Convergence Rate Bounds for Markov Chains Using One-Shot Coupling

Sabrina Sixta and Jeffrey S. Rosenthal

April 15, 2022

The study of Markov chain convergence rates focuses on evaluating how fast a positive recurrent Markov chain converges to its stationary distribution in total variation distance. On one hand, a great deal of progress has been made in bounding the convergence rate for Markov chains defined in discrete state spaces [14]. On the other hand, despite the major developments made in bounding Markov chains in a continuous state space, many applications of continuous state space Markov chains do not have established convergence rate bounds. For example, convergence rate bounds related to Markov chain Monte Carlo (MCMC) models are useful for deciding the size of the burn-in period [5, 4], but many applied MCMC models do not have known upper bounds on their convergence rate [4]. Instead, users rely on ad-hoc convergence diagnostics (e.g. [3]), which offer no guarantees.

Methods using the drift and minorization conditions (eg. [13, 1]), which guarantee geometric ergodicity, are the most studied techniques for bounding Markov chains in continuous state space [11, 5]. Despite the widespread use of bounds generated by the drift and minorization conditions, there are drawbacks. Exogenous variables must be proposed to generate the bounds and bounds using this method do not scale well in high dimensions [10].

One-shot coupling, which was first defined in [12], can provide an upper bound on the convergence rate of a Markov chain while not needing to identify any exogenous sets or functions, and it scales well in high dimensions. One-shot coupling attempts to bring the two Markov chains close together, in what's generally called the 'contracting' phase, and only tries to couple the chain at the last iteration, in the 'coalescing' phase. During the contracting phase the two copies of a Markov chain merge closer together over some predefined metric.

The one-shot coupling method has been applied over a variety of specific examples, namely, a nested gamma model in [7], an image restoration model in [6], and a random walk on the unit sphere in [9]. The method is also used as motivation for a theorem that bounds total variation distance in terms of the Wasserstein distance [8]. This paper looks at developing a general method for directly bounding the total variation distance between two Markov chains using one-shot coupling.

For more details on the following results including proofs, commentary and additional examples, refer to [15].

Notation

We define a Markov chain in terms of iterated random functions [2]. That is, define a family of random functions $\{f_\theta : \theta \in \Theta\}$ such that θ is a random variable and $X_n = f_{\theta_n}(X_{n-1})$. The nth iteration of the Markov chain can be written in terms of $X_0 = x$, as $X_n = (f_{\theta_n} \circ f_{\theta_{n-1}} \ldots \circ f_{\theta_1})(x)$. The stationary distribution of $\{X_n\}_{n\geq 1}$ is denoted as π.

The total variation distance between the laws of two random variables, X and X', defined on the state space $\mathcal{X}$ is $\|\mathcal{L}(X) - \mathcal{L}(X')\| = \sup_{A\subseteq\mathcal{X}} |P(X \in A) - P(X' \in A)|$, where $\mathcal{L}(X)$ represents the law of the random variable X and A is a measurable set. The Markov chain is geometrically ergodic if there exists a $\rho < 1$ and a function $M(x) < \infty$, π-a.e. such that for $X_0 = x$, $\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\| \leq M(x)\rho^n$.

Our goal is to generate geometrically ergodic bounds in total variation distance for Markov chains that can be written as an iterated random function.

Main result

The one-shot coupling method first described in [12] is summarised below. To find an upper bound on the total variation distance between X_N and X'_N we do the following.

1. **Contracting phase:** For $n < N$, set $\theta_n = \theta'_n$ so that the two chains get 'closer' together.

2. **Coalescing phase:** For $n = N$, we specify $j \in \{1, \ldots, |\theta_n|\}$ and set $\theta_{i,n} = \theta'_{i,n}$ for all $i \neq j$. Assume that $j = 1$. We are then left with the random mappings $X_n = f_{(\theta_{1,n},\theta_{-1,n})}(X_{n-1})$ and $X'_n = f_{(\theta_{1,n},\theta'_{-1,n})}(X'_{n-1})$ where X_{n-1} and X'_{n-1} are close to each other in expectation. We apply coupling techniques to find the probability that they are equal.

We propose a general theorem that summarizes the one-shot coupling method for bounding the total variation distance between two copies of a Markov chain.

Theorem (One-Shot Coupling Theorem)**.** *Let $\{X_n\}_{n\geq 1}, \{X'_n\}_{n\geq 1}$ be two copies of a Markov chain such that $X_n = f_{\theta_n}(X_{n-1})$ and $X'_n = f_{\theta'_n}(X'_{n-1})$ where $(\theta_n, \theta'_n)_{n\geq 1}$ are independent random variables with respect to n and the marginal distribution of $\theta_n, \theta'_n \sim \mathcal{D}$, for some distribution $\mathcal{D}$. Suppose that the following two conditions hold for some non-negative integer n_0.*

1. ***Contraction condition:*** *There exists a $D \in (0,1)$ such that for any $n \geq n_0$ when $\theta_{n+1} = \theta'_{n+1} \sim \mathcal{D}$*

$$E[|f_{\theta_{n+1}}(X_n) - f_{\theta_{n+1}}(X'_n)|] \leq DE[|X_n - X'_n|]$$

2. ***Coalescing condition:*** *There exists a $C > 0$ such that for any $n \geq n_0$*

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1}))|| \leq CE[|X_n - X'_n|]$$

Then the total variation distance between the two Markov chains at iteration $n \geq n_0$ is

$$||\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})|| \leq CD^{n-n_0}E[|X_{n_0} - X'_{n_0}|]$$

Using the above theorem, we find the convergence rate upper bound for two families of Markov chains, the random-functional autoregressive process and the randomly scaled iterated random process.

Random-functional autoregressive processes

Random-functional autoregressive processes, $\{X\}_{n\geq 1}$, are of the following form for $g : \mathbb{R}^2 \to \mathbb{R}$

$$X_n = g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} \tag{1}$$

where $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$ are random and $(\theta_{1,n}, \theta_{2,n}) \perp\!\!\!\perp (\theta_{1,m}, \theta_{2,m})$ when $n \neq m$.

We propose the Sideways theorem, which provides an upper bound on the total variation distance for random-functional autoregressive processes.

Theorem (Sideways Theorem)**.** *Let $X_n \in \mathbb{R}$ be a random-functional autoregressive process. That is, X_n is of the following form for $g : \mathbb{R}^2 \to \mathbb{R}$*

$$X_n = g(\theta_{1,n}, X_{n-1}) + \theta_{2,n} \tag{2}$$

where $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$ are random variables and $(\theta_{1,n}, \theta_{2,n}) \perp\!\!\!\perp (\theta_{1,m}, \theta_{2,m})$ when $n \neq m$. Suppose that,

1. ***Contraction condition:*** *There exists a $D \in (0,1)$ such that for $n \geq 0$,*

$$E[|g(\theta_{1,n+1}, X_n) - g(\theta_{1,n+1}, X'_n)|] \leq DE[|X_n - X'_n|]$$

2. ***Attributes of the conditional density $\theta_{2,n}|\theta_{1,n}$:*** *The conditional density of $\theta_{2,n}|\theta_{1,n}$*
   (a) *is bounded above: There exists a $K > 0$ such that for all $(\theta_{1,n}, \theta_{2,n}) \in \mathbb{R}^2$, the conditional density function of $\theta_{2,n}$ is bounded above by K, $f_{\theta_{2,n}}(\theta_{2,n}|\theta_{1,n}) \leq K$.*
   (b) *has at most M local extrema points that are at most $L > 0$ distance apart: For any $\theta_{1,n}$, there are M local maximas and minimas (local extrema points) within the conditional density. The local extrema points are at most L distance apart.*
   (c) *is continuous for any $\theta_{1,n}$*

Then an upper bound on the geometric rate of convergence of the Markov chain is D and the total variation distance between the two copies of the Markov chain, X_n, X'_n, is bounded above as follows,

$$\|\mathcal{L}(X_{n+1}) - \mathcal{L}(X'_{n+1})\| \leq \left(\frac{K(M+1)}{2} + \frac{I_{M>1}}{L}\right) D^n E[|X_0 - X'_0|] \tag{3}$$

The attributes of the conditional density of $\theta_{2,n}|\theta_{1,n}$ serve to prove that the coalescing condition is satisfied. For many examples, conditions 2b and 2c of the Sideways theorem are easily verified if $\theta_{2,n}$ has a defined continuous density.

The following table summarizes various random-functional autoregressive processes that use the Sideways theorem to provide an upper bound on the convergence rate. Refer to [15] for a comparison of the following results to other geometric convergence bounds.

Process	**Upper bound on total variation**
An example of a non-linear autoregressive process $X_n = \frac{1}{2}(X_{n-1} - \sin X_{n-1}) + Z_n$, $Z_n \sim N(0,1)$	$\|\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|\| \leq \sqrt{\frac{2}{3\pi}} E[\|X_0 - X'_0\|] 0.669^{\lfloor (n-1)/2 \rfloor}$
Bayesian regression Gibbs sampler Suppose we have the following observed data $Y \in \mathbb{R}^k$ and $X \in \mathbb{R}^{k \times p}$ where $Y\|\beta, \sigma^2 \sim N_k(X\beta, \sigma^2 I_k)$ for unknown parameters $\beta \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}$. We apply the prior distributions on the unknown parameters, • $\beta\|\sigma^2 \sim N_p(0_p, \frac{\sigma^2}{\lambda} I_p)$, where $\lambda > 0$ is known • $\pi(\sigma^2) \propto 1/\sigma^2$ The Markov chain of interest is on $\beta_n\|\sigma^2_{n-1}, Y$ and $\sigma^2_n\|\beta_n, Y$, which converges to the posterior distribution, $\pi(\beta, \sigma^2\|Y)$	$\\|\mathcal{L}(\beta_n, \sigma_n) - \mathcal{L}(\beta'_n, \sigma'^2_n)\\| \leq KE[\|\sigma_0^2 - \sigma'^2_0\|] \left(\frac{p}{k+p-2}\right)^{n-1}$ where $K = \frac{(C/2)^{\frac{k+2p}{2}}}{\Gamma(\frac{k+2p}{2})} \left(\frac{k+2p+2}{C}\right)^{\frac{k+2p}{2}+1} e^{-\frac{k+2p+2}{2}}$.
Bayesian location model Gibbs sampler Suppose that we are given data points $Y_1, \ldots, Y_J \sim N(\mu, \tau^{-1})$ where μ, τ^{-1} are unknown and $J \geq 3$. Let μ, τ^{-1} have flat priors on $\mathbb{R}$ and $\mathbb{R}_+$. The Markov chain of interest is on $\mu_n\|\tau^{-1}_{n-1}, Y$ and $\tau^{-1}_n\|\mu_n, Y$ which converges to the posterior distribution, $\pi(\mu, \tau^{-1}\|Y)$	$\\|\mathcal{L}(\mu_n, \tau_n^{-1}) - \mathcal{L}(\mu'_n, \tau'^{-1}_n)\\| \leq KE[\|\tau_0^{-1} - \tau'^{-1}_0\|] \left(\frac{1}{J}\right)^{n-1}$ where $K = \frac{(S/2)^{\frac{J-1}{2}}}{\Gamma(\frac{J-1}{2})} \left(\frac{S}{J+1}\right)^{-\frac{J-3}{2}} e^{-\frac{J+1}{2}}$.
Autoregressive normal process in $\mathbb{R}^d$ $\vec{X}_n = A\vec{X}_{n-1} + \vec{Z}_n$, $\vec{Z}_n \sim N(\vec{0}, \Sigma_d^2)$ where Σ_d^2 is a positive semi-definite matrix and A is a diagonalizable matrix.	$\\|\mathcal{L}(\vec{X}_n) - \mathcal{L}(\vec{X}'_n)\\| \leq \sqrt{\frac{d}{2\pi}} \\|\Sigma_d^{-1}\\|_2 \cdot \\|P\\|_2 \\|P^{-1}\\|_2 E[\\|\vec{X}_0 - \vec{X}'_0\\|_2] \max_{1 \leq i \leq d} \|\lambda_i\|^n$ where $A = PDP^{-1}$ is the eigendecomposition, λ_i is the ith eigenvalue of A and $\\|\cdot\\|_2$ denotes the Frobenius norm.

Randomly scaled iterated random functions

Randomly scaled iterated random functions are of the following form for $f : \mathbb{R}^2 \to \mathbb{R}$,

$$X_n = f(\theta_{1,n}, X_{n-1})\theta_{2,n}$$

Where $(\theta_{1,n}, \theta_{2,n})$ are random variables that are i.i.d. with respect to n.

The following table summarizes the application of the one-shot coupling theorem to provide an upper bound on the convergence rate for various randomly scaled iterated random functions. In each example $\{X_n\}_{n \geq 1}$ is the Markov chain we are bounding and the variables $\alpha, \beta, \gamma \in \mathbb{R}$ are constants. $\{Z_n\}_{n \geq 1}$ are i.i.d. random variables with $Z_n > 0$ a.s. and the density of $\log(Z_0)$ is bounded above, has at most M local maxima and minima and is continuous. $\{W_n\}_{n \geq 1}$ are i.i.d. random variables with the density of W_n centred at zero and is monotonically decreasing around zero (like the normal distribution).

Process	**Upper bound on total variation**
Linear ARCH model $X_n = (\beta_0 + \beta X_{n-1})Z_n$, $\alpha, \beta > 0$.	$\|\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|\| \leq \frac{\beta(M+1)}{2\alpha} \sup_x e^x f_{Z_n}(e^x) D^{n-1} E[\|X_0 - X'_0\|]$ where $D = \beta E[Z_0]$.
Asymmetric ARCH model $X_n = \sqrt{(\alpha X_{n-1} + \beta)^2 + \gamma^2} W_n$, where $\alpha > 0$.	$\|\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|\| \leq \frac{\|\alpha\|}{\gamma} D^{n-1} E[\|X_0 - X'_0\|]$ where $D = \|\alpha\| E[\|Z_0\|]$.
GARCH(1,1) model $X_n = \sigma_n W_n$, where $\sigma_n^2 = \alpha^2 + \beta^2 X_{n-1}^2 + \gamma^2 \sigma_{n-1}^2$.	$\|\|\mathcal{L}(X_n) - \mathcal{L}(X'_n)\|\| \leq \frac{D^{n-1}}{\alpha} \sqrt{\beta^2 \|x_0^2 - x_0'^2\| + \gamma^2 \|\sigma_0^2 - \sigma_0'^2\|}$ where $D = \sqrt{\beta^2 E[Z_0^2] + \gamma^2}$.

References

[1] Peter H. Baxendale. “Renewal theory and computable convergence rates for geometrically ergodic Markov chains”. In: *The Annals of Applied Probability* 15.1B (2005), pp. 700–738. DOI: 10.1214/105051604000000710.

[2] Persi Diaconis and David Freedman. “Iterated random functions”. In: *SIAM Review* 41.1 (1999), pp. 45–76. DOI: 10.1137/S0036144598338446.

[3] A. Gelman and D.B. Rubin. “Inference from Iterative Simulation using Multiple Sequences”. In: *Statistical Science* 7.4 (1992), pp. 457–472. DOI: 10.1214/ss/1177011136.

[4] Charles J. Geyer. “Introduction to Markov Chain Monte Carlo”. In: *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall/CRC, 2011, pp. 1–46. DOI: 10.1201/b10905.

[5] James P. Hobert and Galin L. Jones. “Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo”. In: *Statistical Science* 16.4 (2001), pp. 312–334. DOI: 10.1214/ss/1015346317.

[6] Oliver Jovanovski. “Convergence bound in total variation for an image restoration model”. In: *Statistics & Probability Letters* 90 (2014), pp. 11–16. DOI: 10.1016/j.spl.2014.03.007.

[7] Oliver Jovanovski and Neal Madras. “Convergence rates for a hierarchical Gibbs sampler”. In: *Bernoulli* 1.23 (2013), pp. 603–625. DOI: 10.3150/15-BEJ758.

[8] Neal Madras and Denis Sezer. “Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances”. In: *Bernoulli* 16.3 (2010), pp. 882–908. DOI: 10.2307/25735016.

[9] Natesh S. Pillai and Aaron Smith. “Kac’s walk on n-sphere mixes in $n \log n$ steps”. In: *The Annals of Applied Probability* 27.1 (2017), pp. 631–650. DOI: 10.1214/16-AAP1214.

[10] Qian Qin and James P. Hobert. *Wasserstein-based methods for convergence complexity analysis of MCMC with applications*. 2020. arXiv: 1810.08826 [math.ST].

[11] Gareth O. Roberts and Jeffrey S. Rosenthal. “General state space Markov chains and MCMC algorithms”. In: *Probability Surveys* 1 (2004), pp. 20–71. DOI: 10.1214/154957804100000024.

[12] Gareth O. Roberts and Jeffrey S. Rosenthal. “One-shot coupling for certain stochastic recursive sequences”. In: *Stochastic Processes and their Applications* 99 (2002), pp. 195–208. DOI: 10.1016/S0304-4149(02)00096-0.

[13] Jeffrey S. Rosenthal. “Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo”. In: *Journal of the American Statistical Association* 90.430 (1995), pp. 558–566. DOI: 10.2307/2291067.

[14] Laurent Saloff-Coste. “Lectures on finite Markov chains”. In: *Lectures on Probability Theory and Statistics: Ecole d’Eté de Probabilités de Saint-Flour XXVI-1996*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 301–413. DOI: 10.1007/BFb0092621.

[15] Sabrina Sixta and Jeffrey S. Rosenthal. *Convergence rate bounds for iterative random functions using one-shot coupling*. 2021. arXiv: 2112.03982 [stat.CO].