# Arithmetic Error Budgets in Bayesian Metropolis–Hastings

**Ziya ad-Din Gabriella** and **Siddharth Nand** and **Ortwin Athelstan**
Department of Statistics, University of British Columbia
Vancouver, Canada
ziya.gabriella@ubc.ca

## Abstract

If a Bayesian log density is evaluated by the same deterministic routine at every state, its arithmetic error is a potential: the Metropolis–Hastings chain is exact for a tilted posterior. We quantify that tilt in total variation and propagate a standard summation error bound through it, keeping term-evaluation and accumulator precision separate. A closed-form Gaussian example gives the local $u_n\sqrt{n}$ displacement scale, while a bounded-parameter likelihood example shows which assumptions make a global budget finite. Neither scale is universal: both depend on the numerical pipeline and on uniform control of the likelihood terms.

## 1 Introduction

Statements such as "precision fails when $n$ reaches $1/u$" conflate at least three quantities: the precision used to evaluate a likelihood term, the precision used to accumulate terms, and the magnitude of the terms over the parameter space. These distinctions matter even before mixing is considered. If the same computed log density is reused in both ends of an MH proposal, its error changes the invariant target exactly. If the proposal ratio or accept–reject comparison is evaluated separately, an additional pair-dependent error appears.

We first treat the statewise evaluator, then derive its summation budget and describe the assumptions under which that budget is uniform. The posterior-scale calculation and spectral-gap comparison address different questions: the former is local to the posterior's typical scale, while the latter is a global worst-case statement.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

## 2 Related literature

The arithmetic analysis begins with the standard floating-point model and forward-error bounds. Wilkinson's early treatment, Goldberg's computer-science exposition, and Higham's numerical analysis text describe rounding, cancellation, and conditioning from complementary perspectives (Wilkinson, 1963; Goldberg, 1991; Higham, 2002). For summation specifically, Higham compares several accumulation schemes and emphasizes that no scheme is uniformly best; Ogita, Rump, and Oishi develop accurate sum and dot-product algorithms, while Rump surveys rigorous computation using floating-point arithmetic (Higham, 1993; Ogita et al., 2005; Rump, 2010). The *Handbook of Floating-Point Arithmetic* gives a broader account of formats and operations (Muller et al., 2018). IEEE 754 specifies formats and operation semantics, but it does not bound the state-dependent sum of absolute likelihood contributions used in our theorem (IEEE, 2019). Our contribution in this part is not a new summation algorithm: it is to preserve separate term-evaluation and accumulator errors when translating forward error into a posterior distance.

The target being sampled is defined by the Metropolis construction, from its original statistical-mechanical form through Hastings's generalization and later general-state-space treatments (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994). For large data, a separate literature reduces the cost of the accept–reject calculation by using a subset of observations. Early methods include the sequential-test rule of Korattikara, Chen, and Welling and the adaptive subsampling method of Bardenet, Doucet, and Holmes; Bardenet et al. later survey the assumptions and guarantees of tall-data MCMC (Korattikara et al., 2014; Bardenet et al., 2014, 2017). Firefly Monte Carlo instead augments the state so that only a subset of likeli-

hood factors is queried while retaining the exact posterior as its invariant law (Maclaurin and Adams, 2014). These methods address data access or likelihood evaluation. By contrast, our summation theorem starts from a specified deterministic evaluator and bounds the change in its induced target.

Stochastic-gradient methods form another distinct computational regime. Stochastic gradient Langevin dynamics and stochastic gradient Hamiltonian Monte Carlo use noisy gradient information to make large-data transitions computationally feasible (Welling and Teh, 2011; Chen et al., 2014). Their algorithmic questions concern the effect of gradient noise and discretization on the transition dynamics; the statewise identity studied here applies when a single deterministic log-density value is reused at both ends of a proposal. When an implementation does not have that property, the additional error belongs to approximate-kernel perturbation theory (Rudolf and Schweizer, 2018; Medina-Aguayo et al., 2020).

Our posterior-scale calculation is related to local asymptotic normality and the Bernstein–von Mises theorem (van der Vaart, 1998). Misspecified versions clarify that Gaussian posterior limits can persist under model error, but do not identify floating-point error with statistical misspecification (Kleijn and van der Vaart, 2012). Covariance-based sensitivity methods provide a related way to describe how posterior summaries respond to perturbations (Giordano et al., 2018). Here the local Gaussian tilt is exact under its stated convergence and integrability assumptions; it does not derive those assumptions from a hardware implementation. Practical MCMC output analysis, including run-length and Monte Carlo error assessment, addresses the accuracy of estimates from a chain rather than the target displacement analyzed here (Geyer, 1992; Raftery and Lewis, 1992; Flegal and Jones, 2010).

Finally, the global gap comparison belongs to the literature on stability of reversible chains. Holley–Stroock comparison arguments control bounded changes of measure, while classical results study geometric ergodicity and compare spectral gaps for Metropolis kernels (Holley and Stroock, 1987; Mengersen and Tweedie, 1996; Roberts and Rosenthal, 1997; Diaconis and Saloff-Coste, 1993). The present lower bound is correspondingly worst-case: it does not say that a typical posterior's effective sample size changes by the same factor as its spectral-gap bound.

## 3 Statewise computation and its target

Let $\pi_n$ be a probability measure with unnormalized log density $L_n(x)$. Suppose a fixed deterministic evaluator returns

$$\widehat{L}_n(x) = L_n(x) + \psi_n(x).$$

If the MH acceptance log ratio is formed from differences of these computed statewise values, with the proposal correction evaluated exactly, then the computed acceptance log ratio differs from the exact one by $\psi_n(y) - \psi_n(x)$. The resulting kernel is exactly the MH kernel for

$$\widehat{\pi}_n(dx) = \frac{e^{\psi_n(x)}}{\int e^{\psi_n}\, d\pi_n}\, \pi_n(dx).$$

This identity is independent of the numerical method used to produce $\widehat{L}_n$; only determinism and reuse of the same statewise evaluation are needed.

**Theorem 1** (Target-tilt error budget)**.** *If $\psi_n$ is bounded, then*

$$\mathrm{TV}(\widehat{\pi}_n, \pi_n) \le \tanh\left(\frac{\mathrm{osc}(\psi_n)}{4}\right).$$

*Consequently, for $0 \le \epsilon < 1$, a sufficient condition for $\mathrm{TV}(\widehat{\pi}_n, \pi_n) \le \epsilon$ is*

$$\mathrm{osc}(\psi_n) \le 4\,\mathrm{artanh}(\epsilon).$$

*For any bounded measurable $f$,*

$$|\mathbb{E}_{\widehat{\pi}_n} f - \mathbb{E}_{\pi_n} f| \le 2\|f\|_\infty \tanh\left(\frac{\mathrm{osc}(\psi_n)}{4}\right).$$

*Proof.* The Radon–Nikodym derivative is $g = e^{\psi_n}/\mathbb{E}_{\pi_n} e^{\psi_n}$. Put $a = \operatorname{ess\,inf}_{\pi_n} g$ and $b = \operatorname{ess\,sup}_{\pi_n} g$. Then $\mathbb{E}_{\pi_n} g = 1$ and $b/a \le e^{\mathrm{osc}(\psi_n)}$. Convexity of $|t-1|$ on $[a, b]$ gives

$$\mathrm{TV}(\widehat{\pi}_n, \pi_n) \le \frac{(1-a)(b-1)}{b-a}.$$

If $a = b = 1$, the desired bound is immediate. For $r = b/a$, the right-hand side is at most $(\sqrt{r} - 1)/(\sqrt{r} + 1) = \tanh(\log(r)/4)$, by maximizing over $a \in [1/r, 1]$. This proves the total-variation bound; the expectation bound follows from $|\mathbb{E}_{\widehat{\pi}_n} f - \mathbb{E}_{\pi_n} f| \le 2\|f\|_\infty \mathrm{TV}(\widehat{\pi}_n, \pi_n)$. □

## 4 A floating-point bound for sequential summation

Write the log density as

$$L_n(x) = L_{0,n}(x) + \sum_{i=1}^{n} a_i(x),$$

where $L_{0,n}$ contains prior and other non-summed terms. Let $u_e$ bound the relative error in evaluating each summand: $\widehat{a}_i(x) = a_i(x)(1 + \delta_i(x))$ with $|\delta_i(x)| \le u_e$. Let $u_a$ be the unit roundoff of the accumulator. Assume sequential floating-point additions, no overflow or underflow, and $(n-1)u_a < 1$. Define the summation error by

$$d_n^{\mathrm{sum}}(x) = \mathrm{fl}\left(\sum_{i=1}^{n} \widehat{a}_i(x)\right) - \sum_{i=1}^{n} a_i(x).$$

Formally, assume the returned log density satisfies

$$\widehat{L}_n(x) - L_n(x) = d_n^{\mathrm{sum}}(x) + e_{0,n}(x), \quad |e_{0,n}(x)| \le p_n.$$

The remainder $e_{0,n}$ includes every other arithmetic error, including evaluation of non-summed terms and rounding when partial results are combined. Define

$$\gamma_k(u_a) = \frac{ku_a}{1-ku_a}, \quad c_n = u_e + (1+u_e)\gamma_{n-1}(u_a), \quad S_n(x) = \sum_{i=1}^{n} |a_i(x)|.$$

**Theorem 2** (Deterministic evaluation bound)**.** *Under the assumptions above, the statewise log-density error satisfies*

$$|\psi_n(x)| \le p_n + c_n S_n(x).$$

*If* $\sup_x S_n(x) \le B_n$*, then*

$$\mathrm{TV}(\widehat{\pi}_n, \pi_n) \le \tanh\left(\frac{p_n + c_n B_n}{2}\right).$$

*Proof.* The standard sequential-summation model gives

$$\mathrm{fl}\left(\sum_{i=1}^{n} \widehat{a}_i(x)\right) = \sum_{i=1}^{n} \widehat{a}_i(x)(1+\theta_i(x)), \quad |\theta_i(x)| \le \gamma_{n-1}(u_a).$$

Therefore,

$$\left|\mathrm{fl}\left(\sum_i \widehat{a}_i(x)\right) - \sum_i a_i(x)\right| \le \left[u_e + (1+u_e)\gamma_{n-1}(u_a)\right] S_n(x) = c_n S_n(x).$$

Adding the bound $p_n$ for the remaining log-density terms proves the first claim. A uniform absolute bound $|\psi_n(x)| \le p_n + c_n B_n$ implies $\mathrm{osc}(\psi_n) \le 2p_n + 2c_n B_n$. The second claim follows from Theorem 1. □

**Remark 3** (Why there is no universal *un* threshold)**.** *The arithmetic bound depends separately on term evaluation, $u_e$, and accumulation, $u_a$, as well as on $S_n(x)$. For example, if each $|a_i(x)| \le M$, then $B_n \le nM$, while $\gamma_{n-1}(u_a)$ grows with $nu_a$. Pairwise or compensated summation has different error behavior. Without assumptions controlling $S_n$ uniformly over the state space, the theorem gives no finite global target bound. The threshold is model- and algorithm-dependent, not a function of storage precision and n alone.*

**Proposition 4** (Balanced pairwise summation)**.** *Suppose the n terms are summed by a balanced binary tree of maximum depth $d_n = \lceil \log_2 n \rceil$, with accumulator unit roundoff $u_a$. Under the same assumptions on term evaluation and exponent range as in Theorem 2, its conclusion holds with*

$$c_n^{\mathrm{tree}} = u_e + (1+u_e)\gamma_{d_n}(u_a)$$

*in place of $c_n$.*

*Proof.* Each input term passes through at most $d_n$ rounded additions. Consequently its accumulated value has a relative perturbation bounded by $\gamma_{d_n}(u_a)$. Summing the absolute contributions and adding the term-evaluation error gives $[u_e + (1+u_e)\gamma_{d_n}(u_a)]S_n(x)$. The final target bound follows as in Theorem 2. □

**What the summation bound says about scale.** When $nu_a \to 0$, the sequential coefficient satisfies $\gamma_{n-1}(u_a) = (n-1)u_a + O((nu_a)^2)$, whereas the balanced-tree coefficient has $\gamma_{d_n}(u_a) = d_n u_a + O((d_n u_a)^2)$, with $d_n = \lceil \log_2 n \rceil$. If $|a_i(x)| \le M$ uniformly, then $S_n(x) \le nM$: the resulting worst-case accumulation contributions are of order $u_a n^2 M$ for sequential summation and $u_a nM \log n$ for balanced pairwise summation. The term-evaluation contribution is of order $u_e nM$. These are upper-bound scales under the stated model, not predictions of typical error or universal failure thresholds.

**Cancellation example.** For even $n$, take the deterministic summands $a_i = (-1)^{i+1}$. Their exact sum is zero but $S_n = \sum_i |a_i| = n$. The allowed relative perturbations $\delta_i = u_e(-1)^{i+1}$ give $\sum_i a_i \delta_i = nu_e$. Thus a small or vanishing log-likelihood sum does not by itself control the accumulated absolute error. This worst-case example explains why Theorem 2 uses $S_n(x)$ rather than $|\sum_i a_i(x)|$.

For mixed-precision implementations, the accumulator's format determines $u_a$, while storage and per-term evaluation can affect $u_e$. A low-precision storage type paired with a higher-precision accumulator therefore has different guarantees from using the low-precision type for both operations. The table only describes the

Table 1: Illustrative unit roundoffs for common binary formats. The last column is the scale $1/u$ at which the basic quantity $nu$ is of order one; it is not a universal MCMC failure threshold.

| Format | Unit roundoff $u$ | $1/u$ |
|---|---|---|
| binary16 | $2^{-11}$ | $2.0 \times 10^3$ |
| bfloat16 | $2^{-8}$ | $2.6 \times 10^2$ |
| binary32 | $2^{-24}$ | $1.7 \times 10^7$ |
| binary64 | $2^{-53}$ | $9.0 \times 10^{15}$ |

roundoff scales in the elementary summation model; the state-uniform quantity $B_n$ and the non-summed error bound $p_n$ remain model-specific.

**Bounded Gaussian likelihood.** Consider a scalar Gaussian model with parameter $\theta \in [-R, R]$, observations satisfying $|y_i| \le Y$, and known variance $\sigma^2$. Up to terms independent of $\theta$, write

$$a_i(\theta) = -\frac{(y_i - \theta)^2}{2\sigma^2}.$$

Set $K = (Y + R)^2/(2\sigma^2)$. Then

$$S_n(\theta) = \sum_{i=1}^{n} |a_i(\theta)| \le nK.$$

For sequential accumulation, Theorem 2 gives

$$\mathrm{TV}(\widehat{\pi}_n, \pi_n) \le \tanh\left(\frac{p_n + c_n nK}{2}\right)$$

. For a balanced tree, replace $c_n$ by $c_n^{\mathrm{tree}}$ as in Proposition 4. This is a uniform guarantee under the stated parameter and observation bounds, not a bound for an unbounded parameter space: if $\theta \in \mathbb{R}$, then $\sup_\theta S_n(\theta) = \infty$ and this global argument is vacuous.

## 5 Pair-dependent arithmetic error

The state-potential interpretation requires the acceptance log-ratio to be computed from the same deterministic scalar function at each state. Directly computing a pairwise difference, proposal correction, or acceptance comparison can introduce an extra term $\zeta_n(x, y)$:

$$\widehat{\ell}_n(x, y) - \ell_n(x, y) = \psi_n(y) - \psi_n(x) + \zeta_n(x, y).$$

The first term changes the target; the second need not.

A fixed deterministic reduction tree or summation order still defines a statewise function and therefore contributes to $\psi_n$. In contrast, if the arithmetic result varies between evaluations of the same state because of nondeterministic scheduling or randomized rounding, it cannot be represented by one deterministic $\psi_n$; such variation must be included in the pair-dependent or random transition error.

**Corollary 5** (Budget with a pairwise remainder)**.** *Suppose $|\psi_n(x)| \le E_n$ and $|\zeta_n(x, y)| \le d_n$ for all states and proposed pairs. Suppose the MH kernel for the tilted target $\widehat{\pi}_n$ has Dobrushin coefficient at most $\tau_n < 1$, and the perturbed kernel has an invariant distribution $\widetilde{\pi}_n$. Then*

$$\mathrm{TV}(\widetilde{\pi}_n, \pi_n) \le \tanh\left(\frac{E_n}{2}\right) + \frac{d_n}{1 - \tau_n}.$$

*Proof.* The potential has oscillation at most $2E_n$, so Theorem 1 bounds the distance from $\widehat{\pi}_n$ to $\pi_n$. The function $h(t) = 1 \wedge e^t$ is 1-Lipschitz; hence the uniform bound on $\zeta_n$ bounds the proposal-averaged one-step kernel perturbation by $d_n$. The invariant-kernel perturbation argument under Dobrushin contraction then bounds $\mathrm{TV}(\widetilde{\pi}_n, \widehat{\pi}_n)$ by $d_n/(1 - \tau_n)$. The triangle inequality gives the result. □

## 6 Local posterior bias and global mixing

The global oscillation bound can be conservative for a concentrated posterior: a large, nearly constant log-density error may have little effect near the posterior's typical set. A local asymptotic analysis separates this possibility from worst-case global control.

Let $\Pi_n$ be the exact posterior law of $H_n = \sqrt{n}(\theta - \theta_0)$, and suppose $\Pi_n$ converges in total variation to $\Gamma = \mathcal{N}(m, \Sigma)$, where $\Sigma$ is positive definite. Define

$$\Psi_n(h) = \psi_n(\theta_0 + h/\sqrt{n}) - \psi_n(\theta_0),$$
$$\widetilde{\Pi}_n(dh) = \frac{e^{\Psi_n(h)}}{\int e^{\Psi_n}\, d\Pi_n}\, \Pi_n(dh).$$

The subtraction of $\psi_n(\theta_0)$ does not change the normalized tilt. For finite measures, $|\Pi_n - \Gamma|$ denotes the variation measure of their signed difference.

**Theorem 6** (Local linear-tilt limit)**.** *Suppose, for some $b \in \mathbb{R}^d$,*

$$\int \left| e^{\Psi_n(h)} - e^{b^\top h} \right| \Pi_n(dh) \longrightarrow 0,$$
$$\int e^{b^\top h}\, |\Pi_n - \Gamma|(dh) \longrightarrow 0.$$

*Then $\widetilde{\Pi}_n$ converges in total variation to $\mathcal{N}(m + \Sigma b, \Sigma)$, and*

$$\mathrm{TV}(\widetilde{\Pi}_n, \Pi_n) \longrightarrow 2\Phi\left(\frac{\sqrt{b^\top \Sigma b}}{2}\right) - 1,$$

*where $\Phi$ is the standard normal distribution function.*

*Proof.* By the triangle inequality, the assumptions imply convergence in variation of the unnormalized measures $e^{\Psi_n}\Pi_n$ to $e^{b^\top h}\Gamma$. Their total masses therefore converge to $\int e^{b^\top h}\,\Gamma(dh) > 0$, so normalization preserves total-variation convergence. Exponential tilting of $\mathcal{N}(m,\Sigma)$ by $e^{b^\top h}$ gives $\mathcal{N}(m+\Sigma b,\Sigma)$. The stated distance is the total-variation distance between two normal laws with common covariance and mean difference $\Sigma b$. □

**Corollary 7** (Smooth systematic error)**.** *Suppose* $\psi_n(\theta) = u_n G_n(\theta) + k_n$*, where* $u_n \to 0$*,* $u_n\sqrt{n} \to a$*, and*

$$\frac{\nabla G_n(\theta_0)}{n} \longrightarrow g.$$

*Assume that for every fixed* $R < \infty$ *there is* $C_R < \infty$ *such that*

$$\sup_{\|h\|\le R} \left\|\nabla^2 G_n(\theta_0 + h/\sqrt{n})\right\| \le C_R n,$$

*and that the two weighted convergence assumptions of Theorem 6 hold with* $b = ag$*. Then the approximate posterior, expressed in the* $H_n$ *coordinate, has limiting law*

$$\mathcal{N}(m + a\Sigma g, \Sigma).$$

*In particular, its limiting mean displacement in posterior-standard- deviation units is* $|a|\sqrt{g^\top \Sigma g}$*.*

*Proof.* Taylor's theorem gives, uniformly on every compact set of $h$,

$$\Psi_n(h) = \frac{u_n}{\sqrt{n}}\nabla G_n(\theta_0)^\top h + O(u_n\|h\|^2) \longrightarrow a\, g^\top h.$$

The asserted conclusion follows from Theorem 6 and its weighted integrability assumptions. □

**Exact Gaussian illustration.** Let the exact posterior be such that $H_n = \sqrt{n}(\theta - \theta_0) \sim \mathcal{N}(m,\Sigma)$ for every $n$, and define the deterministic error

$$\psi_n(\theta) = u_n n\, g^\top(\theta - \theta_0) + k_n$$

for fixed $g \in \mathbb{R}^d$ and arbitrary constants $k_n$. In the $H_n$ coordinate, $\Psi_n(h) = u_n\sqrt{n}\, g^\top h$. Exponential tilting is exact, so

$$\widetilde{\Pi}_n = \mathcal{N}(m + u_n\sqrt{n}\,\Sigma g, \Sigma)$$

and

$$\mathrm{TV}(\widetilde{\Pi}_n, \Pi_n) = 2\Phi\left(\frac{|u_n|\sqrt{n}\,\sqrt{g^\top\Sigma g}}{2}\right) - 1.$$

This analytic example shows that a smooth, systematic error may cause a posterior-scale displacement of order $u_n\sqrt{n}$, while a constant offset $k_n$ has no effect.

**Proposition 8** (Global spectral-gap comparison)**.** *Let* $P$ *be a reversible MH kernel for* $\pi$*, and let* $P^\psi$ *be the MH kernel using the same proposal and target* $\pi_\psi \propto e^\psi\pi$*, with* $\psi$ *bounded. Then*

$$\mathrm{Gap}(P^\psi) \ge e^{-\operatorname{osc}(\psi)}\,\mathrm{Gap}(P).$$

*Proof.* The off-diagonal reversible flow of an MH kernel is the minimum of the forward and reverse proposal flows. Tilting the target by $e^\psi$ therefore multiplies each such flow, after normalization, by a factor between $e^{\inf\psi}/Z_\psi$ and $e^{\sup\psi}/Z_\psi$. Thus the Dirichlet forms satisfy $\mathcal{E}_{P^\psi}(f,f) \ge e^{\inf\psi}Z_\psi^{-1}\mathcal{E}_P(f,f)$. The same density comparison gives $\mathrm{Var}_{\pi_\psi}(f) \le e^{\sup\psi}Z_\psi^{-1}\,\mathrm{Var}_\pi(f)$. Taking the infimum of the Dirichlet-form-to-variance ratio proves the claim. □

**Remark 9** (Two different precision scales)**.** *If a smooth systematic error has the form in Corollary 7, its local posterior displacement is governed by* $u_n\sqrt{n}$*. By contrast, if* $\operatorname{osc}(\psi_n) = O(u_n n)$*, Proposition 8 gives a global spectral-gap lower bound that can deteriorate when* $u_n n$ *is not small. These are sufficient worst-case scales under different assumptions, not universal thresholds for floating-point arithmetic.*

## 7 Discussion

An assumption-only precision budget must specify a state-uniform bound on absolute likelihood contributions, the accumulation algorithm, any pair-dependent remainder, and (when that remainder is present) a stability condition for the tilted chain. The resulting thresholds are properties of the model class and implementation, not of sample size and storage format alone.

This draft makes no universal claim that precision becomes inadequate at a fixed sample size. A complete statistical result would need to connect the deterministic bounds to a specified asymptotic posterior regime, for example by comparing expectation error with a posterior standard deviation under explicitly stated regularity assumptions. The local result above makes this connection under a smooth systematic error model. It does not derive that model from IEEE arithmetic; doing so, or replacing smoothness by a useful nonsmooth condition, is a necessary step before interpreting $u_n\sqrt{n}$ as a threshold for an actual implementation. No experiment is needed for such a theorem, but the result must account for cancellations and the chosen summation algorithm.

### Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

The authors declare no competing interests.

## References


Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up markov chain monte carlo: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 405–413. PMLR, 2014.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1683–1691. PMLR, 2014.

Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible markov chains. *The Annals of Applied Probability*, 3(3):696–730, 1993. doi: 10.1214/aoap/1177005363.

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.

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

Ryan Giordano, Tamara Broderick, and Michael I. Jordan. Covariances, robustness, and variational bayes. *Journal of Machine Learning Research*, 19 (51):1–49, 2018.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991. doi: 10.1145/103162.103163.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97.

Nicholas J. Higham. The accuracy of floating point summation. *SIAM Journal on Scientific Computing*, 14(4):783–799, 1993. doi: 10.1137/0914050.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2 edition, 2002. doi: 10.1137/1.9780898718027.

Richard Holley and Daniel Stroock. Logarithmic sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5–6):1159–1194, 1987.

IEEE. IEEE standard for floating-point arithmetic. IEEE Std 754-2019, 2019.

B. J. K. Kleijn and A. W. van der Vaart. The bernstein–von mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012. doi: 10.1214/12-EJS675.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the metropolis–hastings budget. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 181–189. PMLR, 2014.

Dougal Maclaurin and Ryan P. Adams. Firefly monte carlo: Exact MCMC with subsets of data, 2014.

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.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.

Jean-Michel Muller, Nicolas Brisebarre, Florent de Dinechin, Claude-Pierre Jeannerod, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, and Serge Torres. *Handbook of Floating-Point Arithmetic*. Birkhäuser, 2 edition, 2018.

Takeshi Ogita, Siegfried M. Rump, and Shin'ichi Oishi. Accurate sum and dot product. *SIAM Journal on Scientific Computing*, 26(6):1955–1988, 2005. doi: 10.1137/030601818.

Adrian E. Raftery and Steven M. Lewis. How many iterations in the gibbs sampler? In José M. Bernardo, James O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, 1992.

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.

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

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. doi: 10.1017/S096249291000005X.

Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994. doi: 10.1214/aos/1176325750.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

James H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice-Hall, 1963.