,
\[
X_{n+1} = g(X_n) + Z_n,
\]
where \( g : X \rightarrow X \), and \( \{Z_n\}_{n=0}^\infty \) is a sequence of iid random vectors with mean zero and identity covariance matrix. See Douc et al. (2004) (and the references therein) for an in-depth discussion of the convergence properties of this family of Markov chains. In this section, we compare the numerical bounds resulting from applications of Corollary 2.1 and Theorem 3.5 to a particular member of the family. Of course, since the Euclidean distance is unbounded, the results of Butkovsky (2014) and Durmus and Moulines (2015) are not applicable.

To establish \((B1)\), we take \( V(x) = (cp)^{-1} \|x\|^2 \), where \( \| \cdot \| \) is the Euclidean norm, and \( c > 0 \) is a parameter that can be tuned. Note that
\[
\psi(x, y) \leq \|x\| + \|y\| \leq \|x\|^2 + \|y\|^2 + \frac{1}{2},
\]
so \((1)\) will hold with \( a = cp \) as long as \( c \geq (2p)^{-1} \), which is assumed in what follows. For every \( x \in X \),
\[
PV(x) = (cp)^{-1} \|g(x)\|^2 + c^{-1}.
\]
To establish \((B2)\), let \( \{(X_n, Y_n)\}_{n=0}^\infty \) be a coupled version of the chain defined by
\[
X_{n+1} = g(X_n) + Z_n \quad \text{and} \quad Y_{n+1} = g(Y_n) + Z_n,
\]
and let \( \hat{P} \) be the corresponding coupling kernel. For each \((x, y) \in X \times X\),

\[
W_\psi(\delta_x P, \delta_y P) \leq \hat{P}_\psi(x, y) = \mathbb{E}\left( \|X_1 - Y_1\| \big| X_0 = x, Y_0 = y \right) = \|g(x) - g(y)\|.
\]

Now, let \( \Lambda : X \times X \to [0, \infty) \) and \( \Gamma : X \times X \to [0, \infty) \) be such that

\[
\Lambda(x, y) \geq \frac{\|g(x)\|^2 + \|g(y)\|^2 + 2p + cp}{\|x\|^2 + \|y\|^2 + cp}, \quad \text{and} \quad \Gamma(x, y) \geq \frac{\|g(x) - g(y)\|}{\|x - y\|} 1_{x\neq y}.
\]

Suppose that (B3) holds. Then by Theorem 3.5, there exists \( \pi \in \mathcal{P}_\psi(X) \) such that for each \( r \in (0, 1) \) that satisfies (10) and each \( x \in X\),

\[
W_\psi(\delta_x P^n, \pi) \leq \left( \frac{\|g(x)\|^2 + \|x\|^2 + cp + p}{1 - \rho_r} \right) \rho_r^n, \quad n \in \mathbb{Z}_+,
\]

where \( \rho_r = \sup_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} < 1 \).

For illustration, let \( X = \mathbb{R} \), and set

\[
g(x) = \frac{x}{2} - \frac{\sin x}{2}.
\]

Then \( \{X_n\}_{n=0}^{\infty} \) is a linear autoregressive chain perturbed by a trigonometric term. We first provide a convergence rate bound based on Corollary 2.1. Letting \( V(x) = x^2 \) for \( x \in X \), we have

\[
PV(x) = \frac{x^2}{4} - \frac{x \sin x}{2} + \left( \frac{\sin x}{2} \right)^2 + 1 \leq \frac{x^2}{2} + \frac{(\sin x)^2}{2} + 1 \leq \frac{x^2}{2} + \frac{3}{2}.
\]

It follows that (A1) holds with \( a = 1, \eta = 1/2 \), and \( L = 3/2 \). Let \( d > 2L/(1 - \eta) = 6 \), and set \( C = \{(x, y) \in X \times X : x^2 + y^2 < d\} \). For \((x, y) \in X \times X\), let

\[
\Gamma(x, y) = \frac{|g(x) - g(y)|}{|x - y|} 1_{x\neq y} + |g'(x)| 1_{x=y} = \frac{1}{2} \left( 1 - \frac{\sin x - \sin y}{x - y} \right) 1_{x\neq y} + \left( \frac{1}{2} - \frac{\cos x}{2} \right) 1_{x=y}.
\]

Then \( |g(x) - g(y)| = \Gamma(x, y)|x - y| \). One can verify that \( \sup_{(x, y) \in X \times X} \Gamma(x, y) = 1 \). Moreover, if \( d \geq 2\pi^2 \), then \( \sup_{(x, y) \in C} \Gamma(x, y) = \Gamma(\pi, \pi) = 1 \). Let \( d \in (6, 2\pi^2) \), and let \( \gamma = \sup_{x^2 + y^2 < d} \Gamma(x, y) \). It can be shown that \( \gamma < 1 \), and (A2) is satisfied with \( K = 1 \). The relation between \( \gamma \) and \( d \) is shown in Figure 1. Since \( K = 1 \), (A3) holds. We can now use (6) to obtain an upper bound on the convergence rate of the chain,
(a) The values of $\gamma = \sup_{x^2 + y^2 < d} \Gamma(x, y)$ when $d$ ranges over $(6, 2\pi^2)$, calculated numerically.

(b) A heat map of $\Gamma(x, y)^r \Lambda(x, y)^{1-r}$ for $r = 0.395$. The supremum of this function is achieved at around $(-2.3, -2.3)$ and $(2.3, 2.3)$, and is about 0.814.

Figure 1: Features of $\Gamma$ and $\Lambda$, defined in (18) and (19).

namely,

$$\rho^A_r = \left[\gamma^r (2L + 1)^{1-r}\right] \lor \left[K^r \left(\frac{\eta d + 2L + 1}{d + 1}\right)^{1-r}\right].$$

Note that this bound depends on $r$ and $d$, both of which can be optimized. The infimum of $\rho^A_r$ is roughly 0.976, and this value occurs when $r \approx 0.856$ and $d \approx 9.2$. This bound can be improved by letting $V(x) = x^2/c$ and optimizing $c \in [1/2, \infty)$, or by finding a sharper drift inequality than (17), but we do not pursue this any further.

We now provide an alternative bound by applying Theorem 3.5 directly. By (17), (B1) holds, and for every $(x, y) \in X \times X$,

$$\frac{PV(x) + PV(y) + 1}{V(x) + V(y) + 1} \leq \frac{x^2/2 + y^2/2 + 4}{x^2 + y^2 + 1} =: \Lambda(x, y).$$

Let $\Gamma : X \times X \to [0, \infty)$ be defined as in (18). It’s easy to see that (B2) is satisfied. To verify (B3), we will evaluate $\sup_{(x,y)\in X\times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r}$ for $r \in (0, 1)$. Figure 1b plots the bivariate function $\Gamma(x, y)^r \Lambda(x, y)^{1-r}$ for a specific value of $r$. It is shown in Appendix B that, for each $r \in (0, 1)$, one can find $\sup_{(x,y)\in X\times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r}$ numerically through optimization over a compact subset of $X \times X$. When $r$ takes the value 0.395 (which is optimal), an upper bound on the chain’s convergence rate is

$$\rho^B_r := \sup_{(x,y)\in X\times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} \approx 0.814.$$

This is a significant improvement on $\rho^A_r$, which is constructed using traditional drift and contraction condi-
tions. The bound can be further improved to \( \rho_r^B = 0.577 \) (with \( r = 0.382 \)) if we let

\[
\Lambda(x, y) = \frac{PV(x) + PV(y) + 1}{V(x) + V(y) + 1} = \frac{[x/2 - \sin(x)/2]^2 + [y/2 - \sin(y)/2]^2 + 3}{x^2 + y^2 + 1}.
\]  

We note that it’s not really fair to compare the second \( \rho_r^B \) with \( \rho_r^A \), since the latter is based on the loosened drift inequality (17).

Of course, in more complicated problems, \( \sup_{(x,y) \in \mathcal{X} \times \mathcal{X}} \Gamma(x,y) \Lambda(x,y)^{1-r} \) would likely be much more difficult to evaluate. Nevertheless, this example shows that generalized drift and contraction may contain useful information that is not available in their traditional counterparts.

Acknowledgment. The second author was supported by NSF Grant DMS-15-11945.

Appendix

A Comparison of \( \rho_r \) and Durmus and Moulines’s Geometric Rate Bound

We begin with a formal statement of Durmus and Moulines’s result.

**Theorem A.1.** ([Durmus and Moulines 2015] Let \( P \) be an Mtk on a Polish metric space \( (\mathcal{X}, \psi, \mathcal{B}) \), where \( \psi \) is bounded above by 1. Suppose that the following two conditions hold:

(C1) There exists a measurable function \( \bar{V} : \mathcal{X} \to [1, \infty) \), \( \eta' \in [0, 1) \), and \( L' \in [0, \infty) \) such that

\[
P \bar{V}(x) \leq \eta' \bar{V}(x) + L', \quad x \in \mathcal{X}.
\]

(C2) There exists \( \gamma' < 1 \) such that for each \( (x,y) \in \mathcal{X} \times \mathcal{X} \),

\[
W_\psi(\delta_x P, \delta_y P) \leq \begin{cases} 
\gamma' \psi(x,y), & (x,y) \in C', \\
\psi(x,y), & (x,y) \notin C',
\end{cases}
\]

where the coupling set \( C' \) is defined to be \( \{(x,y) \in \mathcal{X} \times \mathcal{X} : \bar{V}(x) + \bar{V}(y) \leq (2L' + \delta')/(1 - \eta') \} \), for some \( \delta' > 0 \).
Then \( P \) admits a unique stationary distribution \( \pi \). Moreover, for each \( x \in X \) and every \( n \in \mathbb{Z}_+ \),

\[
W_\psi(\delta_x P^n, \pi) \leq \left( \frac{1}{2} + \frac{1}{(\gamma')^2} \right) \left( 1 + \frac{L'}{1 - \eta'} \right) V(x) (\rho_{DM})^n ,
\]

where

\[
\rho_{DM} = \exp \left( - \frac{\log \lambda \log \gamma'}{\log J - \log \gamma'} \right),
\]

\[
\lambda = 2L'(1 - \eta')/(2L' + \delta') + \eta', \text{ and } J = (2L' + \delta')/(1 - \eta') + 2L'/\lambda.
\]

The following result shows that, in cases where the metric is bounded, Corollary 2.1 can always be used to improve upon \( \rho_{DM} \).

**Proposition A.2.** Let \( P \) be an Mtk on a Polish metric space \((X, \psi, \mathcal{B})\), where \( \psi \) is bounded above by 1, and assume that (\( C_1 \)) and (\( C_2 \)) are satisfied. Then Corollary 2.1 can be used to construct an alternative convergence rate bound, \( \rho_r \), such that \( \rho_r < \rho_{DM} \).

**Proof.** Since \( \bar{V} \geq 1 \), (\( C_1 \)) implies that \( L' \geq (1 - \eta') \pi \bar{V} \geq 1 - \eta' > 0 \) (see Hairer [2006] Proposition 4.24). Thus, \( \lambda \in (\eta', 1) \) and \( J > 1 \).

We now translate (\( C_1 \)) and (\( C_2 \)) into (\( A_1 \)) – (\( A_3 \)) and apply Corollary 2.1. Let \( V(x) = \bar{V}(x) - 1/2 \) for \( x \in X \). Then by (\( C_1 \)) and the boundedness of \( \psi \), (\( A_1 \)) holds with \( a = 1, \eta = \eta' \), and \( L = L' + \eta'/2 - 1/2 \). (Note that \( L \geq 0 \).) Take \( d = (2L + \delta')/(1 - \eta) \), and set

\[
C = \{(x, y) \in X \times X : V(x) + V(y) < d\} = \{(x, y) \in X \times X : \bar{V}(x) + \bar{V}(y) < \frac{2L' + \delta'}{1 - \eta'} \} \subset C'.
\]

By (\( C_2 \)), condition (\( A_2 \)) holds with \( \gamma = \gamma', K = 1, \) and \( C \) given above. Since \( K = 1 \), (\( A_3 \)) also holds, and Corollary 2.1 states that the chain’s \( \psi \)-induced Wasserstein distance to its unique stationary distribution decreases at a geometric rate of (at most)

\[
\rho_r = \left[ \gamma^r (2L + 1)^{1-r} \right] \vee \left( \frac{\eta d + 2L + 1}{d + 1} \right)^{1-r},
\]

where

\[
r \in \left( \frac{\log(2L + 1)}{\log(2L + 1) - \log \gamma}, 1 \right).
\]

Note that

\[
\frac{\eta d + 2L + 1}{d + 1} = \eta' + \frac{2L'}{d + 1} = \lambda.
\]
Now set
\[ r = \frac{\log(2L + 1) - \log \lambda}{\log(2L + 1) - \log \lambda - \log \gamma}. \]

Then
\[ \rho_r = \lambda^{1-r} = \exp \left( -\frac{\log \lambda \log \gamma'}{\log(2L + 1) - \log \lambda - \log \gamma'} \right). \]

Since \( L' \geq 1 - \eta' \), \( J \lambda > \frac{2}{2L} + 2L' > \eta' + 2L' = 2L + 1 \). As a result, \( 0 < \log(2L + 1) - \log \lambda < \log J \), which implies that \( \rho_r < \rho_{DM} \).

Finally, as argued at the end of Section 3, if \( \rho_r \) is calculated based on a set of generalized drift and contraction conditions, then it may be further improved. As demonstrated in Section 4, the convergence rate bound in Theorem 3.5 can be considerably sharper than that in Corollary 2.1, and thus, substantially sharper than \( \rho_{DM} \) as well.

**B Finding \( \rho_r^B \) for the Perturbed Linear Autoregressive Chain**

Let \( X = \mathbb{R} \), and consider the linear autoregressive chain perturbed by a trigonometric term from Section 4.

We now explain how to find
\[ \rho_r^B = \sup_{(x,y) \in X \times X} \Gamma(x,y)^r \Lambda(x,y)^{1-r}, \quad r \in (0, 1), \]

where \( \Lambda \) and \( \Gamma \) are, respectively, given by (19) and (18). The main result is as follows.

**Proposition B.1.** Let \( \Lambda \) and \( \Gamma \) be defined as in (19) and (18). Then for any \( r \in (0, 1) \),
\[ \arg \max_{(x,y) \in X \times X} \Gamma(x,y)^r \Lambda(x,y)^{1-r} \subset C_0 := \{(x,y) : x^2 + y^2 \leq 2\pi^2 \}. \]

**Proof.** Let \( r \in (0, 1) \) and \( (x,y) \in (X \times X) \setminus C_0 \) be arbitrary. It suffices to show that there exists \( (x', y') \in C_0 \) such that
\[ \Gamma(x,y)^r \Lambda(x,y)^{1-r} < \Gamma(x', y')^r \Lambda(x', y')^{1-r}. \]

By the mean value theorem, there exists a point \( \xi \in [x, y] \) (or \( [y, x] \)), as well as a point \( \xi' \in [-\pi, \pi] \) such that
\[ \Gamma(x,y) = \frac{1}{2} - \frac{\cos \xi}{2} = \frac{1}{2} - \frac{\cos \xi'}{2} = \Gamma(\xi', \xi') \]

Note that \( (\xi', \xi') \in C_0 \). Moreover, it’s easy to verify that \( \Lambda(x,y) < \Lambda(\xi', \xi') \). As a result, (21) holds with
Proposition B.1 implies that, to maximize \( \Gamma(x, y)^r \Lambda(x, y)^{1-r} \) over \( X \times X \), we only need to restrict our attention to the compact set \( C_0 \). Since the objective function is uniformly continuous on \( C_0 \), we can solve the problem by optimizing over a sufficiently fine grid.

Finally, assume instead that \( \Lambda \) is given by (20). The analog of Proposition B.1 is as follows.

**Proposition B.2.** Let \( \Lambda \) and \( \Gamma \) be defined as in (20) and (18). Then for any \( r \in (0, 1) \),

\[
\arg \max_{(x, y) \in X \times X} \Gamma(x, y)^r \Lambda(x, y)^{1-r} \subset C_0' := \{ (x, y) : |x| \leq 26, |y| \leq 26 \}.
\]

**Proof.** Let \( r \in (0, 1) \) and \( (x, y) \in (X \times X) \setminus C_0' \) be arbitrary. It suffices to show that there exists \( (x', y') \in C_0' \) such that (21) holds. As in the proof of Proposition B.1, there exists a point \( \xi \in [-2\pi, -\pi] \cup [\pi, 2\pi] \) such that \( \Gamma(x, y) = \Gamma(\xi, \xi) \). Note that \( (\xi, \xi) \in C_0' \). Moreover, \( \xi \) and \( \sin \xi \) have opposite signs. Thus,

\[
\Lambda(\xi, \xi) \geq \frac{\xi^2/2 + 3}{2\xi^2 + 1} > 0.284.
\]

Since \( (x, y) \notin C_0' \), we have \( |x| > 26 \) or \( |y| > 26 \). Without loss of generality, assume that the former holds. Then \( |x| < x^2/26 \), and \( |y| + (x^2 + 1)/|y| > 52 \). It follows that

\[
\Lambda(x, y) \leq \frac{(|x| + 1)^2/4 + (|y| + 1)^2/4 + 3}{x^2 + y^2 + 1} = 0.25 + \frac{0.5|x| + 0.5|y| + 3.25}{x^2 + y^2 + 1} < 0.25 + \frac{x^2}{52(x^2 + y^2 + 1)} + \frac{0.5}{|y| + (x^2 + 1)/|y|} + \frac{3.25}{x^2 + y^2 + 1} < 0.25 + 1/52 + 0.5/52 + 3.25/(26^2 + 1) < 0.284.
\]

Hence, \( \Lambda(x, y) < \Lambda(\xi, \xi) \), and (21) holds with \( (x', y') = (\xi, \xi) \).

**References**

BILLINGSLEY, P. (1999). *Convergence of Probability Measures*. 2nd ed. John Wiley & Sons.

BUTKOVSKY, O. (2014). Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *Annals of Applied Probability* 24 526–552.
DOUC, R., FORT, G., MOULINES, E. and SOULIER, P. (2004). Practical drift conditions for subgeometric rates of convergence. *Annals of Applied Probability* **14** 1353–1377.

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. *Statistics and Computing* **25** 5–19.

EBERLE, A. and MAJKA, M. B. (2018). Quantitative contraction rates for Markov chains on general state spaces. *arXiv:1808.07033*.

HAIRER, M. (2006). Ergodic properties of Markov processes. Lecture notes.

HAIRER, M., MATTINGLY, J. C. and SCHEUTZOW, M. (2011). Asymptotic coupling and a general form of Harris theorem with applications to stochastic delay equations. *Probability Theory and Related Fields* **149** 223–259.

HAIRER, M., STUART, A. M. and VOLLMER, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Annals of Applied Probability* **24** 2455–2490.

JARNER, S. and TWEEDIE, R. (2001). Locally contracting iterated functions and stability of Markov chains. *Journal of Applied Probability* **38** 494–507.

MADRAS, N. and SEZER, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16** 882–908.

NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge, London.

QIN, Q. and HOBERT, J. P. (2018). Wasserstein-based methods for convergence complexity analysis of MCMC with application to Albert and Chib’s algorithm. *arXiv:1810.08826*.

STEINSALTZ, D. (1999). Locally contractive iterated function systems. *Annals of Probability* **27** 1952–1979.

VILLANI, C. (2009). *Optimal Transport: Old and New*, vol. 338 of *Grundlehren der mathematischen Wissenschaften*. Springer.

ZHANG, S. (2000). Existence and application of optimal Markovian coupling with respect to non-negative lower semi-continuous functions. *Acta Mathematica Sinica* **16** 261–270.