Nested $\widehat{R}$: Assessing the convergence of Markov chain Monte Carlo when running many short chains

Charles C. Margossian[*] Matthew D. Hoffman[†] Pavel Sountsov[†]
Lionel Riou-Durand[‡] Aki Vehtari[§] Andrew Gelman[¶]

Abstract. Recent developments in Markov chain Monte Carlo (MCMC) algorithms allow us to run thousands of chains in parallel almost as quickly as a single chain, using hardware accelerators such as GPUs. While each chain still needs to forget its initial point during a warmup phase, the subsequent sampling phase can be shorter than in classical settings, where we run only a few chains. To determine if the resulting short chains are reliable, we need to assess how close the Markov chains are to their stationary distribution after warmup. The potential scale reduction factor $\widehat{R}$ is a popular convergence diagnostic but unfortunately can require a long sampling phase to work well. We present a nested design to overcome this challenge and a generalization called *nested* $\widehat{R}$. This new diagnostic works under conditions similar to $\widehat{R}$ and completes the workflow for GPU-friendly samplers. In addition, the proposed nesting provides theoretical insights into the utility of $\widehat{R}$, in both classical and short-chains regimes.

Keywords: Markov chain Monte Carlo, parallel computation, convergence diagnostics, Bayesian inference, $\widehat{R}$ statistic.

1 Introduction

Over the past decade, much progress in computational power has come from special-purpose single-instruction multiple-data processors such as GPUs. This has motivated the development of GPU-friendly Markov chain Monte Carlo (MCMC) algorithms designed to efficiently run many chains in parallel (e.g., Lao et al., 2020; Hoffman et al., 2021; Sountsov and Hoffman, 2021; Hoffman and Sountsov, 2022; Riou-Durand et al., 2023). These methods often address shortcomings in pre-existing samplers designed with CPUs in mind: for example, ChEES-HMC (Hoffman et al., 2021) is a GPU-friendly alternative to the popular but control-flow-heavy no-U-turn sampler (NUTS) (Hoffman and Gelman, 2014). With these novel samplers, we can sometimes run thousands of chains almost as quickly as a single chain on modern hardware (Lao et al., 2020).

In practice, MCMC operates in two phases: a warmup phase that reduces the bias of the Monte Carlo estimators and a sampling phase during which the variance decreases

[*]Center for Computational Mathematics, Flatiron Institute
[†]Google Research
[‡]Laboratoire de Mathématiques de l'INSA Rouen Normandie
[§]Department of Computer Science, Aalto University
[¶]Department of Statistics and Political Science, Columbia University

with the number of samples collected. There are two ways to increase the number of samples: run a longer sampling phase or run more chains. Practitioners often prefer running a longer sampling phase because each chain needs to be warmed up and so the total number of warmup operations increases linearly with the number of chains. However, with GPU-friendly samplers, it is possible to efficiently run many chains in parallel. As a result, the higher computational cost for warmup only marginally increases the algorithm's runtime (Lao et al., 2020). It is then possible to trade the length of the sampling phase for the number of chains (Rosenthal, 2000). When running hundreds or thousands of chains, we can rely on a much shorter sampling phase than when running only 4 or 8 chains. This defines the *many-short-chains* regime of MCMC.

The length of the warmup phase is a crucial control parameter of MCMC. If the warmup is too short, the chains will not be close enough to their stationary distribution and the first iterations of the sampling phase will have an unacceptable bias. On the other hand, if the warmup is too long, we waste precious computation time. Both concerns are exacerbated in the many-short-chains regime. A large bias at the beginning of the sampling phase implies that the entire (short) chain carries a large bias, but running a longer warmup comes at a relatively high cost, since the warmup phase dominates the computation.

To check if the warmup phase is sufficiently long, practitioners often rely on convergence diagnostics (Cowles and Carlin, 1996; Robert and Casella, 2004; Gelman and Shirley, 2011; Gelman et al., 2013). Here, several notions of convergence for MCMC may be considered. When studying the warmup length, the emphasis is often on the total variation distance D_{TV} between the distribution of the first sample obtained after warmup and the target distribution p. Colloquially, has the Markov chain sufficiently approached its stationary distribution during warmup? We may also consider the bias of the Monte Carlo estimator, for any quantity of interest, and check that this bias is small, even negligible, before we start sampling. Convergence in D_{TV} relates to convergence in bias, though the two notions are not equivalent; see Roberts and Rosenthal (2004, Proposition 3). Unfortunately, neither D_{TV} nor the bias can be measured, and so these quantities must be monitored by indirect means.

In the multiple-chains setting, the most popular convergence diagnostic may well be the potential scale reduction factor $\widehat{R}$ (Gelman and Rubin, 1992; Brooks and Gelman, 1998; Vehtari et al., 2021). The driving idea behind $\widehat{R}$ is to compare multiple independent chains initialized from an overdispersed distribution and check that, despite the different initialization, each Markov chain still produces Monte Carlo estimators in close agreement. In other words, we check how well the Markov chain "forgets" its starting point with the understanding that, once the influence of the initialization vanishes, the chain must have reached its stationary distribution and the bias decayed to 0.

In this paper, we formally describe "forgetfulness" as the decay of a *nonstationary variance* (to be defined), show how to monitor it, and demonstrate how the nonstationary variance relates to bias decay. Furthermore, we show that, in the many-short-chains regime, $\widehat{R}$ may do a poor job monitoring the nonstationary variance and we propose a generalization of $\widehat{R}$, called the *nested* $\widehat{R}$, to adress this shortcoming.

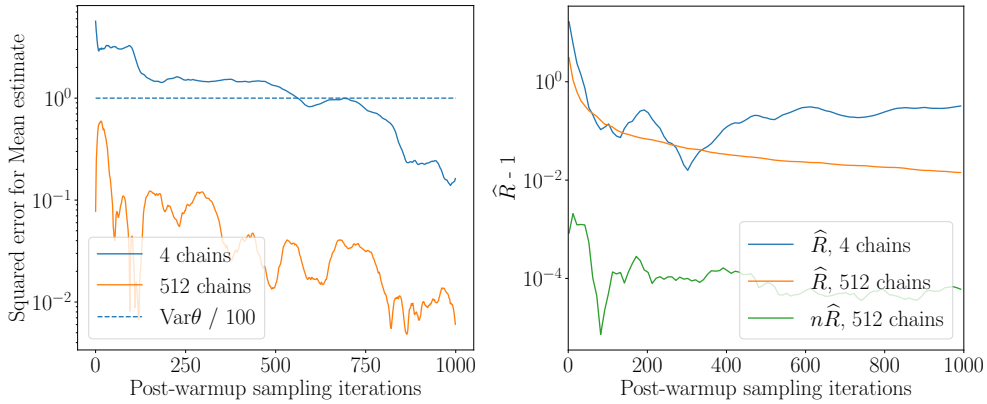

Figure 1: *MCMC for* $\mathbb{E}(\theta_1)$ *in the Rosenbrock distribution. (left) Squared error of Monte Carlo estimators after warmup. (right)* $\widehat{R}$ *computed with 4 or 512 chains, and* $\widehat{R}_{\mathrm{n}}$ *with 4 groups of 128 subchains.*

1.1 A motivating problem

We first examine the behavior of $\widehat{R}$ in the classic regime and the many-short-chains regime of MCMC. Consider the Rosenbrock distribution $p(\theta_1, \theta_2)$,

$$\theta_1 \sim \text{normal}(0, 1) \; ; \quad \theta_2 \mid \theta_1 \sim \text{normal}(0.03(\theta_1^2 - 100), 1), \tag{1}$$

and suppose we wish to estimate $\mathbb{E}_p(\theta_1)$, and achieve a squared error below $\text{Var}_p(\theta_1)/100$. This corresponds to the expected squared error attained with 100 independent samples from $p(\theta_1, \theta_2)$. We run ChEES-HMC (Hoffman et al., 2021) on a T4 GPU (available for free on Google Colaboratory[1]), first using 4 chains and then using 512 chains. We discard the first 100 iterations as part of the warmup phase and then run 1000 sampling iterations. On a GPU, running ChEES-HMC with 512 chains takes ∼20% longer than running 4 chains.[2] Using 4 chains, we require a sampling phase of ∼600-700 iterations to achieve our target precision (Figure 1, left). With 512 warmed-up chains, the target error is attained after 1 sampling iteration.

Next we compute the convergence diagnostic $\widehat{R}$, and we check whether or not $\widehat{R}$ is close to 1, for example below the threshold 1.01 proposed by Vehtari et al. (2021). But even after running 1000 sampling iterations, we find that $\widehat{R} > 1.01$ (Figure 1, right), whether we run 4 chains or 512 chains. Clearly, the computation cost to reduce $\widehat{R}$ to 1.01 far exceeds the cost to achieve our target precision, especially when running many chains. Figure 1 further suggests that, whether we run 4 chains or 512 chains, $\widehat{R}$ decays to 1 at the same rate, even if $\widehat{R}$ is noisier when computed with fewer chains.

[1] `https://colab.google/`

[2] Run time is evaluated using the Python command `%%timeit`, which reports an average ± standard deviation run time of: $2.42 \pm 0.03s$ for 4 chains and $3.07 \pm 0.04s$ for 512 chains, evaluated by running the code 7 times.

In this paper, we show that for $\widehat{R}$ to go to 1, the variance of the Monte Carlo estimator generated by a *single* chain must decrease to 0. This criterion cannot be met if each individual chain is short. Crucially, $\widehat{R}$ is a measure of mixing of the chains, which is a separate question than whether the final Monte Carlo estimator, obtained by averaging all the chains, has an acceptable error. For a simple example, consider a large number of chains started at random positions from the target distribution (or, equivalently, warmed up long enough to approximate independent draws from the target). Inference can be fine after a single post-warmup iteration, even though it could take a long time for the chains to mix.

1.2 Main Ideas: nonstationary variance and nested $\widehat{R}$

To address this issue, we introduce *nested* $\widehat{R}$, denoted $\widehat{R}_{\mathrm{n}}$. The key idea is to compare clusters of chains or *superchains* rather than individual chains. In order to still track the influence of the initialization, we require all the chains within a superchain to start at the same location. After a sufficiently long warmup, $\widehat{R}_{\mathrm{n}}$ decays to 1 with the number of subchains, even when each chain remains short. We show this on our motivating example, where we split the 512 chains into 4 superchains of 128 subchains; see the right panel of Figure 1.

Our approach can be motivated by an analysis of the Monte Carlo estimator's variance. Consider a state space Θ over which the target distribution p is defined, and suppose we want to estimate $\mathbb{E}(f(\theta))$, where $\theta \in \Theta$ and f maps θ to a univariate variable. In practice were are interested in multiple such functions f. Let $\bar{f}^{(1)}$ be the Monte Carlo estimator generated by a single Markov chain and let Γ be the distribution of $\bar{f}^{(1)}$; that is,

$$\bar{f}^{(1)} \sim \Gamma. \tag{2}$$

Γ is characterized by an initial draw from a starting distribution, $\theta_0 \sim p_0$, and then γ, the construction of the Markov chain starting at θ_0. The process γ includes the warmup phase (which is discarded) and the sampling phase. Then by the law of total variance

$$\mathrm{Var}_{\Gamma}\ \bar{f}^{(1)} = \underbrace{\mathrm{Var}_{p_0}\left[\mathbb{E}_{\gamma}(\bar{f}^{(1)} \mid \theta_0)\right]}_{\text{nonstationary variance}} + \underbrace{\mathbb{E}_{p_0}\left[\mathrm{Var}_{\gamma}(\bar{f}^{(1)} \mid \theta_0)\right]}_{\text{persistent variance}}. \tag{3}$$

We call the first term on the right side of eq. (3) the *nonstationary variance* and propose to use it as a formal measure of how well the Markov chain forgets its starting point. Indeed, if the warmup phase is sufficiently long, the initialization should only have a negligible influence on $\mathbb{E}(\bar{f}^{(1)})$ and so the nonstationary variance should be close to 0 (even when the sampling phase is short). Furthermore, the nonstationary variance acts as a "proxy clock" for the bias, since both quantities decay as the length of the warmup phase increases. We will illustrate, in an example, that both the squared bias and the nonstationary variance decay at the same rate. It is therefore useful to monitor the nonstationary variance in order to establish whether the warmup phase is sufficiently long for the bias to decay.

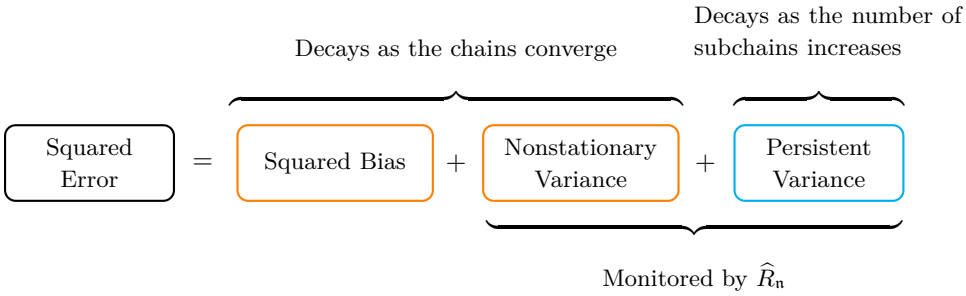

Figure 2: *Expected squared error for the Monte Carlo estimator generated by a superchain. The estimator is obtained by taking the sample mean of a superchain, after discarding the warmup phase.*

The second term on the right side of eq. (3) is the *persistent variance*: even if the chain reaches stationarity during warmup, the persistent variance remains large and can only decay after a long sampling phase. Fortunately, the persistent variance can be averaged out and its contribution to the final Monte Carlo estimator decreases to 0 as we increase the number of chains. Therefore, in the many-short-chains regime, the quantity of primary interest is the nonstationary variance due to its relationship to the squared bias, which cannot be reduced by increasing the number of chains.

To understand the behavior of $\widehat{R}_{\mathfrak{n}}$, we must now analyze the variance of the Monte Carlo estimator returned by a superchain, i.e. a cluster of chains initialized at the same point and then run independently. We show that in this case, the persistent variance decreases linearly with the number of subchains and so becomes small if we run a large number of chains. On the other hand, the nonstationary variance remains unchanged, reflecting the unchanged bias. $\widehat{R}_{\mathfrak{n}}$ can then effectively monitor the nonstationary variance of the Markov chain, even when each individual chain is short, provided we run many chains. Figure 2 summarizes the driving idea behind $\widehat{R}_{\mathfrak{n}}$.

1.3 Outline and contributions

The paper is organized as follows:

- We define $\widehat{R}_{\mathfrak{n}}$ as a generalization of $\widehat{R}$ and analyze the variance of the Monte Carlo estimator produced by a superchain. We showcase $\widehat{R}_{\mathfrak{n}}$'s ability to effectively monitor the nonstationary variance. In the edge case where each Markov chain contains one sample, we propose an exact correction to remove the influence of the persistent variance.
- We then consider a setting where the behavior of $\widehat{R}_{\mathfrak{n}}$ can be analyzed analytically: the continuous time limit of MCMC when targeting a Gaussian distribution, attained by a Langevin diffusion. In this context, we show that the nonstationary

variance decays at the same exponential rate as the squared bias.

- We empirically study the variance of $\widehat{R}_{\mathfrak{n}}$ at various phases of the MCMC warmup. Given a fixed total number of chains, we find that no choice of superchain size uniformly minimizes the variance of $\widehat{R}_{\mathfrak{n}}$ and we provide some recommendations. Finally, we showcase the use of $\widehat{R}_{\mathfrak{n}}$ across a range of Bayesian modeling problems, including ones where mixing is slow, the target is multimodal, or the parameter space is moderately high-dimensional ($d = 501$).

Throughout the paper, details of proofs are relegated to Appendix A. Code to reproduce the figures and experiments in this paper can be found at `https://github.com/charlesm93/nested-rhat`.

1.4 Related work

There has recently been a renewed interest in $\widehat{R}$ and its practical implementation (Vehtari et al., 2021; Vats and Knudson, 2021; Moins et al., 2023), with an emphasis on the classical regime of MCMC with 8 chains or fewer. $\widehat{R}$ computes a ratio of two standard deviations and is straightforward to evaluate. However, when applied to samples which are neither independent nor identically distributed—as is the case for MCMC—it becomes difficult to understand which quantity $\widehat{R}$ measures. It is moreover unclear how to choose a threshold for $\widehat{R}$ to decide if the Markov chains have converged. These issues were recently raised by Vats and Knudson (2021) and Moins et al. (2023), who studied the convergence of $\widehat{R}$ in the limit of infinitely long chains. But such an asymptotic analysis tacitly applies to stationary Markov chains and convergence of the Monte Carlo estimator itself during the sampling phase, rather than convergence in D_{TV} or in bias during the warmup. As prescribed by the many-short-chains regime, we took limits in another direction: an infinite number of finite nonstationary Markov chains. This perspective elicits the nonstationary variance, sheds light on the advantages and limitations of $\widehat{R}$, and motivates $\widehat{R}_{\mathfrak{n}}$.

Beyond running a long warmup phase, many strategies have been proposed to control the bias of Monte Carlo estimators. Examples include annealed importance sampling (Neal, 2001) and sequential Monte Carlo (Del Moral et al., 2006). More recently, unbiased MCMC has been proposed as a paradigm to construct unbiased estimators (Glynn and Rhee, 2014; Jacob et al., 2020). This strategy relies on transition kernels which allow pairs of Markov chains to couple after a finite time. Once a coupling occurs, we can construct unbiased estimators of expectation values. In this sense, the *coupling time* replaces the traditional warmup phase. Designing such transition kernels with short coupling times is an active area of research (Heng and Jacob, 2019; Jacob et al., 2020; Nguyen et al., 2022), but at present it is not always possible to find a kernel that couples quickly. For example, Hamiltonian Monte Carlo (HMC; Neal, 2012; Betancourt, 2018) methods using many integration steps per proposal are often the only viable option for sampling from poorly conditioned high-dimensional posteriors over a continuous space. Unfortunately the coupling HMC kernel of Heng and Jacob (2019) only couples

quickly if the problem is sufficiently well conditioned that HMC converges rapidly with a relatively small number of integration steps per proposal.

Methods such as Stein thinning (Riabiz et al., 2022) can remove strongly biased samples produced during the early stages of MCMC, and in some sense automatically discard a warmup phase (South et al., 2021). However, Stein methods can be computationally expensive and may not scale well in high dimensions. Furthermore, if the total length of the Markov chain is too short, no thinning can remove the bias incurred by the initialization—a problem which, arguably, $\widehat{R}_{\mathrm{n}}$ can detect and so the proposed diagnostic may be used conjointly with Stein methods.

2 Nested $\widehat{R}$

The motivation for $\widehat{R}_{\mathrm{n}}$ is to construct a diagnostic which does not conflate poor convergence and short sampling phase. $\widehat{R}_{\mathrm{n}}$ works by comparing Monte Carlo estimators whose persistent variance decreases with the number of chains. However, the bias does not decrease with the number of chains and this should be reflected in the nonstationary variance. Our solution is introduce superchains, which are groups of chains initialized at the same point and then run independently.

Consider an MCMC process over a state space Θ, which converges in D_{TV} to p for any choice of initialization; Roberts and Rosenthal (2004) review the conditions of irreducibility, aperiodicity, Harris recurrence under which the Markov chains converge. We denote the starting distribution p_0. Each Markov chain is initialized at a point drawn from p_0, and comprises $\mathcal{W}$ warmup iterations, which are discarded, and N sampling iterations. The warmup phase can simply be successive applications of a fixed transition kernel or it can involve an adaptation of the transition kernel (Andrieu and Thoms, 2008). In the latter case, we require that, as $\mathcal{W} \to \infty$, the Markov chain still converges in D_{TV} to p; this can be achieved by choosing an adaptation scheme that preserves the stationary distribution (e.g Gilks et al., 1994; Hoffman and Sountsov, 2022) or by freezing the transition kernel after a finite number of warmup iterations. The transition kernel stays fixed during the sampling phase.

So far, we have only considered a standard setup for MCMC. We now introduce the concept of superchain.

Definition 2.1. *(Superchain) We call superchain a collection of M Markov chains. Furthermore,*

1. *We say the superchain is* constrained *if all its subchains are initialized at the same point,*
2. *We say the superchain is* naive *if all its subchains are initialized independently.*

In either case, conditional on the initial point, the Markov chains are independent.

Throughout the paper, we primarily focus on constrained superchains and drop the

"constrained" prefix for brevity. Naive superchains are more straightforward to construct and serve as a benchmark.

We generate K superchains, each compromised of M subchains. In the case $M = 1$ where each superchain only contains a single subchain, we recover the classic regime of MCMC. We denote as $\theta^{(nmk)}$ the nth draw from chain m of superchain k. Suppose our goal is to estimate $\mathbb{E}(f(\theta))$ for some function f that maps θ to a univariate variable and assume furthermore that $\mathrm{Var}_p f(\theta) < \infty$. The sample mean of each superchain is

$$\bar{f}^{(..k)} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} f\left(\theta^{(nmk)}\right). \tag{4}$$

Moving forward, we write $f^{(nmk)} = f(\theta^{(nmk)})$ to alleviate the notation. The final Monte Carlo estimator is obtained by averaging all the superchains,

$$\bar{f}^{(\cdots)} = \frac{1}{K} \sum_{k=1}^{K} \bar{f}^{(..k)}. \tag{5}$$

We now define $\widehat{R}_{\mathfrak{n}}$.

Definition 2.2. *Consider the estimator $\widehat{B}_{\mathfrak{n}}$ of the between-superchain variance,*

$$\widehat{B}_{\mathfrak{n}} = \frac{1}{K-1} \sum_{k=1}^{K} \left(\bar{f}^{(..k)} - \bar{f}^{(\cdots)}\right)^2, \tag{6}$$

and the estimator $\widehat{W}_{\mathfrak{n}}$ of the within-superchain variance,

$$\widehat{W}_{\mathfrak{n}} \quad = \quad \frac{1}{K} \sum_{k=1}^{K} \left(\tilde{B}_k + \tilde{W}_k\right), \tag{7}$$

where we require either $N > 1$ or $M > 1$, and we introduce the estimators $\tilde{B}_k$ of the between-chain variance and $\tilde{W}_k$ of the within-chain variance; that is

$$\begin{aligned}
\tilde{B}_k &\triangleq \begin{cases} \frac{1}{M-1} \sum_{m=1}^{M} \left(\bar{f}^{(\cdot mk)} - \bar{f}^{(\cdot\cdot k)}\right)^2 & \text{if } M > 1, \\ 0 & \text{if } M = 1, \end{cases} \\
\tilde{W}_k &\triangleq \begin{cases} \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N-1} \sum_{n=1}^{N} \left(f^{(nmk)} - \bar{f}^{(\cdot mk)}\right)^2 & \text{if } N > 1, \\ 0 & \text{if } N = 1. \end{cases}
\end{aligned}$$

Then $\widehat{R}_{\mathfrak{n}}$ is the ratio between the total sample standard deviation across all super chains and the average within-superchain sample standard deviation, that is

$$\widehat{R}_{\mathfrak{n}} \triangleq \sqrt{\frac{\widehat{W}_{\mathfrak{n}} + \widehat{B}_{\mathfrak{n}}}{\widehat{W}_{\mathfrak{n}}}} = \sqrt{1 + \frac{\widehat{B}_{\mathfrak{n}}}{\widehat{W}_{\mathfrak{n}}}}. \tag{8}$$

Remark 2.3. *In the edge case where $M = 1$, $\widehat{R}_{\mathfrak{n}}$ reduces to $\widehat{R}$.*[3]

A common practice to assess convergence of the chains is to check that $\widehat{R} \leq 1 + \epsilon$ for some $\epsilon > 0$. Recommended choices of ϵ have evolved over time, starting at $\epsilon = 0.1$ (Gelman and Rubin, 1992) and recently using the more conservative value $\epsilon = 0.01$ (Vehtari et al., 2021). In the proposed nested design, we now check whether $\widehat{R}_{\mathfrak{n}} \leq 1 + \epsilon$ and therefore,

$$\widehat{B}_{\mathfrak{n}} \leq 2\epsilon \, \widehat{W}_{\mathfrak{n}} + \mathcal{O}(\epsilon^2). \tag{9}$$

This inequality establishes a tolerance value for $\widehat{B}_{\mathfrak{n}}$, scaled by the within-superchain-variance $\widehat{W}_{\mathfrak{n}}$. Moreover, we want to check that despite the distinct initialization and seed, the sample means of each superchain are in good agreement, as measured by $\widehat{B}_{\mathfrak{n}}$.

3 Properties of nested $\widehat{R}$

We now analyze the properties of $\widehat{R}_{\mathfrak{n}}$ theoretically and illustrate its use on several examples. For all formal statements, we assume the superchains are constrained, meaning all their subchains are initialized at the same point.

3.1 Which quantity does $\widehat{R}_{\mathfrak{n}}$ measure?

To answer this question, we consider the asymptotics of $\widehat{R}_{\mathfrak{n}}$ along K, the number of superchains. Applying the law of large numbers,

$$\widehat{B}_{\mathfrak{n}} \xrightarrow[K \to \infty]{\text{a.s}} B_{\mathfrak{n}} \triangleq \text{Var}_{\Gamma}(\bar{f}^{(\cdot\cdot k)}). \tag{10}$$

Then by the law of total variance,

$$B_{\mathfrak{n}} = \text{Var}_{p_0}\left[\mathbb{E}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k)\right] + \mathbb{E}_{p_0}\left[\text{Var}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k)\right]. \tag{11}$$

Because the chains are identically distributed and conditionally independent, we have

$$\mathbb{E}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k) = \mathbb{E}_{\gamma}(\bar{f}^{(\cdot mk)} \mid \theta_0^k) \tag{12}$$

$$\text{Var}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k) = \frac{1}{M}\text{Var}_{\gamma}(\bar{f}^{(\cdot mk)} \mid \theta_0^k). \tag{13}$$

Plugging these results back into (11), we obtain the following result.

Theorem 3.1. *Consider a constrained superchain initialized at $\theta_0^k \sim p_0$ and made of M chains. Then the variance of the superchain's sample mean is*

$$B_{\mathfrak{n}} = \underbrace{\text{Var}_{p_0}\left[\mathbb{E}_{\gamma}(\bar{f}^{(\cdot mk)} \mid \theta_0^k)\right]}_{\textit{nonstationary variance}} + \underbrace{\frac{1}{M}\mathbb{E}_{p_0}\left[\text{Var}_{\gamma}(\bar{f}^{(\cdot mk)} \mid \theta_0^k)\right]}_{\textit{persistent variance}}. \tag{14}$$

[3]The original $\widehat{R}$ uses a slightly different estimator for the within-chain variance $\tilde{W}_k$ when computing the numerator in $\widehat{R}$. There the sample variance $\tilde{W}_k$ is scaled by $1/N$, rather than $1/(N-1)$. This explains why occasionally $\widehat{R} < 1$. This is of little concern when N is large, but we care about the case where N is small, and we therefore adjust the $\widehat{R}$ statistic slightly.

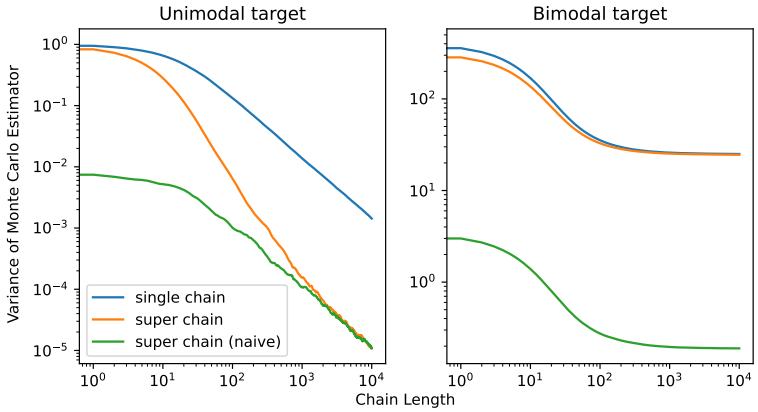

Figure 3: *Variance of Monte Carlo estimators,* $\bar{f}^{(\cdot\cdot k)}$*, constructed using a single chain, a constrained superchain of 1024 subchains initialized at the same point, or a naive superchain of 1024 independent subchains. When the Markov chains converge, the variance of a superchain decreases to the variance of a naive superchain. This transition occurs because the nonstationary variance decays to 0. If the Markov chains do not converge, the variance of a superchain stays large.*

Remark 3.2. *From Theorem 3.1, we immediately see that the constrained superchain has the same nonstationary variance as a single chain, but the persistent variance decreases linearly with* M*.*

We demonstrate the behavior of $B_{\mathfrak{n}}$ on two target distributions:

- A standard Gaussian, $p = \text{normal}(0, 1)$, with initial distribution, $p_0 = \text{normal}(2, 1)$.
- An unbalanced mixture of two Gaussians, $p = 0.3\,\text{normal}(-10, 1) + 0.7\,\text{normal}(10, 1)$, with initial distribution, $p_0 = \text{normal}(0, 20)$.

As benchmarks, we use B, the variance of the Monte Carlo estimator generated by a single chain, and $\tilde{B}_{\mathfrak{n}}$, the variance of the Monte Carlo estimator generated by a naive superchain. Because there are no initialization constraints, the nonstationary variance of the naive superchain also scales as $1/M$ (but the bias stays the same!).

We run Hamiltonian Monte Carlo (HMC; Neal, 2012) on a GPU and use $M = 1028$ chains for each superchain. Figure 3 shows the results. When the chains converge, $B_{\mathfrak{n}}$ first behaves like B and then transitions to behaving like $\tilde{B}_{\mathfrak{n}}$, as the nonstationary variance decays. When the chains do not converge, the transition does not occur and $B_{\mathfrak{n}}$ stays large, giving a clear indication that the chains do not converge. One drawback is that the constrained superchain has a larger variance than the naive superchain, especially during the early stages of MCMC; however, if the Markov chains converge, the nonstationary variance vanishes, and the expected squared error is dominated by the persistent variance, which is the same for both types of superchains.

We end this section with two corollaries of Theorem 3.1. To show these corollaries, we first require the following lemma on the asymptotic limit of $\widehat{W}_{\mathfrak{n}}$.

Lemma 3.3. *In the limit of an infinite number of superchains,*

$$\lim_{K\to\infty} \widehat{W}_{\mathfrak{n}} = W_{\mathfrak{n}} \triangleq \mathbb{E}_{p_0} Var_{\gamma}(\bar{f}^{(\cdot mk)} \mid \theta_0^k) + W', \tag{15}$$

where

$$W' \triangleq \begin{cases} \frac{1}{N-1} \sum_{n=1}^{N} Var_{\Gamma} f^{(nmk)} - Var_{\Gamma} \bar{f}^{(\cdot mk)} + (\mathbb{E}_{\Gamma} f^{(nmk)})^2 - (\mathbb{E}_{\Gamma} f^{(\cdot mk)})^2 & \text{if } N > 1 \\ 0 & \text{if } N = 1, \end{cases}$$

The first corollary of Theorem 3.1 examines the behavior of $\widehat{R}_{\mathfrak{n}}$ after a long warmup phase, specifically in the limit where the Markov chains have converged to their stationary distribution.

Corollary 3.4. *Suppose that all chains within a superchain start at the same point $\theta_0^k \sim p_0$ and that each chain is made of M subchains. Assume further that the length of the sampling phase $N > 1$ and the chains have positive autocorrelation. Then, for an infinitely long warmup phase,*

$$\lim_{\mathcal{W}\to\infty} \sqrt{1 + \frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}}} \geq \sqrt{1 + \frac{1}{M} \frac{1 - 1/N}{\text{ESS}_{(1)} + 1/N}}, \tag{16}$$

where $\text{ESS}_{(1)}$ *is the effective sample size for a single chain of length N.*

From the above, we see that without the nesting design (i.e. $M = 1$ case), $\widehat{R}$ can only decay to 1 if $\text{ESS}_{(1)}$ is large, because we need to kill off the persistent variance. This explains the results seen in Figure 1 and echoes the observation by Vats and Knudson (2021) that for stationary Markov chains, $\widehat{R}$ (or rather the quantity measured by $\widehat{R}$) is a one-to-one map with the ESS. We emphasize however that this equivalence does not hold when the chains are not stationary and $\mathcal{W} < \infty$.

One persistent problem is that, even though we want to monitor the nonstationary variance, we instead measure the total variance. The second corollary of Theorem 3.1 provides a means to adress this problem. This corollary examines the special case where $N = 1$, which is mathematically convenient and takes the logic of the many-short-chains regime to its extreme, with the variance reduction entirely handled by the number of chains.

Corollary 3.5. *Suppose that all chains within a superchain start at the same point $\theta_0^k \sim p_0$ and that each chain is made of M subchains. Assume further that $N = 1$. Then the persistent variance, scaled by $W_{\mathfrak{n}}$, is*

$$\frac{\mathbb{E}_{p_0} Var_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k)}{W_{\mathfrak{n}}} = \frac{1}{M}, \tag{17}$$

and furthermore

$$\sqrt{1 + \frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}}} = \sqrt{1 + \frac{1}{M} + \frac{\text{Var}_{p_0} \mathbb{E}_{\gamma}(f^{(1mk)} \mid \theta_0^k)}{\mathbb{E}_{p_0} \text{Var}_{\gamma}(f^{(1mk)} \mid \theta_0^k)}}. \tag{18}$$

We can now express a threshold on $\widehat{R}_{\mathfrak{n}}$ as a threshold on the nonstationary variance, scaled by the within-chain variance. Conversely, we can express a bound on the (scaled) nonstationary variance as an equivalent bound on $\widehat{R}_{\mathfrak{n}}$. This result does not depend on the transition kernel, nor on the length of the warmup phase, and it applies to nonstationary Markov chains.

Remark 3.6. *Such a result is not available for $\widehat{R}$, which by definition must be computed for $N > 1$.*

3.2 Bias and nonstationary variance

We have established that $\widehat{R}_{\mathfrak{n}}$ monitors the nonstationary variance. Now the primary goal of the warmup phase is to reduce the bias. In this section, we examine the connection between the nonstationary variance and the bias in an illustrative example. We first define the idea of reliability for a convergence diagnostic.

Definition 3.7. *For a univariate random variable f, we say an MCMC process is (δ, δ')-reliable for $\widehat{R}_{\mathfrak{n}}$ if*

$$\frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}} \leq \delta \implies \frac{(\mathbb{E}_\Gamma \bar{f}^{(\cdot m)} - \mathbb{E}_p f)^2}{\mathrm{Var}_p f} \leq \delta'. \tag{19}$$

Since $\widehat{R}$ is a special case of $\widehat{R}_{\mathfrak{n}}$, we have also defined reliability for $\widehat{R}$. Gelman and Rubin (1992) tackled the question of $\widehat{R}$'s reliability by using an overdispersed initialization. Here, we provide a formal proof that in the Gaussian case (δ, δ')-reliability is equivalent to using an initial distribution p_0 with a large variance relative to the initial bias. The Gaussian case provides intuition for unimodal targets and can be a reasonable approximation after rank normalization of the samples (Vehtari et al., 2021).

Let $p = \mathrm{normal}(\mu, \sigma^2)$. To approximate a large class of MCMC processes, we consider the solution $(X_t)_{t \geq 0}$ of the Langevin diffusion targeting p defined by the stochastic differential equation

$$dX_t = -(X_t - \mu)dt + \sqrt{2\sigma}\, dW_t, \tag{20}$$

where $(W_t)_{t \geq 0}$ is a standard Brownian motion. The convergence of MCMC toward continuous-time stochastic processes has been widely studied, notably by Gelman et al. (1997) and Roberts and Rosenthal (1998), who established scaling limits of random walk Metropolis and the Metropolis adjusted Langevin algorithm toward Langevin diffusions. Similar studies have been conducted for Hamiltonian Monte Carlo and its extensions; see, e.g., Beskos et al. (2013), Riou-Durand and Vogrinc (2022). Typically, the solution of a continuous-time process after a time $T > 0$ is approximated by a Markov chain, discretized with a time step $h > 0$ and run for $\lfloor T/h \rfloor$ steps. We consider here a Gaussian initial distribution

$$p_0 = \mathrm{normal}(\mu_0, \sigma_0^2). \tag{21}$$

In this setup, the bias and the variance of the Monte Carlo estimator admit an analytical form. The solution X_T is interpreted as an approximation as $h \to 0$ of the setting of

parallel chains for $\mathcal{W} = \lfloor T/h \rfloor$ iterations and $N = 1$, i.e., $\theta^{(1mk)} = X_T$. This scenario is the simplest one to analyze and illustrates an edge case of the many-short-chains regime.

The main result of this section states that the squared bias and scaled nonstationary variance decay at a rate $\sim e^{-2T}$, providing justification as to why the latter can be used as a proxy clock for the former.

Theorem 3.8. *Let $(X_t)_{t\geq 0}$ be the solution to (20), which describes a Langevin diffusion targeting $p = \text{normal}(\mu, \sigma)$, starting from $X_0 \sim p_0 = \text{normal}(\mu_0, \sigma_0)$. Then for any warmup time $T > 0$, the bias is*

$$\mathbb{E}\bar{\theta}^{(1\cdot k)} - \mathbb{E}_p X = \mathbb{E}X_T - \mathbb{E}_p X = (\mu_0 - \mu)e^{-T}. \tag{22}$$

Furthermore,

$$\frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}} = \frac{\text{Var}_{p_0}\mathbb{E}(X_T \mid X_0)}{\mathbb{E}_{p_0}\text{Var}(X_T \mid X_0)} = \frac{1}{M} + \frac{\sigma_0^2}{\sigma^2(e^{2T} - 1)}. \tag{23}$$

We end this section with a corollary that states that the initialization needs to be sufficiently dispersed for $\widehat{R}_{\mathfrak{n}}$ to be reliable.

Corollary 3.9. *Let $\delta > 0$ and $\delta' > 0$. Assume the conditions stated in Theorem 3.8. If $(\mu_0 - \mu)^2/\sigma^2 \leq \delta'$, $\widehat{R}_{\mathfrak{n}}$ is trivially (δ, δ')-reliable. If $(\mu_0 - \mu)^2/\sigma^2 > \delta'$, then $\widehat{R}_{\mathfrak{n}}$ is (δ, δ')-reliable if and only if*

$$\sigma_0^2 > \left(\delta - \frac{1}{M}\right)\left(\frac{(\mu - \mu_0)^2}{\delta'\sigma^2} - 1\right)\sigma^2. \tag{24}$$

Remark 3.10. *If $\delta < 1/M$, then the condition $B_{\mathfrak{n}}/W_{\mathfrak{n}} < \delta$ cannot be verified. Hence $\widehat{R}_{\mathfrak{n}}$ is reliable in a trivial sense: we never erroneously claim convergence because we never claim convergence.*

Appendix B provides additional results on the reliability of $\widehat{R}_{\mathfrak{n}}$: we numerically test the lower bound provided by Corollary 3.9 on a Gaussian and a mixture of Gaussians, and we then derive a similar bound for $\widehat{R}$ when $N > 1$.

3.3 Variance of $\widehat{R}_{\mathfrak{n}}$

In the previous sections, we focused on the quantity measured by $\widehat{R}_{\mathfrak{n}}$, attained in the limit of a large number of superchains, that is $K \to \infty$. We now assume a finite number of superchains. Even if $\widehat{R}_{\mathfrak{n}}$ is (δ, δ')-reliable, it may be too noisy to be useful. An example of this arises in the multimodal case, where all K superchains may initialize *by chance* in the attraction basin of the same mode. In this scenario, K is too small and we drastically underestimate $B_{\mathfrak{n}} = \text{Var}_\Gamma \bar{f}^{(\cdot\cdot k)}$. One would need enough superchains with distinct initializations to find multiple modes and diagnose the chains' poor mixing.

When using $\widehat{R}$ it seems reasonable to increase the number of chains as much as possible. The question is more subtle for $\widehat{R}_{\mathfrak{n}}$: given a fixed total number of chains KM

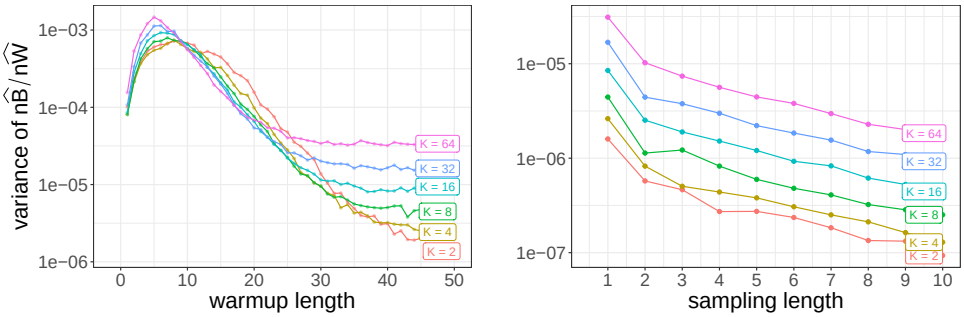

Figure 4: *Variance of* $\widehat{B}_{\mathfrak{n}}/\widehat{W}_{\mathfrak{n}}$ *when running* $KM = 2048$ *chains split into* K *superchains. (left)* $\widehat{B}_{\mathfrak{n}}/\widehat{W}_{\mathfrak{n}}$ *is computed using* $N = 1$ *samples after a varying warmup length. No choice of* K *uniformly minimizes the variance across all phases of MCMC. (right) This time, the chains are stationary. Increasing the length of the sampling phase* N *reduces the variance of* $\widehat{B}_{\mathfrak{n}}/\widehat{W}_{\mathfrak{n}}$.

how many superchains should we run? In general no choice of K, given KM, minimizes the variance of $\widehat{B}_{\mathfrak{n}}/\widehat{W}_{\mathfrak{n}}$ across all stages of MCMC.

We demonstrate this phenomenon when targeting a Gaussian with $KM = 2048$ chains (Figure 4, left). Here the choice $K = 2$ minimizes the variance once the chains are stationary but nearly maximizes it during the early stages of the warmup phase. For K small, all the chains may be in agreement by chance even when they are not close to the stationary distribution. Increasing K helps avoid this scenario. On the other hand, a large K results in a large variance once the chains approach their stationary distribution. This is because the persistent variance remains large if K is large (and therefore M is small), even once the nonstationary variance vanishes. In other words $\widehat{B}_{\mathfrak{n}}$ remains noisy because of a large nuisance term. We now see the competing forces at play when choosing K. Empirically, we will find that several choices of K work well across a collection of problems (Section 4).

The variance of $\widehat{R}_{\mathfrak{n}}$ can be further reduced by increasing the length of the sampling phase N. The right panel of Figure 4 demonstrates the reduction in variance as N varies from 1 to 10. For many problems, running 10 more iterations is computationally cheap. A large N also makes $B_{\mathfrak{n}}$ a sharper upper bound on the nonstationary variance. However, if $N > 1$, we cannot use Corollary 3.5 to exactly characterize and correct for the persistent variance.

3.4 Error tolerance and threshold for $\widehat{R}_{\mathrm{n}}$

Ultimately, our goal is to control the error of our Monte Carlo estimators. Consider the bias-variance decomposition for an estimator returned by a superchain,

$$\mathbb{E}_{\Gamma}\left((\bar{f}^{(\cdot\cdot k)} - \mathbb{E}_p f)^2\right) = \underbrace{(\mathbb{E}_{\Gamma}\bar{f}^{(\cdot\cdot k)} - \mathbb{E}_p f)^2}_{\text{squared bias}} + \underbrace{\mathrm{Var}_{p_0}(\mathbb{E}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k))}_{\text{nonstationary variance}} + \underbrace{\mathbb{E}_{p_0}(\mathrm{Var}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k))}_{\text{persistent variance}}. \tag{25}$$

Much of the MCMC literature focuses on the final term, with the assumption that we have run the Markov chain's warmup "long enough" for them to be approximately stationary, at which point the bias (and the nonstationary variance) become negligible. The error tolerance can then be expressed in terms of the MCSE and the ESS (e.g. Flegal et al., 2008; Gelman et al., 2013; Vats et al., 2019; Vehtari, 2022). What constitutes an appropriate ESS is problem-dependent and also subject to academic discussion (e.g Mackay, 2003; Gelman and Shirley, 2011; Vats et al., 2019; Vehtari, 2022; Margossian and Gelman, 2023).

If the Markov chains are not close to stationarity, variance alone cannot characterize the expected squared error, and so we must first check for approximate convergence. We may do so by setting a tolerance τ on the (scaled) nonstationary variance,

$$\frac{\mathrm{Var}_{p_0}\left(\mathbb{E}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k)\right)}{\mathbb{E}_{p_0}\left(\mathrm{Var}_{\gamma}(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k)\right)} \leq \tau. \tag{26}$$

Since with $\widehat{R}_{\mathrm{n}}$ we measure the total variance rather than the nonstationary variance, we need to correct for the persistent variance by either running more subchains, or falling back on the $N = 1$ case to apply Corollary 3.5, which tells us that the scaled persistent variance is exactly $1/M$. We use the latter and construct our threshold as

$$\widehat{R}_{\mathrm{n}} \leq \sqrt{1 + \frac{1}{M} + \tau}. \tag{27}$$

The choice of τ itself, just like the tolerable expected squared error, depends on the problem. We propose to make the nonstationary variance—and so, by the proxy clock heuristic, the squared bias—small next to the tolerable squared error. Then the error is dominated by the persistent variance and can be characterized by estimators of the MCSE. We will demonstrate such an approach on a collection of examples.

4 Numerical experiments

We demonstrate an MCMC workflow using $\widehat{R}_{\mathrm{n}}$ on a diversity of applications mostly drawn from the Bayesian literature. Our focus is on producing accurate estimates for the first moment for the model parameters. We consider six targets which represent a diversity of applications, notably in Bayesian modeling. Table 1 summarises the target

Target	d	**Description**
Rosenbrock	2	A joint normal distribution nonlinearly transformed to have high curvature (Rosenbrock, 1960). See Equation 1. This target produces Markov chains with a large autocorrelation.
Eight Schools	10	The posterior for a hierarchical model of the effect of a test-preparation program for high school students (Rubin, 1981). Fitting such a model with MCMC requires a careful reparameterization (Papaspiliopoulos et al., 2007).
German Credit	25	The posterior of a logistic regression applied to a numerical version of the German credit data set (Dua and Graff, 2017).
Pharmacokinetics	45	The posterior for a hierarchical model describing the absorption of a drug compound in patients during a clinical trial (e.g Wakefield, 1996; Margossian et al., 2022), using data simulated over 20 patients. This model uses a likelihood based on an ordinary differential equation.
Bimodal	100	An unbalanced mixture of two well-separated Gaussians. With standard MCMC, each Markov chain "commits" to a single mode, leading to bias sampling. Even after a long compute time, the Markov chains fail to converge.
Item Response	501	The posterior for a model to assess students' ability based on test scores (Gelman and Hill, 2007). The model is fitted to the response of 400 students to 100 questions, and the model estimates (i) the difficulty of each question and (ii) each student's. aptitude. This problem has a relatively high dimension.

Table 1: Target distributions for our numerical experiments.

distributions with more details available in Appendix C. The bimodal example provides a case where the chains fail to converge after a reasonable amount of compute time. As our MCMC algorithm, we run ChEES-HMC (Hoffman et al., 2021), which is an adaptive HMC sampler, designed to run efficiently on GPUs. ChEES-HMC pools information across all Markov chains to set the tuning parameters for HMC, specifically (i) the step size of the integrator solving Hamilton's equations of motion and (ii) the number of steps the integrator takes, thereby determining the length of the Hamiltonian trajectory. Fixing (ii) across all chains ensures ensures that, at a given iteration, each chain requires the same amount of compute and so the operation can be efficiently parallelized on a GPU. The algorithm is implemented in TensorFlow Probability (TensorFlow Probability Development Team, 2023).

4.1 Monte Carlo squared error after convergence

We construct a model of the squared error for stationary Markov chains, and use this as a benchmark for the empirical squared error. In the stationarity limit, i.e. $\mathcal{W} \to \infty$, the subchains within a superchain are no longer correlated. A central limit theorem may then be taken along the total number of chains. Then for $N = 1$, the scaled squared error approximatively follows a χ^2 distribution,

$$E^2 \triangleq \frac{KM}{\text{Var}_p f}(\bar{f}^{(1\cdot\cdot)} - \mathbb{E}_p f)^2 \overset{\text{approx.}}{\sim} \chi^2_1. \tag{28}$$

High-precision estimates of $\mathbb{E}_p f$ and $\text{Var}_p f$ are computed using long MCMC runs (Appendix C). We use the above approximation to jointly model the expected squared error at stationarity across all dimensions. This is somewhat of a simplification, since we do not account for the correlations between dimensions.

After a sufficiently long warmup phase, the Markov chains are nearly stationary and E^2 approximately follows a χ^2_1 distribution. This should ideally be reflected by $\widehat{R}_{\mathfrak{n}} \approx 1$. However if the warmup phase is too short and the squared error large because of the Markov chain's bias, we expect to see $\widehat{R}_{\mathfrak{n}} \gg 1$, per the proxy clock heuristic.

4.2 Results when running 2048 chains with $K = 16$ superchains

Suppose we target an ESS of $\sim$2000. In this case, we may run 2048 chains, broken into $K = 16$ superchains and $M = 128$ subchains. For each target distribution, we compute $\widehat{R}_{\mathfrak{n}}$ using $N = 1$ draw after warmups of varying lengths,

$$\mathbf{W} = (10, 20, 30, \ldots, 100, 200, 300, \ldots, 1000).$$

$\mathbf{W}$ contains both warmup lengths that are clearly too short to achieve convergence and lengths after which approximate convergence is expected when running HMC. That way, the behavior of $\widehat{R}_{\mathfrak{n}}$ can be examined on both nonstationary and (nearly) stationary Markov chains. At the end of each warmup window, we record $\widehat{R}_{\mathfrak{n}}$ for each dimension and the corresponding E^2 using a single draw per chain. We repeat our experiment 10 times for each model.

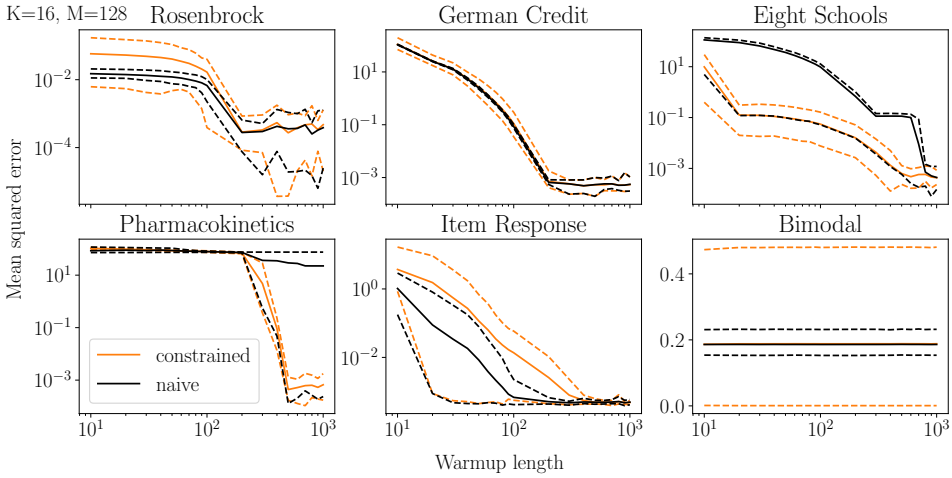

Figure 5: *Mean squared error, scaled by the posterior variance, when using a constrained superchain with* $K = 16$ *distinct initializations and a naive superchain. The results are computed across 10 seeds. The solide is the average MSE, and the dotted lines represent the best and worst runs across 10 seeds.*

While helpful for diagnosing convergence, constraining all subchains within a superchain to start at the same location increases the total Monte Carlo variance (Theorem 3.1). We empirically investigate the repercusions of this initialization scheme on the mean squared error (MSE). Overall, we do not see a drastic difference between using a constrained or a naive superchain, especially as the warmup length approaches 1000 iterations (Figure 5). For the Item Response model, the MSE decays faster with a naive superchain on average. For the hierarchical models (Eight Schools and Pharmacokinetics) the naive superchains has a slower decaying MSE. This phenomenon is consistent with past observations that for models with an intricate posterior geometry, one or more chains can get stuck in difficult regions of the probability space, usually well in the tails of the target distribution (Hoffman et al., 2021; du Ché and Margossian, 2023). The adaptation strategy of ChEES-HMC is to reduce the step size of HMC to insure that Markov chains stuck in difficult regions can still move forward ("no chains left behind" heuristic). Now, ChEES-HMC uses a common transition kernel for all chains, and a conservative step size, while helpful for stuck chains, may be suboptimal for other chains that may already have reached regions where the posterior probability mass concentrates. Unfortunately, increasing the number of distinct initializations increases the probability that at least one chain gets stuck in a difficult region, particularly during the early warmup.

In practice, the MSE itself cannot be measured and we rely on convergence diagnostics. We plot in Figure 6 E^2 against $\widehat{R}_{\mathfrak{n}}$ for a constrained superchain, and we observe a clear correlation between $\widehat{R}_{\mathfrak{n}}$ and E^2. After a short warmup, the chains are far from their stationary distribution: this manifests as both a large E^2 and a large $\widehat{R}_{\mathfrak{n}}$. When

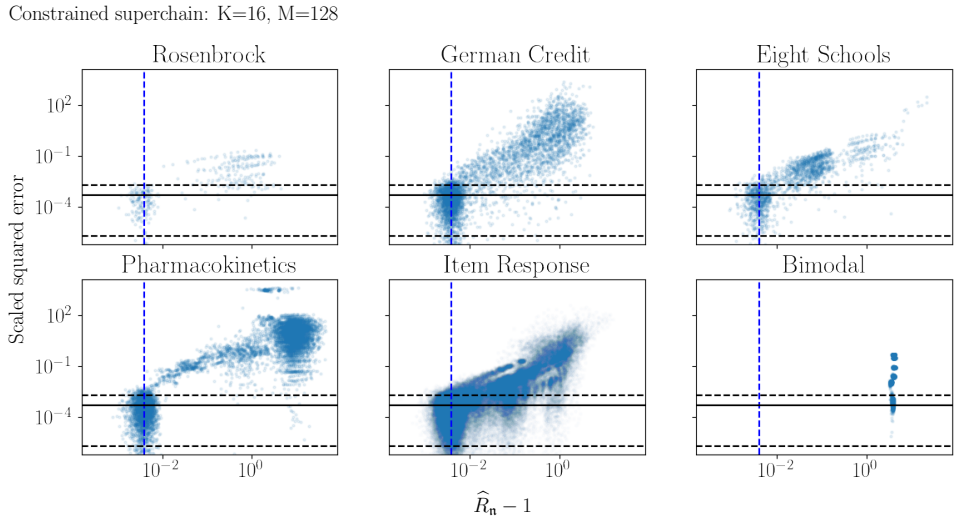

Figure 6: *Scaled squared error against* $\widehat{R}_{\mathfrak{n}}$*, using* $K = 16$*,* $M = 128$ *and* $N = 1$*, with constrained superchains. For short warmup phases, the Markov chains are far from their stationary distribution, which manifests as a large squared error and a large* $\widehat{R}_{\mathfrak{n}}$*. Once* $\widehat{R}_{\mathfrak{n}}$ *is close to 1, the scaled squared error behaves as we would expect from stationary Markov chains (Section 4.1). The vertical blue line corresponds to the proposed threshold* $\epsilon \approx 0.004$*. The horizontal lines show the median (solid line) and 0.9 coverage (dashed lines) for the squared error of a stationary Markov chain.*

$\widehat{R}_{\mathfrak{n}}$ is close to 1, the squared error is smaller and approaches the distribution we would expect from stationary Markov chains (Section 4.1). For the bimodal target, the Markov chains fail to converge after a warmup of 1000 iterations and our experiment only reports observations where E^2 and $\widehat{R}_{\mathfrak{n}}$ are both large.

As a tolerance on the scaled nonstationary variance, we consider $\tau = 10^{-4}$, which corresponds to a fifth of the scaled variance we tolerate with an ESS of 2000.[4] The tolerance for $\widehat{R}_{\mathfrak{n}}$ is then $\sim$1.004, which is smaller than the recommended threshold of 1.01 for $\widehat{R}$. Bear in mind the tolerance we chose is relative to our target ESS, which can change between problems. Figure 7 plots the fraction of times the scaled squared error E^2 is above the 0.95^{th} quantile of a χ^2_1 distribution for $\widehat{R}_{\mathfrak{n}}$ below varying thresholds. For $\widehat{R}_{\mathfrak{n}} \leq \sqrt{1 + 1/M + \tau}$, this fraction approaches 0.05 for all models but can be larger.

Finally, we examine what would happen if we computed $\widehat{R}_{\mathfrak{n}}$ with naive superchains, that is without constraining all the subchains to start at the same point (Figure 8). In this setting, $\widehat{R}_{\mathfrak{n}}$ hardly corrolates with E^2, and there is no threshold of $\widehat{R}_{\mathfrak{n}}$ under which

[4] An ESS of 2000 corresponds to a relative variance, $\text{Var}_\Gamma \bar{f}^{(\cdots)}/\text{Var}_p f = 1/2000 = 5 \times 10^{-4}$, and the tolerance on the nonstationary variance, scaled by the estimator $\widehat{W}_{\mathfrak{n}}$ of $\text{Var}_p f$, is set to a fraction of this quantity.

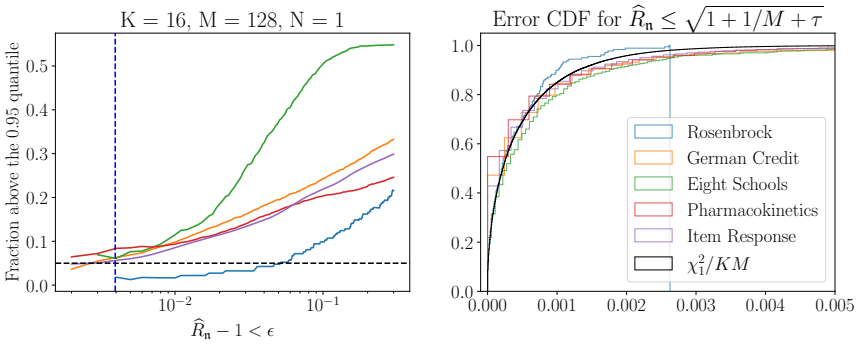

Figure 7: *(left) Fraction of Monte Carlo estimates with squared error above the* 95^{th} *quantile of the stationary error distribution (Section 4.1). The vertical blue line is the prescribed threshold* $\sqrt{1 + 1/M + \tau}$*. (right) For* $\widehat{R}_{\text{n}}$ *close to 1, the empirical CDF approaches the theoretical CDF for stationary Markov chains.*

the scaled squared error follows the distribution we expected from stationary Markov chains.

4.3 Results when varying the number of superchains K

We extend the analysis in the previous section to the case where the total number of chains stays fixed at $KM = 2048$, but we vary the number of superchains K. As before, we first examine the scaled MSE, and find no clear disadvantage to using constrained superchains (Figure 9). For the Rosenbrock, German Credit, and Item Response models, increasing the number of distinct initializations improves, to varying degrees, the rate at which the MSE decreases. For the hierarchical models (Eight Schools and Pharmacokinetics), there is no clear trend.

We next examine whether $\widehat{R}_{\text{n}}$ can diagnose convergence of the Markov chains (Figure 10). The threshold for $\widehat{R}_{\text{n}}$ is adjusted as M varies, although the tolerance on the nonstationary variance τ remains fixed. For $K \in \{8, 16, 64, 256\}$ we find that the 95^{th} quantile of E^2 is in reasonable agreement with the 95^{th} quantile of a χ^2_1, albeit slightly larger. The results are less stable for the extreme choices $K = 2$ and $K = 1024$. For $K = 2$ we expect the variance of $\widehat{R}_{\text{n}}$ to be large during the early stages of MCMC, while for $K = 1024$, the variance is high near stationarity (Section 3.3). Overall, there is a broad range of choices for K away from these extremes that yield a functioning convergence diagnostic in the considered examples.

5 Discussion

While CPU clock speeds stagnate, parallel computational resources continue to get cheaper. The question of how to make effective use of these parallel resources for MCMC

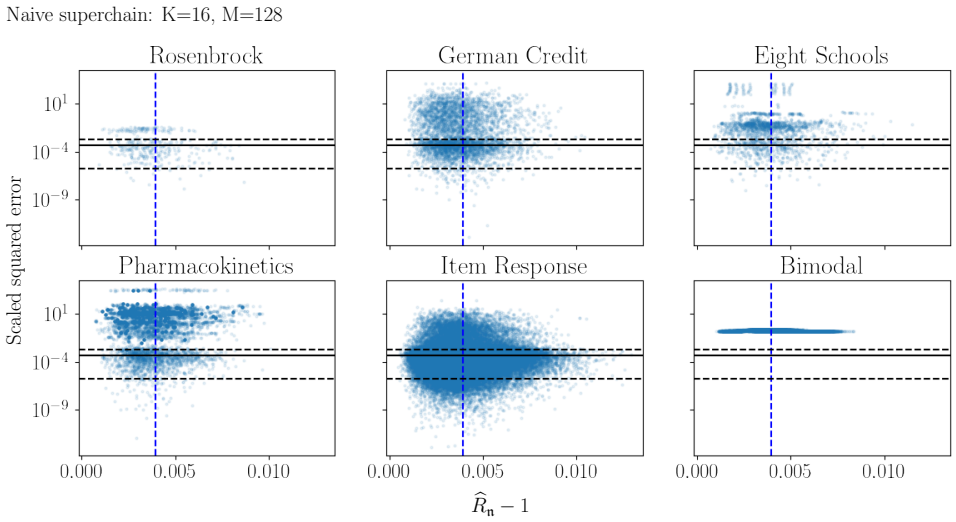

Figure 8: *Scaled squared error against* $\widehat{R}_{\mathbf{n}}$*, using* $K = 16$*,* $M = 128$*, and* $N = 1$*, with a naive superchain. Without the constrained initialization, there is no useful correlation between* $\widehat{R}_{\mathbf{n}}$ *and the squared error.*

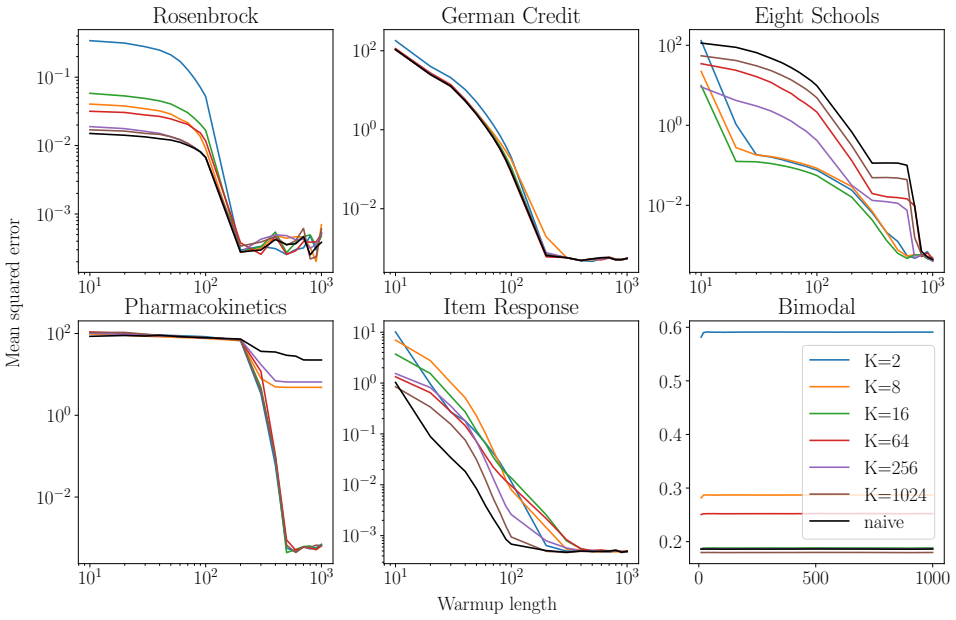

Figure 9: *Mean squared error, scaled by the posterior variance, when using* K *constrained superchains for a total of 2048 chains or using 2048 independently initialized chains (naive setting).*

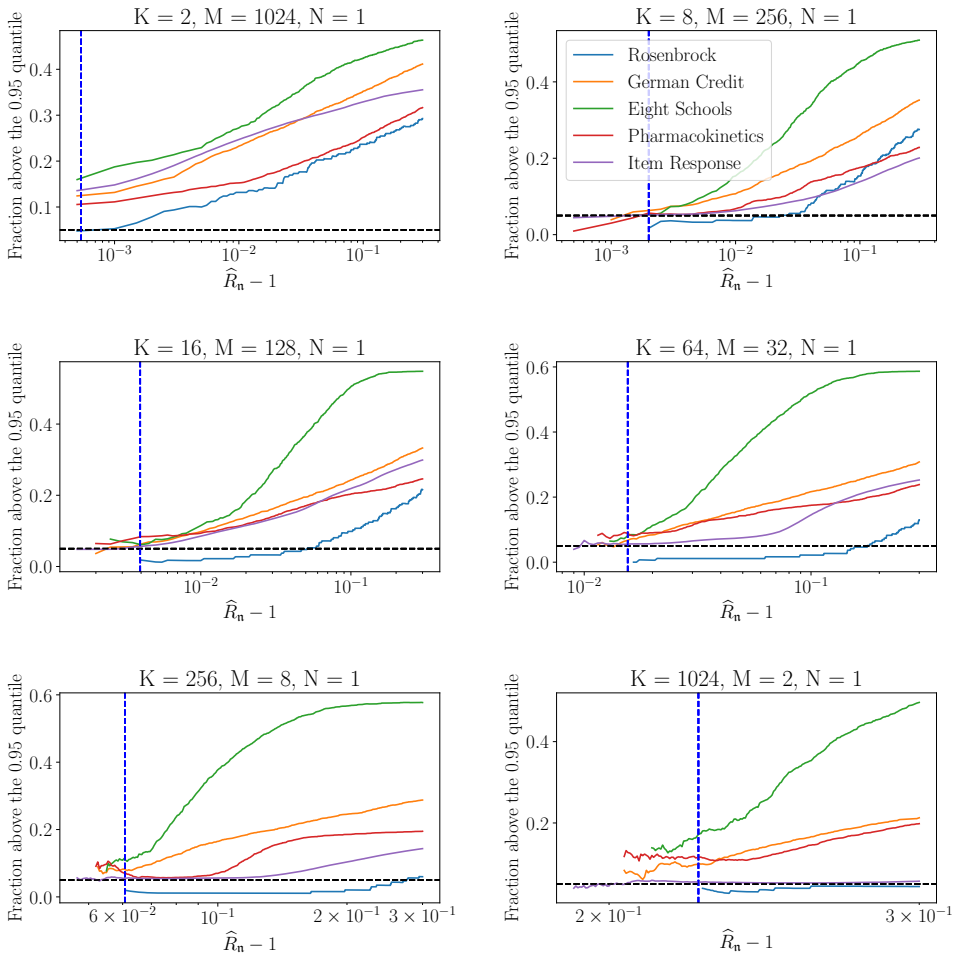

Figure 10: *Fraction of Monte Carlo estimates with squared error above the 95th quantile of the stationary error distribution (Section 4.1) when varying K. For K between 8 and 256, the fraction approaches the expected 0.05 (horizontal dotted line) for stationary Markov chains, past the prescribed threshold (vertical blue line).*

remains an outstanding challenge. This paper tackles the problem of assessing approximate convergence. We propose a new convergence diagnostic, $\widehat{R}_{\mathfrak{n}}$, which is straightforward to implement for a broad class of MCMC algorithms and works for both long and short chains, in the sense that $\widehat{R}$ works for long chains. A small $\widehat{R}_{\mathfrak{n}}$ (or $\widehat{R}$) does not guarantee convergence to the stationary distribution or that the bias has decayed to 0. Still $\widehat{R}$ has empirically proven its usefulness in applied statistics, machine learning, and many scientific disciplines. Our analysis reveals that potential success (or failure) of $\widehat{R}_{\mathfrak{n}}$ and $\widehat{R}$ is best understood by studying (i) the relation between the nonstationary variance and the squared bias, and (ii) how well $\widehat{R}_{\mathfrak{n}}$ monitors the nonstationary variance.

In addition to working in the many-short-chains regime, $\widehat{R}_{\mathfrak{n}}$ provides more guidance to choose a threshold, notably in the $N = 1$ case (Corollary 3.5). Unlike $\widehat{R}$, the proposed $\widehat{R}_{\mathfrak{n}}$ requires a partition of the chains into superchains. No choice of partition uniformly minimizes the variance of $\widehat{R}_{\mathfrak{n}}$ during all phases of MCMC. Based on our numerical experiments, we recommend using $K = 8 - 256$ initializations, when running $KM = 2048$ chains.

The nesting design we introduce opens the prospect of generalizing other variations on $\widehat{R}$, including multivariate $\widehat{R}$ (Brooks and Gelman, 1998; Vats and Knudson, 2021; Moins et al., 2023), rank-normalized $\widehat{R}$ (Vehtari et al., 2021) and local $\widehat{R}$ (Moins et al., 2023). Nesting can further be used for less conventional convergence diagnostics, such as R^*, which uses classification trees to compare different chains (Lambert and Vehtari, 2022).

A direction for future work is to adaptively set the warmup length using $\widehat{R}_{\mathfrak{n}}$. This would follow a long tradition of using diagnostics to do early stopping of MCMC (Geweke, 1992; Cowles and Carlin, 1996; Cowles et al., 1998; Jones et al., 2006; Zhang et al., 2020). Still, standard practice remains to prespecify the warmup length. This means the warmup length is rarely optimal, which is that much more exasperating in the many-short-chains regime, where the warmup dominates the computation.

6 Acknowledgments

We thank the TensorFlow Probability team at Google, especially Alexey Radul. We also thank Marylou Gabrié and Sam Livingstone for helpful discussions; Rif A. Saurous, Andrew Davison, Owen Ward, Mitzi Morris, and Lawrence Saul for helpful comments on the manuscript; and the U.S. Office of Naval Research and Research Council of Finland for partial support. LRD was supported by the EPSRC grant EP/R034710/1. Much of this work was done while CM was at Google Research and in the Department of Statistics at Columbia University, and while LRD was in the Department of Statistics at the University of Warwick.

References

Andrieu, C. and Thoms, J. (2008). "A tutorial on adaptive MCMC." *Statistics and Computing*, 18: 343–376. 7

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). "Optimal tuning of the hybrid Monte Carlo algorithm." *Bernoulli*, 19(5A): 1501–1534. 12

Betancourt, M. (2018). "A conceptual introduction to Hamiltonian Monte Carlo." *arXiv:1701.02434v1*. 6

Brooks, S. P. and Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7: 434–455. 2, 23

Cowles, M. K. and Carlin, B. P. (1996). "Markov chain Monte Carlo convergence diagnostics: A comparative review." *Journal of the American Statistical Association*, 91: 883–904. 2, 23

Cowles, M. K., Roberts, G. O., and Rosenthal, J. S. (1998). "Possible biases induced by MCMC convergence diagnostics." *Journal of Statistical Computation and Simulation*, 64: 87–104. 23

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers." *Journal of the Royal Statistical Society, Series B*, 68: 411–436. 6

du Ché, S. and Margossian, C. C. (2023). "Parallelization for Markov chain Monte Carlo with heterogeneous runtimes." *BayesComp,* `https://charlesm93.github.io/files/Bayescomp_ode_chains.pdf`. 18

Dua, D. and Graff, C. (2017). "UCL machine learning repository." URL `http://archive.ics.ucl.edu/ml` 16, 38

Flegal, J. M., Haran, M., and Jones, G. L. (2008). "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science*, 250–260. 15

Gardiner, C. W. (2004). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, 3rd edition*. Springer-Verlag, Berlin. 31

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, 3rd edition*. CRC Press. 2, 15

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms." *Annals of Applied Probability*, 7(1): 110–120. 12

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel-Hierarchical Models*. Cambridge University Press. 16

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences (with discussion)." *Statistical Science*, 7: 457–511. 2, 9, 12

Gelman, A. and Shirley, K. (2011). "Inference from simulations and monitoring con-

vergence." In *Handbook of Markov chain Monte Carlo*, chapter 4. CRC Press. 2, 15

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments." In *Bayesian Statistics 4*, 169–193. Oxford University Press. 23

Gilks, W. R., Roberts, G. O., and George, E. I. (1994). "Adaptive direction sampling." *Journal of the Royal Statistical Society: Series D*, 43(1): 179–189. 7

Glynn, P. W. and Rhee, C.-H. (2014). "Exact estimation for Markov chain equilibrium expectations." *Journal of Applied Probability*, 51: 377–389. 6

Heng, J. and Jacob, P. E. (2019). "Unbiased Hamiltonian Monte Carlo with couplings." *Biometrika*, 106: 287 – 302. 6

Hoffman, M. and Sountsov, P. (2022). "Tuning-free generalized Hamiltonian Monte Carlo." *Artificial Intelligence and Statistics*. 1, 7

Hoffman, M. D. and Gelman, A. (2014). "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15: 1593–1623. 1

Hoffman, M. D., Radul, A., and Sountsov, P. (2021). "An adaptive MCMC scheme for setting trajectory lengths in Hamiltonian Monte Carlo." *Artificial Intelligence and Statistics*. 1, 3, 17, 18, 37

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). "Unbiased Markov chain Monte Carlo methods with couplings." *Journal of the Royal Statistical Society, Series B*, 82: 543–600. 6

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). "Fixed-width output analysis for Markov chain Monte Carlo." *Journal of the American Statistical Association*, 101: 1537–1547. 23

Lambert, B. and Vehtari, A. (2022). "R^*: A robust MCMC convergence diagnostic with uncertainty using decision tree classifiers." *Bayesian Analysis*, 17: 353–379. 23

Lao, J., Suter, C., Langmore, I., Chimisov, C., Saxena, A., Sountsov, P., Moore, D., Saurous, R. A., Hoffman, M. D., and Dillon, J. V. (2020). "tfp.mcmc: Modern Markov chain Monte Carlo tools built for modern hardware." *arXiv:2002.01184*. 1, 2

Mackay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. 15

Margossian, C. C. and Gelman, A. (2023). "For how many iterations should we run Markov chain Monte Carlo." *arXiv:2311.02726*. 15

Margossian, C. C., Zhang, Y., and Gillespie, W. R. (2022). "Flexible and efficient Bayesian pharmacometrics modeling using Stan and Torsten, part I." *CPT: Pharmacometrics & Systems Pharmacology*, 11: 1151–1169. 16

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. (2023). "On the use of a local $\widehat{R}$ to improve MCMC convergence diagnostic." *Bayesian Analysis*. 6, 23

Neal, R. M. (2001). "Annealed importance sampling." *Statistics and Computing*, 11: 125–139. 6

— (2012). "MCMC using Hamiltonian dynamics." In *Handbook of Markov Chain Monte Carlo*. CRC Press. 6, 10

Nguyen, T. D., Trippe, B. L., and Broderick, T. (2022). "Many processors, little time: MCMC for partitions via optimal transport couplings." *Artificial Intelligence and Statistics*, 151: 3483–3514. 6

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). "A general framework for the parametrization of hierarchical models." *Statistical Science*, 22: 59–73. 16

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. (2022). "Optimal thinning of MCMC output." *Journal of the Royal Statistical Society: Series B*, 84: 1059–1081. 7

Riou-Durand, L., Sountsov, P., Vogrinc, J., Margossian, C. C., and Power, S. (2023). "Adaptive tuning for Metropolis adjusted Langevin trajectories." *Artificial Intelligence and Statistics*. 1

Riou-Durand, L. and Vogrinc, J. (2022). "Metropolis adjusted Langevin trajectories: A robust alternative to Hamiltonian Monte Carlo." *arXiv:2202.13230*. 12

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer. 2

Roberts, G. O. and Rosenthal, J. S. (1998). "Optimal scaling of discrete approximations to Langevin diffusions." *Journal of the Royal Statistical Society, Series B*, 60: 255–268. 12

— (2004). "General state space Markov chains and MCMC algorithms." *Probability Surveys*, 1: 20 – 71. 2, 7

Rosenbrock, H. H. (1960). "An automatic method for finding the greatest or least value of a function." *Computer Journal*, 3: 175–184. 16

Rosenthal, J. S. (2000). "Parallel computing and Monte Carlo algorithms." *Far East Journal of Theoretical Statistics*, 4: 207–236. 2

Rubin, D. B. (1981). "Estimation in parallelized randomized experiments." *Journal of Educational Statistics*, 6: 377–400. 16, 38

Sountsov, P. and Hoffman, M. D. (2021). "Focusing on difficult directions for learning HMC trajectory lengths." *arXiv:2110.11576*. 1

Sountsov, P., Radul, A., and contributors (2020). "Inference Gym." URL https://pypi.org/project/inference_gym 37

South, L. F., Riabiz, M., Teymur, O., and Oates, C. J. (2021). "Post-processing of MCMC." *Annual Review of Statistics and its Application*, 9: 1–30. 7

TensorFlow Probability Development Team (2023). "TensorFlow Probability." URL https://www.tensorflow.org/probability 17

Vats, D., Flegal, J. M., and Jones, G. L. (2019). "Multivariate output analysis for Markov chain Monte Carlo." *Biometrika*, 106: 321–337. 15

Vats, D. and Knudson, D. (2021). "Revisiting the Gelman-Rubin diagnostic." *Statistical Science*, 36: 518–529. 6, 11, 23

Vehtari, A. (2022). "Bayesian workflow book - Digits."
URL https://avehtari.github.io/casestudies/Digits/digits.html 15

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion)." *Bayesian Analysis*, 16: 667–718. 2, 3, 6, 9, 12, 23

Wakefield, J. (1996). "The Bayesian analysis of population pharmacokinetic models." *Journal of the American Statistical Association*, 91: 62–75. 16

Zhang, Y., Gillespie, B., Bales, B., and Vehtari, A. (2020). "Speed up population Bayesian inference by combining cross-chain warmup and within-chain parallelization." In *American Conference on Pharmacometrics*. 23

Appendix

Appendix A: Proofs

Here we provide the proofs for formal statements throughout the paper.

A.1 Proofs for Section 2: "Nested $\widehat{R}$"

Proof of Lemma 3.3: Asymptotic limit of $\widehat{W}_{\mathbf{n}}$

Proof. Recall the superchains are independent. Then applying the law of large numbers along K yields,

$$\widehat{W}_{\mathbf{n}} \xrightarrow[K\to\infty]{a.s} \mathbb{E}_\Gamma \tilde{B}_k + \mathbb{E}_\Gamma \tilde{W}_k.$$

Now

$$\begin{aligned}
\mathbb{E}_\Gamma \tilde{B}_k &= \frac{1}{M-1}\sum_{m=1}^{M} \mathbb{E}_\Gamma \left(\bar{f}^{(\cdot mk)} - \bar{f}^{(\cdot\cdot k)}\right)^2 \\
&= \frac{1}{M-1}\sum_{m=1}^{M} \left(\mathbb{E}_\Gamma[\bar{f}^{(\cdot mk)}]^2 + \mathbb{E}_\Gamma[\bar{f}^{(\cdot\cdot k)}]^2 - 2\mathbb{E}_\Gamma(\bar{f}^{(\cdot mk)}\bar{f}^{(\cdot\cdot k)})\right),
\end{aligned}$$

and

$$\begin{aligned}
\sum_{m=1}^{M} \mathbb{E}_\Gamma(\bar{f}^{(\cdot mk)}\bar{f}^{(\cdot\cdot k)}) &= M\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_\Gamma(\bar{f}^{(\cdot mk)}\bar{f}^{(\cdot\cdot k)}) \\
&= M\mathbb{E}_\Gamma\left(\frac{1}{M}\sum_{m=1}^{M} \bar{f}^{(\cdot mk)}\bar{f}^{(\cdot\cdot k)}\right) \\
&= M\mathbb{E}_\Gamma[\bar{f}^{(\cdot\cdot k)}]^2.
\end{aligned}$$

Plugging this back in yields,

$$\begin{aligned}
\mathbb{E}_\Gamma \tilde{B}_k &= \frac{1}{M-1}\sum_{m=1}^{M} \left(\mathbb{E}_\Gamma[\bar{f}^{(\cdot mk)}]^2 - \mathbb{E}_\Gamma[\bar{f}^{(\cdot\cdot k)}]^2\right) \\
&= \frac{1}{M-1}\sum_{m=1}^{N} \left(\mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)} + [\mathbb{E}_\Gamma \bar{f}^{(\cdot mk)}]^2 - \mathrm{Var}_\Gamma \bar{f}^{(\cdot\cdot k)} - [\mathbb{E}_\Gamma \bar{f}^{(\cdot\cdot k)}]^2\right) \\
&= \frac{1}{M-1}\sum_{m=1}^{N} \left[\mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)} - \mathrm{Var}_\Gamma \bar{f}^{(\cdot\cdot k)}\right] + \left[(\mathbb{E}_\Gamma \bar{f}^{(\cdot mk)})^2 - (\mathbb{E}_\Gamma \bar{f}^{(\cdot\cdot k)})^2\right].
\end{aligned}$$

The second term vanishes, because $\mathbb{E}_\Gamma \bar{f}^{(\cdot mk)} = \mathbb{E}_\Gamma \bar{f}^{(\cdot\cdot k)}$. Then

$$\mathbb{E}\tilde{B}_k = \frac{1}{M-1}\sum_{m=1}^{M} \mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)} - \mathrm{Var}_\Gamma \bar{f}^{(\cdot\cdot k)}$$

$$= \frac{M}{M-1}\left(\mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)} - \mathrm{Var}_\Gamma \bar{f}^{(\cdot\cdot k)}\right),$$

We next apply the law of total variance:

$$\mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)} = \mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k) + \mathrm{Var}_{p_0}\mathbb{E}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k),$$

and

$$\begin{aligned}\mathrm{Var}_\Gamma \bar{f}^{(\cdot\cdot k)} &= \mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k) + \mathrm{Var}_{p_0}\mathbb{E}_\gamma(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k) \\ &= \frac{1}{M}\mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k) + \mathrm{Var}_{p_0}\mathbb{E}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k),\end{aligned}$$

where the second line follows from noting that, conditional on θ_0^k, the chains are independent, and that $\mathbb{E}_\gamma(\bar{f}^{(\cdot\cdot k)} \mid \theta_0^k) = \mathbb{E}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)$. Plugging this result back, we get

$$\begin{aligned}\mathbb{E}\tilde{B}_k &= \frac{M}{M-1}\left[\mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k) - \frac{1}{M}\mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)\right] \\ &= \mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k).\end{aligned}$$

Next, if $N = 1$, $\tilde{W} = 0$. If $N > 1$, we obtain, following a similar approach as above,

$$\begin{aligned}\tilde{W}_k &= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N-1}\sum_{n=1}^{N}\left(f^{(nmk)} - \bar{f}^{(\cdot mk)}\right)^2 \\ &= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N-1}\sum_{n=1}^{N}\left(f^{(nmk)}\right)^2 - \left(\bar{f}^{(\cdot mk)}\right)^2.\end{aligned}$$

Then, taking expectations and expanding the square inside the expectation,

$$\mathbb{E}_\Gamma \tilde{W}_k = \begin{cases}\frac{1}{N-1}\sum_{n=1}^{N}\mathrm{Var}_\Gamma f^{(nmk)} - \mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)} + (\mathbb{E}_\Gamma f^{(nmk)})^2 - (\mathbb{E}_\Gamma f^{(\cdot mk)})^2 & \text{if } N > 1 \\ 0 & \text{if } N = 1,\end{cases} \tag{29}$$

with the right side corresponding to our definition of W'. Thus

$$W_{\mathfrak{n}} = \mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k) + W',$$

as desired. □

Remark A.1. *The second term on the right side of eq. (29) is a drift term: the (expected) sample variance increases because the samples do not have the same mean.*

Proof of Corollary 3.4: stationary lower bound for $\widehat{R}_{\mathfrak{n}}$

Proof. For $\mathcal{W} \to \infty$, the chains are stationary and have "forgotten" their initialization, that is $\mathrm{Var}_{p_0}\mathbb{E}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k) = 0$. Thus

$$\mathbb{E}_\Gamma \tilde{B}_k = \mathbb{E}_{p_0}\mathrm{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k) = \mathrm{Var}_\Gamma \bar{f}^{(\cdot mk)}.$$

Also, because the samples are now all identically distributed, the drift term in $\mathbb{E}_\Gamma \tilde{W}$ goes to 0. We also have for any n that $\text{Var} f^{(nmk)} = \sigma^2$, where the constant σ^2 is the variance of the stationary distribution, p. Furthermore, due the chain's positive autocorrelation,

$$\text{Var}_\Gamma \bar{f}^{(\cdot mk)} \geq \frac{\sigma^2}{N}. \tag{30}$$

Thus

$$W_{\mathfrak{n}} \leq \frac{1}{N-1} \text{Var}_\Gamma \bar{f}^{(\cdot mk)} + \frac{N}{N-1}\sigma^2.$$

Then

$$\frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}} \geq \frac{\text{Var}_\Gamma \bar{\theta}^{(\cdot mk)}}{M\left(\frac{1}{N-1}\text{Var}_\Gamma \bar{\theta}^{(\cdot mk)} + \frac{N}{N-1}\sigma^2\right)}.$$

Noting that for stationary chains $\text{ESS}_{(1)} = \sigma^2 / \text{Var}_\Gamma \bar{\theta}^{(\cdot mk)}$,

$$\frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}} \geq \frac{1}{M\left(\frac{1+N\text{ESS}_{(1)}}{N-1}\right)} == \frac{1}{M}\frac{1-1/N}{\text{ESS}_{(1)} + 1/N}.$$

□

Proof of Corollary 3.5: correction for persistent variance when $N = 1$

The proof of Corollary 3.5 follows from Theorem 3.1 and Lemma 3.3.

Proof. When $N = 1$, $\tilde{W}_k = 0$. Thus $W_{\mathfrak{n}} = \mathbb{E}_{p_0} \text{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)$ and

$$\begin{aligned}
\frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}} &= \frac{\mathbb{E}_{p_0}\text{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)}{M\mathbb{E}_{p_0}\text{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)} + \frac{\text{Var}_{p_0}\mathbb{E}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)}{\mathbb{E}_{p_0}\text{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)} \\
&= \frac{1}{M} + \frac{\text{Var}_{p_0}\mathbb{E}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)}{\mathbb{E}_{p_0}\text{Var}_\gamma(\bar{f}^{(\cdot mk)} \mid \theta_0^k)}.
\end{aligned}$$

Finally, $\bar{f}^{(\cdot mk)} = f^{(1mk)}$, given that $N = 1$. □

Proof of Theorem 3.8: bias and $B_{\mathfrak{n}}/W_{\mathfrak{n}}$ in continuous limit

We prove Theorem 3.8 which gives us an exact expression for the bias and the ratio $B_{\mathfrak{n}}/W_{\mathfrak{n}}$. The Monte Carlo estimator $\bar{X}_T$ is the average of M diffusion processes evaluated at time T. The processes are initialized at the same point $X_0 \sim p_0$ but then run independently, according to the Langevin diffusion process targeting $p = \text{normal}(\mu, \sigma)$.

Proof. Let Ψ denote the stochastic process which generates X_T, and further break this process into (i) p_0, the process which draws x_0 and (ii) ψ the process which generates

X_T conditional on x_0. Following the same arguments as in the previous sections of the Appendix, we leverage Corollary 3.5,

$$\frac{B_{\mathfrak{n}}}{W_{\mathfrak{n}}} = \frac{1}{M} + \frac{\text{Var}_{p_0}\mathbb{E}_{\psi}(X_T \mid X_0)}{\mathbb{E}_{p_0}\text{Var}_{\psi}(X_T \mid X_0)}.$$

It is well known that Ornstein-Uhlenbeck SDEs, such as the Langevin diffusion SDE targeting a Gaussian, admit explicit solutions; see Gardiner (2004). The solution of eq. (20) given $X_0 = x_0$ is given for $T > 0$ by

$$X_T = e^{-T}x_0 + \mu(1 - e^{-T}) + \sqrt{2\sigma}\int_0^T e^{-(T-s)}\text{d}W_s. \tag{31}$$

Now, integrating with respect to $X_0 \sim p_0 = \text{normal}(\mu_0, \sigma_0)$ yields

$$\mathbb{E}_{p_0}\mathbb{E}_{\psi}(X_T \mid X_0) = \mu_0 e^{-T} + \mu(1 - e^{-T}),$$

then

$$\text{Var}_{p_0}\mathbb{E}_{\psi}(X_T \mid X_0) = \text{Var}_{p_0}\left(X_0 e^{-T} + \mu(1 - e^{-T})\right) = e^{-2T}\sigma_0^2,$$

and

$$\mathbb{E}_{p_0}\text{Var}_{\psi}(X_T \mid X_0) = \mathbb{E}_{p_0}\sigma^2(1 - e^{-2T}) = \sigma^2(1 - e^{-2T}),$$

from which the desired result follows. □

Proof of Corollary 3.9: initialization conditions under which $\widehat{R}_{\mathfrak{n}}$ is reliable

Theorem 3.9 provides conditions in the continuous limit under which $\widehat{R}_{\mathfrak{n}}$ is (δ, δ')-reliable and formalizes the notion of overdispersed intializations.

Proof. The bias is given by

$$\mathbb{E}(X_T) - \mathbb{E}_p(\mu) = (\mu_0 - \mu)e^{-T},$$

which is a monotone decreasing function of T. The time at which the scaled squared bias is below δ' is obtained by solving

$$\frac{(\mu_0 - \mu)^2 e^{-2T}}{\sigma^2} \leq \delta'.$$

If $(\mu_0 - \mu)^2/\sigma^2 \leq \delta'$, then the above condition is verified for any T and $\widehat{R}_{\mathfrak{n}}$ is trivially (δ, δ')-reliable. Suppose now that $(\mu_0 - \mu)^2/\sigma^2 > \delta'$. Then we require

$$T \geq \frac{1}{2}\log\left(\frac{(\mu_0 - \mu)^2}{\delta'\sigma^2}\right) \triangleq T^*.$$

It remains to ensure that for $T < T^*$, $B_{\mathfrak{n}}/W_{\mathfrak{n}} > \delta$. Plugging in T^* in the expression from Lemma 3.8 and noting $B_{\mathfrak{n}}/W_{\mathfrak{n}}$ is monotone decreasing in T, we have

$$\sigma_0^2 > \left(\delta - \frac{1}{M}\right)\left(e^{2T^*} - 1\right)\sigma^2,$$

which is the wanted expression. □

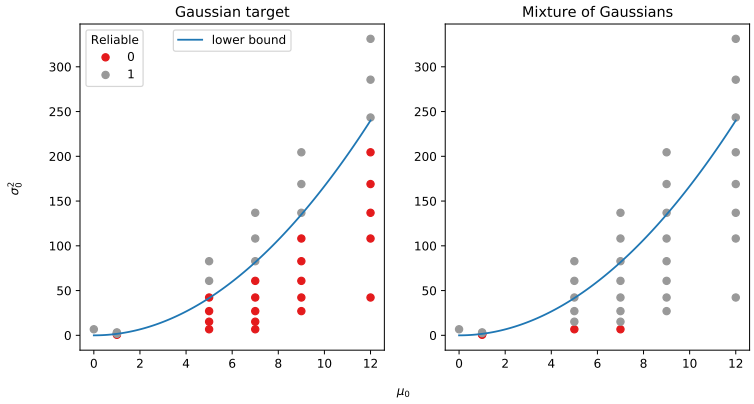

Figure 11: (δ, δ')*-reliability of* $\widehat{R}_{\mathfrak{n}}$ *for varying initial bias* μ_0 *and initial variance* σ_0^2.

Appendix B: Additional results on the reliability of $\widehat{R}_{\mathfrak{n}}$ and $\widehat{R}$

This appendix provides additional results on the reliability of $\widehat{R}_{\mathfrak{n}}$ (Section 3.2). We numerically test the lower bound on the initial variance, provided by Corollary 3.9, on a Gaussian target and a mixture of two Gaussians. We then derive a reliability condition for $\widehat{R}$, tackling the $N > 1$ case where we have more than one sample per chain.

B.1 Numerical evaluation of the reliability of $\widehat{R}_{\mathfrak{n}}$

We numerically evaluate the (δ, δ')-reliability of $\widehat{R}_{\mathfrak{n}}$ on two examples:

(i) a standard Gaussian, which conforms to the assumption of our theoretical analysis.

(ii) a mixture of two Gaussians, which constitutes a canonical example where $\widehat{R}$ potentially fails. In this example the Markov chains fail to mix, hence reliability is achieved if $\widehat{R}_{\mathfrak{n}} \geq \sqrt{1+\delta}$.

We approximate the Langevin diffusion using the Metropolis adjusted Langevin algorithm (MALA) algorithm, which is equivalent to HMC with a one-step leapfrog integrator. The step size is 0.04, which is chosen to be as small as possible while ensuring that $\widehat{R}_{\mathfrak{n}}$ reports convergence after $\sim 2 \times 10^4$ iterations for the standard Gaussian target. We set $M = 16$, $\delta \approx 0.1$ and $\delta' = \delta/5$. Reliability is defined in terms of $B_{\mathfrak{n}}$ and $W_{\mathfrak{n}}$, which are the asymptotic limits of $\widehat{B}_{\mathfrak{n}}$ and $\widehat{W}_{\mathfrak{n}}$ when $K \to \infty$. We approximate this limit by using $K = 1024$ superchains for a total of 16384 chains.

The results are shown in Figure 11. The theoretical lower bound (Theorem 3.9) is accurate when using a standard Gaussian target. When targeting a mixture of Gaussians, the lower bound is too conservative and $\widehat{R}_{\mathfrak{n}}$ is reliable even when using an "underdispersed" initialization. This is because we use a large number of distinct initializations and, even with a small initial variance, we typically find both modes and identify poor mixing. This suggests the failure of $\widehat{R}$ on multimodal targets is often due to using too few distinct initializations, rather than an underdispersed initialization.

B.2 Reliability condition for $\widehat{R}$ in the continuous limit

We here conduct an analysis for $\widehat{R}$ similar to the one conducted for $\widehat{R}_{\mathfrak{n}}$ in Section 3.2. That is we examine the reliability of $\widehat{R}$ when approximating the MCMC chain by a Langevin diffusion which targets a Gaussian.

There are three differences when compared to our study of $\widehat{R}_{\mathfrak{n}}$: (i) each Monte Carlo estimator is now made up of only one chain and all chains are independent, (ii) the chains do not include warmup, i.e., $\mathcal{W} = 0$, and (iii) the length of each chain is chosen as $N = \lfloor T/h \rfloor$ for a small value step size h such that for each chain m, the distribution of $\theta^{(\lfloor t/h \rfloor m)}$ can be approximated by the Langevin solution defined in (31). The distribution of each (within chain) estimator can therefore be approximated by the distribution of

$$\overline{X}_T = \frac{1}{T} \int_0^T X_s \mathrm{d}s. \tag{32}$$

In this framework, the limits of B and W as $h \to 0$ yield

$$B = \mathrm{Var}_\Psi(\bar{X}_T), \qquad W = \frac{1}{T} \int_0^T \mathbb{E}\left[\left(X_t - \bar{X}_T \right)^2 \right].$$

This next lemma provides an exact expression for B/W.

Lemma B.1. *Suppose we initialize a process at $X_0 \sim p_0$, which evolves according to (20) from time $t = 0$ to $t = T > 0$, and let $\bar{X}_T$ be defined as above. Denote its distribution as Ψ. Let*

$$\begin{aligned} \rho_T &\triangleq \frac{1}{T}(1 - e^{-T}), \\ \xi_T &\triangleq \frac{1}{2T}\left(1 - e^{-2T}\right). \\ \eta_T &\triangleq \frac{2}{T}(1 - \rho_T). \end{aligned}$$

Then

$$\mathbb{E}_\Psi \bar{X}_T - \mathbb{E}_p X = (\mu_0 - \mu)\rho_T, \tag{33}$$

and

$$\frac{B}{W} = \frac{(\sigma_0^2 - \sigma^2)\rho_T^2 + \sigma^2 \eta_T}{(\sigma_0^2 - \sigma^2 + (\mu_0 - \mu)^2)(\xi_T - \rho_T^2) + \sigma^2(1 - \eta_T)}. \tag{34}$$

Proof. Let Ψ denote the stochastic process which generates X_t. Once again, we exploit the explicit solution (31) to the Langevin SDE. We begin with the numerator.

$$\begin{aligned}
B &= \mathrm{Var}_\Psi(\bar{X}_T) \\
&= \frac{1}{T^2}\int_0^T\int_0^T \mathrm{Cov}_\Psi(X_s, X_t)\mathrm{d}s\mathrm{d}t \\
&= \frac{1}{T^2}\int_0^T\int_0^T \left((\sigma_0^2-\sigma^2)e^{-(s+t)} + \sigma^2 e^{-|s-t|}\right)\mathrm{d}s\mathrm{d}t \\
&= (\sigma_0^2-\sigma^2)\rho_T^2 + 2\left(\frac{\sigma^2}{T} + (-1+e^{-T})\frac{\sigma^2}{T^2}\right) \\
&= (\sigma_0^2-\sigma^2)\rho_T^2 + \sigma^2\eta_T.
\end{aligned}$$

Next we have

$$W = \frac{1}{T}\int_0^T \mathbb{E}\left[\left(X_t - \bar{X}_T\right)^2\right].$$

Following the same steps to prove Lemma 3.3, we have

$$W = \frac{1}{T}\int_0^T \left[\mathrm{Var}_\Psi(X_t) - \mathrm{Var}_\Psi(\bar{X}_T)\right] + \left[(\mathbb{E}_\Psi(X_t))^2 - (\mathbb{E}_\Psi(\bar{X}_T))^2\right]\mathrm{d}t.$$

Computing each term yields

$$\begin{aligned}
\mathbb{E}_\Psi(X_t) &= \mu + (\mu_0-\mu)e^{-t} \\
\mathrm{Var}_\Psi(X_t) &= \sigma^2 + (\sigma_0^2-\sigma^2)e^{-2t} \\
\mathbb{E}_\Psi(\bar{X}_T) &= \mu + (\mu_0-\mu)\rho_T \\
\mathrm{Var}_\Psi(\bar{X}_T) &= \sigma^2\eta_T + (\sigma_0^2-\sigma^2)\rho_T^2
\end{aligned}$$

Constructing W term by term,

$$\begin{aligned}
\frac{1}{T}\int_0^T (\mathrm{Var}_\Psi(X_t) - \mathrm{Var}_\Psi(\bar{X}_T))\mathrm{d}t &= \frac{1}{T}\int_0^T \sigma^2(1-\eta_T) + (\sigma^2-\sigma_0^2)(e^{-2t}-\rho_T^2)\mathrm{d}t \\
&= \sigma^2(1-\eta_T) + (\sigma_0^2-\sigma^2)(\xi_T - \rho_T^2),
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1}{T}\int_0^T ((\mathbb{E}_\Psi(X_t))^2 - (\mathbb{E}_\Psi(\bar{X}_T))^2)\mathrm{d}t &= \frac{1}{T}\int_0^T (\mu_0-\mu)(e^{-t}-\rho_T)(2\mu + (\mu_0-\mu)(e^{-t}+\rho_T))\mathrm{d}t \\
&= (\mu_0-\mu)^2(\xi_T - \rho_T^2).
\end{aligned}$$

Putting it all together, we have

$$W = (\sigma_0^2 - \sigma^2 + (\mu_0-\mu)^2)(\xi_T - \rho_T^2) + \sigma^2(1-\eta_T).$$

□

Remark B.2. *Taking the limit at* $T \to 0$ *yields*

$$\lim_{T\to 0} \rho_T = \lim_{T\to 0} \xi_T = \lim_{T\to 0} \eta_T = 1$$

Thus $\lim_{T\to 0} B = \sigma_0^2$ *and* $\lim_{T\to 0} W = 0$*, therefore* $\lim_{T\to 0} B/W = +\infty$.

The above limits can be calculated by Taylor expanding the exponential.

We now state the main result of this section, which provides a lower bound on σ_0^2 in order to insure $\widehat{R}$ is (δ, δ')-reliable. Unlike in the $\widehat{R}_{\mathfrak{n}}$ case the proof requires some additional assumptions.

Theorem B.3. *If* $(\mu-\mu_0)^2/\sigma^2 \leq \delta'$*, then* $\widehat{R}$ *is always* (δ, δ')*-reliable. Suppose now that* $(\mu - \mu_0)^2/\sigma^2 > \delta'$*. Let* T^* *solve*

$$\frac{(\mu - \mu_0)^2 \rho_T^2}{\sigma^2} = \delta',$$

for T*. Assume:*

(A1) B/W *is monotone decreasing (conjecture: this is always true).*

(A2) δ *verifies the upper bound*

$$\delta < \frac{1}{\frac{1}{2}T^* \coth\left(\frac{T^*}{2}\right) - 1}$$

where coth *is the hyperbolic cotangent.*

Then $\widehat{R}$ *is* (δ, δ')*-reliable if and only if*

$$\sigma_0^2 \geq \frac{\delta(\xi_{T^*} - \rho_{T^*}^2)(\mu - \mu_0)^2 + [\delta(1 + \rho_{T^*}^2 - \eta_{T^*} - \xi_{T^*}) - (\eta_{T^*} - \rho_{T^*}^2)]\sigma^2}{(1+\delta)\rho_{T^*}^2 - \delta\xi_{T^*}}. \tag{35}$$

Proof. When $(\mu - \mu_0)^2/\sigma^2 \leq \delta'$, (δ, δ')-reliability follows from the fact ρ_T and thence the bias are monotone decreasing.

Consider now the case where $(\mu - \mu_0)^2/\sigma^2 > \delta'$. To alleviate the notation, assume without loss of generality that $\mu = 0$. Per Assumption (A1), it suffices to check that for $T = T^*$, $B/W \geq \delta$. Per Lemma B.1, this is equivalent to

$$\frac{(\sigma_0^2 - \sigma^2)\rho_{T^*}^2 + \sigma^2\eta_{T^*}}{(\sigma_0^2 - \sigma^2 + \mu_0^2)(\xi_{T^*} - \rho_{T^*}^2) + \sigma^2(1 - \eta_{T^*})} \geq \delta$$

$$\iff \quad \frac{\sigma_0^2\rho_{T^*}^2 + (\eta_{T^*} - \rho_{T^*}^2)\sigma^2}{\sigma_0^2(\xi_{T^*} - \rho_{T^*}^2) + (\mu_0^2 - \sigma^2)(\xi_{T^*} - \rho_{T^*}^2) + \sigma^2(1 - \eta_{T^*})} \geq \delta$$

$$\begin{aligned}
\iff \qquad & \sigma_0^2(\rho_{T^*}^2 + \delta(\rho_{T^*}^2 - \xi_{T^*})) && \geq \delta\left[(\mu_0^2 - \sigma^2)(\xi_{T^*} - \rho_{T^*}^2)\right] + \delta(1 - \eta_{T^*})\sigma^2 \\
& && \quad -(\eta_{T^*} - \rho_{T^*}^2)\sigma^2 \\
\iff \qquad & \sigma_0^2[(1+\delta)\rho_{T^*}^2 - \delta\xi_{T^*}] && \geq \delta\mu_0^2(\xi_{T^*} - \rho_{T^*}^2) \\
& && \quad +[\delta(1 + \rho_{T^*}^2 - \eta_{T^*} - \xi_{T^*}) - (\eta_{T^*} - \rho_{T^*}^2)]\sigma^2.
\end{aligned}$$

To complete the proof, we need to show that $\left((1+\delta)\rho_{T^*}^2 - \delta\xi_{T^*}\right)$ is positive. This will not always be true, hence the requirement for Assumption (A2). We arrive at this condition by expressing ξ_T in terms of ρ_T^2.

$$\begin{aligned}
\xi_T &= \frac{\xi_T}{\rho_T}\rho_T \\
&= \frac{1}{2}\left(\frac{1 - e^{-2T}}{1 - e^{-T}}\right)\rho_T \\
&= \frac{1}{2}\left(\frac{1 - e^{-T} + e^{-T} - e^{-2T}}{1 - e^{-T}}\right)\rho_T \\
&= \frac{1}{2}\left(1 + e^{-T}\frac{1 - e^{-t}}{1 - e^{-T}}\right)\rho_T \\
&= \frac{1}{2}(1 + e^{-T})\rho_T \\
&= \frac{1}{2}\left(\frac{1 + e^{-T}}{\rho_T}\right)\rho_T^2 \\
&= \frac{1}{2}\left(T\frac{1 + e^{-T}}{1 - e^{-T}}\right)\rho_T^2 \\
&= \frac{1}{2}T\coth(T/2)\rho_T^2.
\end{aligned}$$

Thus

$$(1+\delta)\rho_{T^*}^2 - \delta\xi_{T^*} = \rho_{T^*}^2\left[1 + \delta - \frac{\delta}{2}T\coth(T/2)\right],$$

which by assumption (A2) is positive. □

Remark B.4. *On the right side of* (35)*, all terms in parenthesis in the numerator are positive, meaning the numerator comprises a positive term scaled by* δ*,*

$$\delta\left[(\xi_{T^*} - \rho_{T^*}^2)(\mu - \mu_0)^2 + (1 + \rho_{T^*}^2 - \eta_{T^*} - \xi_{T^*})\sigma^2\right],$$

and a negative term,

$$-(\eta_{T^*} - \rho_{T^*}^2)\sigma^2.$$

This second term appears in the expression for B *(Lemma B.1), which we can rewrite as*

$$B = \sigma_0^2\rho_T^2 + \sigma^2(\eta_T - \rho_T^2) \geq \sigma^2(\eta_T - \rho_T^2).$$

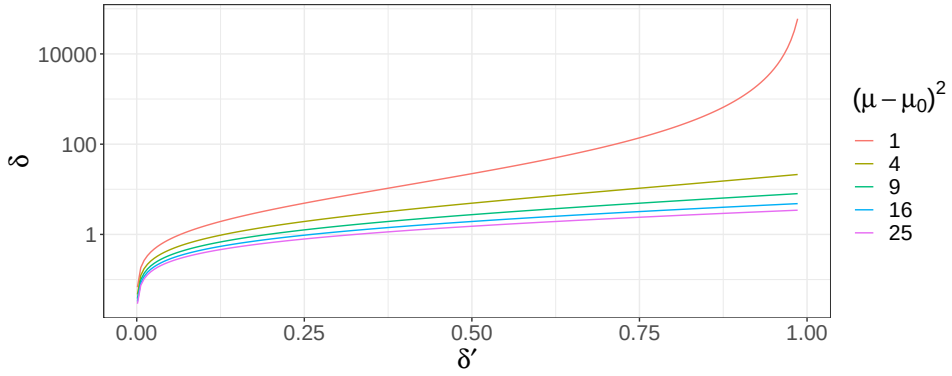

Figure 12: *Upper bound on δ to verify Assumption (A2) in Theorem B.3*

This lower bound does not cancel with W, ensuring that B/W is nonzero. Hence for

$$\delta \leq \frac{(\eta_{T^*} - \rho_{T^*}^2)\sigma^2}{(\xi_{T^*} - \rho_{T^*}^2)(\mu - \mu_0)^2 + (1 + \rho_{T^*}^2 - \eta_{T^*} - \xi_{T^*})\sigma^2}$$

the reliability condition is always met, including even when $\sigma_0^2 = 0$. This is comparable to the $\delta < 1/M$ case in Theorem 3.9.

We can understand Assumption (A2) as a requirement that δ be not too large compared to δ', since a smaller δ' implies a larger T^*. *Why* such a requirement exists remains conceptually unclear. We simulate the upper bound for $\delta' \in (0, 1)$, $\sigma^2 = 1$ and $|\mu_0 - \mu| \in \{1, 2, \cdots, 5\}$ (Figure 12). In the studied cases, the upper bound on δ is always at least $\sim$3 times δ' and potentially orders of magnitude larger.

Appendix C: Description of models

We provide details on the target distributions used in Section 4. For the Rosenbrock, German Credit, Eight Schools, and Item Response models, precise estimates of the mean and variance along each dimension can be found in the Inference Gym (Sountsov et al., 2020). For the Pharmacokinetics example, we compute benchmark means and variances using 2048 chains, each with 1000 warmup and 1000 sampling iterations. We run MCMC using TensorFlow Probability's implementation of ChEES-HMC (Hoffman et al., 2021). The resulting effective sample size is between 60,000 and 100,000 depending on the parameters. For the bimodal target, the correct mean and variance are worked out analytically.

C.1 Rosenbrock (Dimension = 2)

The Rosenbrock distribution is a nonlinearly transformed normal distribution with highly non-convex level sets; see Equation 1.

C.2 German Credit (Dimension = 25)

"German Credit" is a Bayesian logistic regression model applied to a dataset from a machine learning repository (Dua and Graff, 2017). There are 24 features and an intercept term. The joint distribution over (θ, y) is

$$\begin{aligned} \theta &\sim \text{normal}(0, I), \\ y_n &\sim \text{Bernoulli}\left(\frac{1}{1 + e^{-\theta^T x_n}}\right), \end{aligned}$$

where $I \in \mathbb{R}^{24 \times 24}$ is the identity matrix. Our goal is to sample from the posterior distribution $p(\theta \mid y)$.

C.3 Eight Schools (Dimension = 10)

"Eight Schools" is a Bayesian hierarchical model describing the effect of a program to train students to perform better on a standardized test, as measured by performance across 8 schools (Rubin, 1981). We estimate the group mean and the population mean and variance. To avoid a funnel shaped posterior density, we use a non-centered parameterization:

$$\begin{aligned} \mu &\sim \text{normal}(5, 3) \\ \sigma &\sim \text{normal}^+(0, 10) \\ \eta_n &\sim \text{normal}(0, 1) \\ \theta_n &= \mu + \eta_n \sigma \\ y_n &\sim \text{normal}(\theta_n, \sigma_n), \end{aligned}$$

with the posterior distribution taken over μ and η.

C.4 Pharmacokinetic (Dimension = 45)

"Pharmacokinetics" is a one-compartment model with first-order absorption from the gut that describes the diffusion of a drug compound inside a patient's body. Oral administration of a bolus drug dose induces a discrete change in the drug mass inside the patient's gut. The drug is then absorbed into the *central compartment*, which represents the blood and organs into which the drug diffuses profusely. This diffusion process is described by the system of differential equations:

$$\begin{aligned} \frac{\mathrm{d}m_{\text{gut}}}{\mathrm{d}t} &= -k_1 m_{\text{gut}} \\ \frac{\mathrm{d}m_{\text{cent}}}{\mathrm{d}t} &= k_1 m_{\text{gut}} - k_2 m_{\text{cent}}, \end{aligned}$$

which admits the analytical solution, when $k_1 \neq k_2$,

$$m_{\text{gut}}(t) = m_{\text{gut}}^0 \exp(-k_1 t)$$

$$m_{\text{cent}}(t) \quad = \quad \frac{\exp(-k_2 t)}{k_1 - k_2} \left(m^0_{\text{gut}} k_1 (1 - \exp[(k_2 - k_1)t] + (k_1 - k_2) m^0_{\text{cent}}) \right).$$

Here m^0_{gut} and m^0_{cent} are the initial conditions at time $t = 0$.

A patient typically receives multiple doses. To model this, we solve the differential equations between dosing events, and then update the drug mass in each compartment, essentially resetting the boundary conditions before we resume solving the differential equations. In our example, this means adding m_{dose}, the drug mass administered by each dose, to $m_{\text{gut}}(t)$ at the time of the dosing event. We label the dosing schedule as x.

Each patient receives a total of 3 doses, taken every 12 hours. Measurements are taken at times $t = (0.083, 0.167, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8)$ hours after each dosing event.

We simulate data for 20 patients and for each patient, indexed by n, we estimate the coefficients (k_1^n, k_2^n). We use a hierarchical prior to pool information between patients and estimate the population parameters $(k_1^{\text{pop}}, k_2^{\text{pop}})$ with a non-centered parameterization. The full Bayesian model is:

$$\begin{aligned}
&\textit{hyperpriors:} \\
k_1^{\text{pop}} &\sim \text{lognormal}(\log 1, 0.1) \\
k_2^{\text{pop}} &\sim \text{lognormal}(\log 0.3, 0.1) \\
\sigma_1 &\sim \text{lognormal}(\log 0.15, 0.1) \\
\sigma_2 &\sim \text{lognormal}(\log 0.35, 0.1) \\
\sigma &\sim \text{lognormal}(-1, 1) \\
&\textit{hierarchical priors:} \\
\eta_1^n &\sim \text{normal}(0, 1) \\
\eta_2^n &\sim \text{normal}(0, 1) \\
k_1^n &= k_1^{\text{pop}} \exp(\eta_1^n \sigma_1) \\
k_2^n &= k_2^{\text{pop}} \exp(\eta_2^n \sigma_2) \\
&\textit{likelihood:} \\
y_n &\sim \text{lognormal}(\log m_{\text{cent}}(t, k_1^n, k_2^n, x), \sigma).
\end{aligned}$$

We fit the model on the unconstrained scale, meaning the Markov chains explore the parameter space of, for example, $\log k_1^{\text{pop}} \in \mathbb{R}$, rather than $k_1^{\text{pop}} \in \mathbb{R}^+$.

C.5 Mixture of Gaussians (Dimension = 100)

A mixture of two well-separated 100-dimensional normals,

$$\theta \sim 0.3 \text{ MVN}(-\mu, I) + 0.7 \text{ MVN}(\mu, I),$$

where μ is the 100-dimensional vector of 5's.

C.6 Item Response Theory (Dimension = 501)

The posterior of a model to estimate students abilities when taking an exam. There are $J = 400$ students and $L = 100$ questions. The model parameters are the mean student ability $\delta \in \mathbb{R}$, the ability of each individual student $\boldsymbol{\alpha} \in \mathbb{R}^J$ and the difficulty of each question $\boldsymbol{\beta} \in \mathbb{R}^L$. The observations are the binary matrix $Y \in \mathbb{R}^{J \times L}$, with $y_{j\ell}$ the response of student j to question ℓ. The joint distribution is

$$
\begin{aligned}
\delta &\sim \text{normal}(0.75, 1) \\
\boldsymbol{\alpha} &\sim \text{MVN}(0, I) \\
\boldsymbol{\beta} &\sim \text{MVN}(0, I) \\
y_{j\ell} &\sim \text{Bernoulli}\left[\text{logit}^{-1}(\alpha_j - \beta_\ell + \delta)\right].
\end{aligned}
$$