Finite-Run Certification for Approximate Metropolis–Hastings: Bounds and Limits

Antonino Ougein[1] **Siddharth Nand**[1] **Ziya ad-Din Gabriella**[1]

[1]Department of Statistics, University of British Columbia, Vancouver, Canada, Corresponding author: antonino.ougein@ubc.ca

Abstract

When can a finite Metropolis–Hastings run be certified as close to its target? Under a known Wasserstein contraction and an analytic one-step error bound, we give a law-level guarantee for every run horizon. A common-proposal coupling converts acceptance error into the one-step term, and a potential–residual split separates transient error from stationary target shift. Conversely, two irreducible two-state chains can have almost indistinguishable short paths but well-separated invariant laws. The certificate is therefore useful only when its structural constants are known independently of the observed trajectory.

1 INTRODUCTION

Convergence diagnostics estimate properties of a realized simulation, but a finite chain can fail to reveal important parts of its invariant distribution. This paper considers a different object: a deterministic bound on the distance between the law of a finite run and the intended target. The bound is based on analytic control of the exact chain's contraction and the one-step discrepancy introduced by an approximate transition rule.

The result is intentionally assumption-based. To obtain a numerical certificate, the contraction rate, initial distance, and kernel error bound must be known from the model and algorithm. The theorem does not estimate these quantities from the simulated path. This avoids statistical calibration, but also makes explicit when a certificate cannot be obtained.

2 RELATED LITERATURE

Quantitative convergence of Markov chains is usually established from drift and minorization conditions, spectral information, or explicit couplings. Classical MCMC results give geometric rates for Metropolis-type kernels and general-state-space chains [Mengersen and Tweedie, 1996, Roberts and Rosenthal, 1997, Rosenthal, 1995, Roberts and Rosenthal, 2004]. Meyn and Tweedie's treatment and Baxendale's computable-rate results make clear that obtaining constants is a separate task from proving qualitative ergodicity [Meyn and Tweedie, 2009, Baxendale, 2005]. For uniformly ergodic kernels, Mitrophanov gives sensitivity bounds under perturbations [Mitrophanov, 2005]; Wasserstein perturbation theory extends this style of comparison under contraction assumptions [Rudolf and Schweizer, 2018]. Noisy-MCMC analyses study related approximate transitions, including Monte Carlo within Metropolis [Alquier et al., 2016, Medina-Aguayo et al., 2020]. Our recursion is a direct contractive specialization, with its one-step error expressed in the acceptance probabilities of the two MH kernels.

Convergence diagnostics answer a different question. Multiple-chain diagnostics beginning with Gelman and Rubin, later monitoring recommendations, and comparative reviews assess convergence from simulated output [Gelman and Rubin, 1992, Brooks and Gelman, 1998, Cowles and Carlin, 1996]. Rank-normalized $\widehat{R}$ and multivariate output analysis improve practical diagnostics and Monte Carlo error assessment [Vehtari et al., 2021, Vats et al., 2019]; classical discussions also address chain behavior and run length [Geyer, 1992, Flegal and Jones, 2010]. These tools are valuable for analyzing a run, but do not by themselves provide a deterministic upper bound on the law of the chain at a specified time. Conversely, our result is a law-level bound and does not provide a confidence interval for a realized time average.

Coupling supplies both a proof technique and a family of simulation methods. Standard coupling texts develop the

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 286:98–102.

maximal-coupling and transport arguments used to compare transition laws [Lindvall, 1992, Thorisson, 2000]. L-lag couplings turn meeting events into data-dependent non-asymptotic estimates of finite-time error, including comparisons between exact and approximate MCMC [Biswas et al., 2019]. Coupled unbiased MCMC instead uses a telescoping correction to remove bias from selected estimators [Jacob et al., 2020]. Both approaches use a realized coupled trajectory; the certificate here is analytic once contraction and one-step constants are supplied.

For continuous-state chains, reflection couplings can establish Wasserstein contraction for diffusions under suitable conditions, while spectral methods give finite-state and reversible-chain convergence information [Eberle, 2016, Fill, 1991]. Recent convergence tools are complementary, not automatic inputs: a diffusion contraction theorem or a spectral-gap estimate does not verify the required Wasserstein contraction for an arbitrary Metropolis–Hastings proposal.

Finally, the two-state lower bound is related in spirit, but not in scope, to impossibility results for approximate MCMC such as the subsampling-focused no-free-lunch analysis of Johndrow et al. [2020]. Our construction concerns a short realized path and the invariant law itself; it shows why analytic structural assumptions cannot be replaced by path inspection alone.

3 THE DISTRIBUTIONAL RECURSION

Let (X, d) be a Polish metric space. For probability measures with finite first moment, let W_1 denote the Wasserstein distance with cost d. Let P be an exact Markov kernel with invariant distribution π, and let $\widetilde{P}$ be an approximate kernel. Assume the exact kernel is contractive:

$$W_1(P_x, P_{x'}) \le \rho\, d(x, x'), \qquad 0 \le \rho < 1,$$

where $P_x = P(x, \cdot)$. Assume also the one-step discrepancy is uniformly bounded:

$$\sup_x W_1(\widetilde{P}(x, \cdot), P(x, \cdot)) \le b.$$

For an initial law μ_0, write $\mu_t = \mu_0 \widetilde{P}^t$ and assume $W_1(\mu_0, \pi) \le D_0 < \infty$.

Theorem 1 (Finite-time law certificate)**.** *For every integer $t \ge 0$,*

$$W_1(\mu_t, \pi) \le \rho^t D_0 + b\,\frac{1-\rho^t}{1-\rho}.$$

Consequently, for every L-Lipschitz function f for which the expectations exist,

$$\left| \int f\, d\mu_t - \int f\, d\pi \right| \le L\left(\rho^t D_0 + b\,\frac{1-\rho^t}{1-\rho}\right).$$

Proof. By the triangle inequality, the one-step bound, and contraction of P,

$$\begin{aligned} W_1(\mu_{t+1}, \pi) &= W_1(\mu_t \widetilde{P}, \pi P) \\ &\le W_1(\mu_t \widetilde{P}, \mu_t P) + W_1(\mu_t P, \pi P) \\ &\le b + \rho\, W_1(\mu_t, \pi). \end{aligned}$$

Iterating this recursion from $W_1(\mu_0, \pi) \le D_0$ gives the stated geometric sum. The expectation bound follows from the Kantorovich–Rubinstein inequality. □

Corollary 2 (Sufficient run horizon)**.** *Let $B = b/(1-\rho)$ and suppose $0 < \rho < 1$. The bound in Theorem 1 is $B + \rho^t(D_0 - B)$. If $D_0 > B$ and $\epsilon > B$, it is at most ϵ whenever*

$$t \ge \max\left\{0, \left\lceil \frac{\log((D_0 - B)/(\epsilon - B))}{-\log \rho} \right\rceil \right\}.$$

If $D_0 > B$ and $\epsilon \le B$, this bound cannot reach the tolerance at any finite horizon. If $D_0 \le B$, the displayed certificate is smallest at $t = 0$; further iterations do not improve this particular bound.

Proof. The theorem's right-hand side equals $B + \rho^t(D_0 - B)$. For $D_0 > B$, solving this expression's inequality with ϵ gives the stated horizon; it remains strictly above B at every finite t. For $D_0 \le B$, it is nondecreasing in t, so its minimum is the initial bound D_0. □

Theorem 3 (Finite-time total-variation certificate)**.** *Define the Dobrushin coefficient*

$$\delta(P) = \sup_{x,x'} \mathrm{TV}(P(x, \cdot), P(x', \cdot)).$$

Suppose $\delta(P) \le \tau < 1$ and $\sup_x \mathrm{TV}(\widetilde{P}(x, \cdot), P(x, \cdot)) \le b_{\mathrm{TV}}$. Then, for $\mu_t = \mu_0 \widetilde{P}^t$ and invariant law π of P,

$$\mathrm{TV}(\mu_t, \pi) \le \tau^t\, \mathrm{TV}(\mu_0, \pi) + b_{\mathrm{TV}}\,\frac{1-\tau^t}{1-\tau}.$$

Proof. Dobrushin contraction gives $\mathrm{TV}(\mu P, \nu P) \le \tau\, \mathrm{TV}(\mu, \nu)$ for probability measures μ, ν. The one-step error assumption gives $\mathrm{TV}(\mu \widetilde{P}, \mu P) \le b_{\mathrm{TV}}$ for every μ. Therefore

$$\mathrm{TV}(\mu_{t+1}, \pi) \le b_{\mathrm{TV}} + \tau\, \mathrm{TV}(\mu_t, \pi).$$

Iteration of this recursion yields the stated geometric sum. □

Remark 4 (What is certified)**.** *Theorem 1 bounds the distribution of the chain at time t, and hence the bias of expectations of Lipschitz functions. It is not a concentration bound for a time average or a guarantee about the error of one realized output. A certificate can be evaluated without running the chain if valid values of ρ, b, and D_0 are available. If these constants are unknown, the theorem alone does not produce a numerical certificate.*

4 CHARGING MH ACCEPTANCE ERROR

Suppose P and $\widetilde{P}$ use the same proposal kernel Q and acceptance probabilities $a(x,y)$ and $\widetilde{a}(x,y)$, respectively. Couple the transitions by drawing the same proposal $Y \sim Q(x,\cdot)$ and the same uniform random variable for both acceptance decisions.

Proposition 5 (Common-proposal discrepancy)**.** *For every state x,*

$$W_1(\widetilde{P}_x, P_x) \le \int d(x,y)\,|\widetilde{a}(x,y) - a(x,y)|\,Q(x,dy),$$

where $\widetilde{P}_x = \widetilde{P}(x,\cdot)$. The same coupling also gives

$$\mathrm{TV}(\widetilde{P}_x, P_x) \le \int |\widetilde{a}(x,y) - a(x,y)|\,Q(x,dy).$$

If the log acceptance ratios differ by $\varepsilon(x,y)$, so that $a(x,y) = h(\ell(x,y))$ and $\widetilde{a}(x,y) = h(\ell(x,y) + \varepsilon(x,y))$ with $h(s) = 1 \wedge e^s$, then

$$W_1(\widetilde{P}_x, P_x) \le \int d(x,y)\,\min\{1, |\varepsilon(x,y)|\}\,Q(x,dy).$$

Proof. Under the common-randomness coupling, the two next states differ only when one chain accepts the proposal and the other rejects it. Given $Y = y$, this event has probability $|\widetilde{a}(x,y) - a(x,y)|$, and the distance between the resulting states is $d(x,y)$. The mismatch probability bounds total variation, and the expected coupling cost proves the Wasserstein inequality. The function h is 1-Lipschitz and takes values in $[0,1]$, so the acceptance-probability difference is at most $\min\{1, |\varepsilon(x,y)|\}$. □

Combining Proposition 5 with Theorem 1 yields a deterministic bound whenever

$$b \;\ge\; \sup_x \int d(x,y)\,\min\{1, |\varepsilon(x,y)|\}\,Q(x,dy)$$

can be bounded analytically. For a point initialization x_0, one may take $D_0 = \int d(x_0,z)\,\pi(dz)$ when this target moment is known to be finite and bounded. These conditions are substantive: strong log-concavity alone does not automatically verify every required contraction and proposal-displacement constant for every MH variant.

Lemma 6 (Sharp bound for a bounded target tilt)**.** *If ψ is bounded and $\pi_\psi \propto e^\psi \pi$, then*

$$\mathrm{TV}(\pi_\psi, \pi) \le \tanh\!\left(\frac{\mathrm{osc}(\psi)}{4}\right).$$

Proof. Let $g = d\pi_\psi/d\pi$, $a = \operatorname{ess\,inf}_\pi g$, and $b = \operatorname{ess\,sup}_\pi g$. Then $\int g\,d\pi = 1$ and $b/a \le e^{\mathrm{osc}(\psi)}$. Convexity of $|t-1|$ on $[a,b]$ yields

$$\mathrm{TV}(\pi_\psi, \pi) \le \frac{(1-a)(b-1)}{b-a}.$$

If $a = b = 1$, the distance is zero. Otherwise set $r = b/a$ and write $b = ra$. The displayed upper bound is maximized over $a \in [1/r, 1]$ at $a = r^{-1/2}$, where it equals $(\sqrt{r}-1)/(\sqrt{r}+1) = \tanh(\log(r)/4)$. Since $r \le e^{\mathrm{osc}(\psi)}$, the lemma follows. □

5 SEPARATING TARGET SHIFT FROM TRANSITION ERROR

Theorem 7 (Finite-time potential–residual certificate)**.** *Suppose the approximate acceptance log ratio is $\ell(x,y) + \psi(y) - \psi(x) + r(x,y)$, and let P^ψ be the exact MH kernel for $\pi_\psi \propto e^\psi \pi$, and let $\widetilde{P}$ be the MH kernel using the approximate log ratio. Assume*

$$a^\psi(x,y) = h(\ell(x,y) + \psi(y) - \psi(x))\,,$$
$$\widetilde{a}(x,y) = h(\ell(x,y) + \psi(y) - \psi(x) + r(x,y))\,.$$

Assume also

$$W_1(P^\psi_x, P^\psi_{x'}) \le \rho\, d(x,x'), \qquad 0 \le \rho < 1,$$

and define

$$b_r = \sup_x \int d(x,y)|\widetilde{a}(x,y) - a^\psi(x,y)|\,Q(x,dy).$$

If $W_1(\mu_0, \pi_\psi) \le D_0$ and $B_\psi = W_1(\pi_\psi, \pi) < \infty$, then for all $t \ge 0$,

$$W_1(\mu_0 \widetilde{P}^t, \pi) \le \rho^t D_0 + b_r \frac{1-\rho^t}{1-\rho} + B_\psi.$$

Moreover,

$$b_r \le \sup_x \int d(x,y)\,\min\{1, |r(x,y)|\}\,Q(x,dy).$$

If the metric space has finite diameter D, then $B_\psi \le D\tanh(\mathrm{osc}(\psi)/4)$.

Proof. The shared-proposal coupling in Proposition 5, applied to $\widetilde{P}$ and P^ψ, bounds their one-step Wasserstein distance at state x by the integrand defining b_r. The triangle-inequality recursion relative to P^ψ therefore gives

$$W_1(\mu_0 \widetilde{P}^t, \pi_\psi) \le \rho^t D_0 + b_r \frac{1-\rho^t}{1-\rho}.$$

Adding B_ψ proves the first assertion. The range and 1-Lipschitz property of h give the residual bound. If the diameter is D, coupling the common mass of π_ψ and π gives $W_1(\pi_\psi, \pi) \le D\ \mathrm{TV}(\pi_\psi, \pi)$; Lemma 6 supplies the stated estimate. □

Corollary 8 (Pure potential error)**.** *If $r = 0$, then $\widetilde{P} = P^\psi$ and*

$$W_1(\mu_0 \widetilde{P}^t, \pi) \le \rho^t W_1(\mu_0, \pi_\psi) + W_1(\pi_\psi, \pi).$$

On a metric space of diameter D, the stationary term is at most $D\tanh(\mathrm{osc}(\psi)/4)$.

Proof. When $r = 0$, the one-step residual term b_r vanishes. The conclusion follows from Theorem 7 and Lemma 6. □

6 A SHORT PATH CANNOT REVEAL EVERY MODE

The following finite-state construction illustrates why a finite path cannot, without structural restrictions, identify the invariant law. Fix a horizon $T \geq 1$ and $0 < \alpha < 1$, and observe states at times $0, 1, \ldots, T$. On $\mathsf{X} = \{0, 1\}$ with distance $d(0, 1) = 1$, define two chains, both initialized at 0, with $q = \alpha/(2T)$. In both chains the transition probability from 0 to 1 is q. Chain A returns from 1 to 0 with probability q; chain B returns with probability q^2. Both chains are irreducible and aperiodic, with

$$\pi_A(\{1\}) = \frac{1}{2}, \qquad \pi_B(\{1\}) = \frac{1}{1+q}.$$

Proposition 9 (Trajectory indistinguishability)**.** *The total-variation distance between the length-T path laws of chains A and B is at most $\alpha/2$, while*

$$\mathrm{TV}(\pi_A, \pi_B) = \frac{1-q}{2(1+q)}.$$

Thus, for any estimator of the invariant law based only on that finite path, at least one of the two chains has probability at least $(1 - \alpha/2)/2$ of an estimation error at least $(1 - q)/[4(1 + q)]$ in total variation.

Proof. Couple the chains to agree until the first visit to state 1, using the same randomness for their common transition from 0. Their path laws can differ only if state 1 is visited by time T, an event with probability at most $1 - (1 - q)^T \leq Tq = \alpha/2$. The stationary probability of state 1 for a two-state chain with transition probabilities p_{01} and p_{10} is $p_{01}/(p_{01} + p_{10})$, which gives the displayed invariant laws.

For the estimation claim, the two target laws are separated by $D = (1 - q)/(2(1 + q))$. Let E be the event that the estimator is within $D/2$ of π_A. Under chain A, an error of at least $D/2$ occurs on E^c. Under chain B, on E the triangle inequality forces an error of at least $D/2$ relative to π_B. The sum of these two error probabilities is at least one minus the total-variation distance between the path laws, and is therefore at least $1 - \alpha/2$. At least one of the probabilities is at least half that amount. □

7 SCOPE OF THE GUARANTEE

For fixed $\rho < 1$, Theorem 1 has limiting bound $b/(1 - \rho)$ as $t \to \infty$. The decomposed result has limiting contribution $b_r/(1 - \rho) + B_\psi$. These are bounds on the law at a fixed time and, through Wasserstein duality, on expectations of Lipschitz functions.

They are not concentration bounds for a time average and do not certify a realized sample mean. The contraction rate, initial distance, and one-step error must be proved or bounded from model assumptions. Establishing useful constants for a named MCMC algorithm is the remaining step between these inequalities and an operational certificate.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors declare no competing interests.

References

Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy Monte Carlo: convergence of markov chains with approximate transition kernels. *Statistics and Computing*, 26(1–2):29–47, 2016. doi: 10.1007/s11222-014-9521-x.

Peter H. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic markov chains. *The Annals of Applied Probability*, 15(1B):700–738, 2005. doi: 10.1214/105051604000000710.

Niloy Biswas, Pierre E. Jacob, and Paul Vanetti. Estimating convergence of markov chains with L-lag couplings, 2019.

Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4): 434–455, 1998. doi: 10.1080/10618600.1998.10474787.

Mary Kathryn Cowles and Bradley P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996. doi: 10.1080/01621459.1996.10476956.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166 (3–4):851–886, 2016. doi: 10.1007/s00440-015-0673-1.

James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process. *The Annals of Applied Probability*, 1(1):62–87, 1991. doi: 10.1214/aoap/1177005981.

James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in markov chain monte carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010. doi: 10.1214/09-AOS735.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ss/1177011136.

Charles J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992. doi: 10.1214/ss/1177011138.

Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020. doi: 10.1111/rssb.12336.

James E. Johndrow, Natesh S. Pillai, and Aaron Smith. No free lunch for approximate MCMC, 2020.

Torgny Lindvall. *Lectures on the Coupling Method*. Wiley, 1992.

Felipe Medina-Aguayo, Daniel Rudolf, and Nikolaus Schweizer. Perturbation bounds for monte carlo within metropolis via restricted approximations. *Stochastic Processes and their Applications*, 130(4):2200–2227, 2020. doi: 10.1016/j.spa.2019.06.015.

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the hastings and metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996. doi: 10.1214/aos/1033066201.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2 edition, 2009.

A. Yu. Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Gareth O. Roberts and Jeffrey S. Rosenthal. Geometric ergodicity and hybrid markov chains. *Electronic Communications in Probability*, 2:13–25, 1997. doi: 10.1214/ECP.v2-981.

Gareth O. Roberts and Jeffrey S. Rosenthal. General state space markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004. doi: 10.1214/154957804100000024.

Jeffrey S. Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995. doi: 10.1080/01621459.1995.10476548.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for markov chains via wasserstein distance. *Bernoulli*, 24 (4A):2610–2639, 2018. doi: 10.3150/17-BEJ938.

Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, 2000.

Dootika Vats, James M. Flegal, and Galin L. Jones. Multivariate output analysis for markov chain monte carlo. *Biometrika*, 106(2):321–337, 2019. doi: 10.1093/biomet/asz002.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi: 10.1214/20-BA1221.