Assessing the Convergence of Markov chain Monte Carlo when running many short chains

Charles C. Margossian[1], Matthew D. Hoffman[2], Pavel Sountsov[2], Lionel Riou-Durand[3], Aki Vehtari[4] and Andrew Gelman[5]
[1]Center for Computational Mathematics, Flatiron Institute, USA; [2]Google Research, USA; [3]Department of Statistics, University of Warwick, UK;
[4]Department of Computer Science, Aalto University, Finland; [5]Department of Statistics and Political Science, Columbia University, USA.

> *"All human wisdom is contained in these two words: wait and hope."*
> — Alexandre Dumas, *The Count of Monte Cristo*

The goal of this work is to wait less and not rely too much on hope.

Contributions:

- *Nested-*$\widehat{R}$: a generalization of $\widehat{R}$ to diagnose convergence in the many-short-chains regime.
- An asymptotic analysis of $\widehat{R}$ and $\mathfrak{n}\widehat{R}$ for non-stationary chains.

The many-short-chains regime

MCMC runs in two phases:

1. A warmup phase, made of $\mathcal{W}$ iterations, to reduce the bias.
2. A sampling phase, made of N iterations, to reduce the variance.

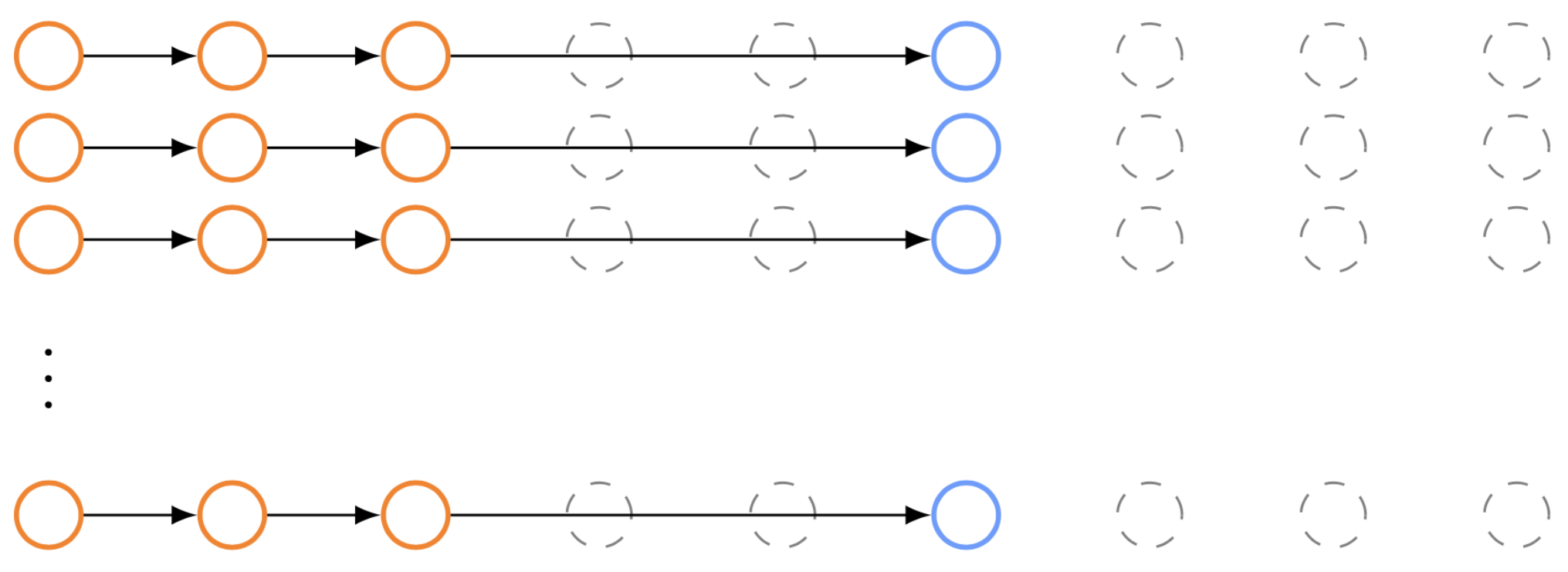


Figure 1. *The many short-chains regime. Running many chains allows us to shorten the* *sampling phase*. *With cross-chain adaptation, we may also reduce the length of the* *warmup phase*.

GPU-friendly samplers:

- Synchronize run times of Markov chains to work well on GPUs.
- Cross-chain adaptation: pool information between chains to tune the sampler (Figure 2).
- Examples: ChEES-HMC [1] and adaptive MALT [6].
- High performance implementation using `TensorFlow Probability` [3].

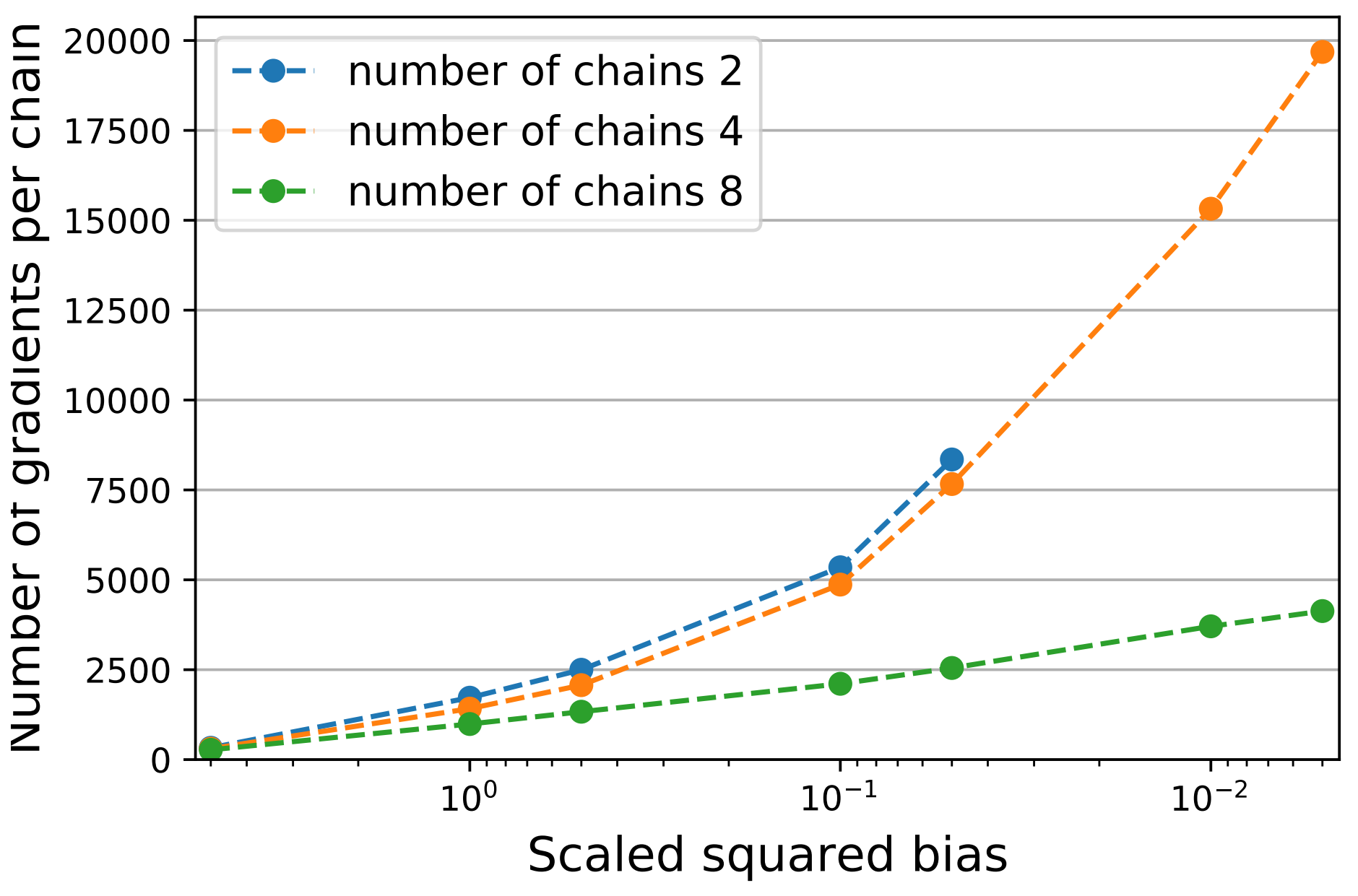


Figure 2. *Using more chains and doing cross-chain adaptation reduces the average number of gradient evaluations per chain required to achieve a target bias. In this example, we target an ill-conditioned Gaussian using ChEES-HMC [1] over 1,000 iterations. With 2 chains, a target bias below* 10^{-2} *cannot be achieved in 1,000 iterations.*

Limitation of $\widehat{R}$ and motivation for $\mathfrak{n}\widehat{R}$

Let $\theta^{(nm)}$ be the nth iteration of the mth chain.

Definition of $\widehat{R}$:

$$\widehat{B} = \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{\theta}^{(\cdot m)} - \bar{\theta}^{(\cdot\cdot)}\right)^2}_{\text{Variance of single chain mean}}; \quad \widehat{W} = \underbrace{\frac{1}{M}\sum_{n=1}^{N}\frac{1}{N-1}\sum_{n=1}^{N}\left(\bar{\theta}^{(nm)} - \bar{\theta}^{(\cdot m)}\right)^2}_{\text{Within-chain variance}}; \quad \widehat{R} = \sqrt{1 + \widehat{B}/\widehat{W}}.$$

- Recommendation: check if $\widehat{R} \leq 1.01$ to assess convergence [8].
- *Even for stationary chains*, we still need long chains for $\widehat{R}$ to decay to 1.01 (Figure 3).

Proposition. Suppose the chains are stationary (i.e. the warmup length $\mathcal{W} \to \infty$). Then, if the chains have a non-negative auto-correlation,

$$\lim_{\mathcal{W}, M \to \infty} \widehat{R} \geq \sqrt{1 + 1/\text{ESS}^{(1)}},$$

where M is the number of chains and $\text{ESS}^{(1)}$ is the effective sample size *per chain*.

Nested-$\widehat{R}$**.**

- Use the variance of the Monte Carlo estimator computed by a group of chains or *superchain*.
- Partition the Markov chains in to K superchains, each made of M subchains initialized at the same point, θ_0^k.
- $\theta^{(nmk)}$ is now the nth iteration in the mth subchain of the kth superchain.

$$\mathfrak{n}\widehat{B} = \underbrace{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{\theta}^{(\cdot\cdot k)} - \bar{\theta}^{(\cdot\cdot\cdot)}\right)^2}_{\text{Variance of superchain mean}}; \quad \underbrace{\mathfrak{n}\widehat{W} = \frac{1}{K}\sum_{k=1}^{K}\widehat{B} + \widehat{W}}_{\text{Within-superchain variance}} \; ; \quad \mathfrak{n}\widehat{R} = \sqrt{1 + \frac{\mathfrak{n}\widehat{B}}{\mathfrak{n}\widehat{W}}}.$$

- For stationary chains, $\mathfrak{n}\widehat{R}$ decays to 1.01 either for large N or large M (Figure 3).
- $\widehat{R}$ is a special case of $\mathfrak{n}\widehat{R}$ where each superchain only has one subchain ($M = 1$).

Analysis of nested-$\widehat{R}$

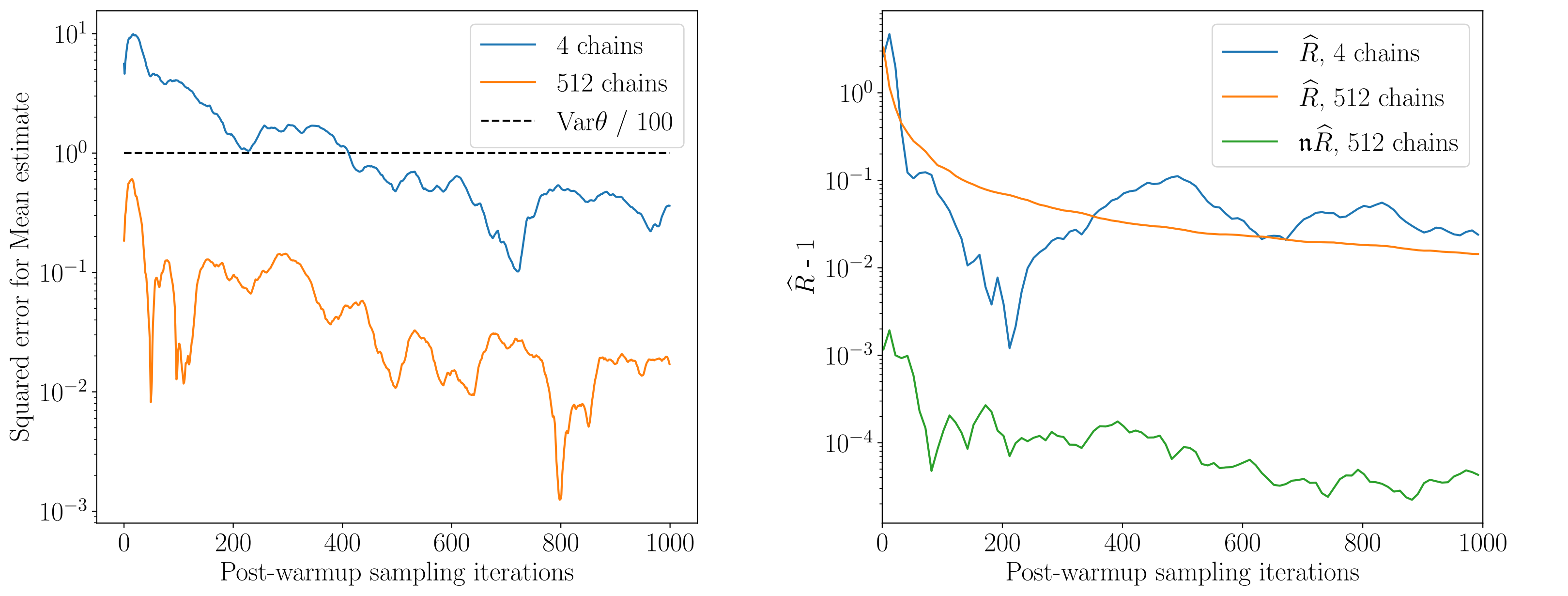


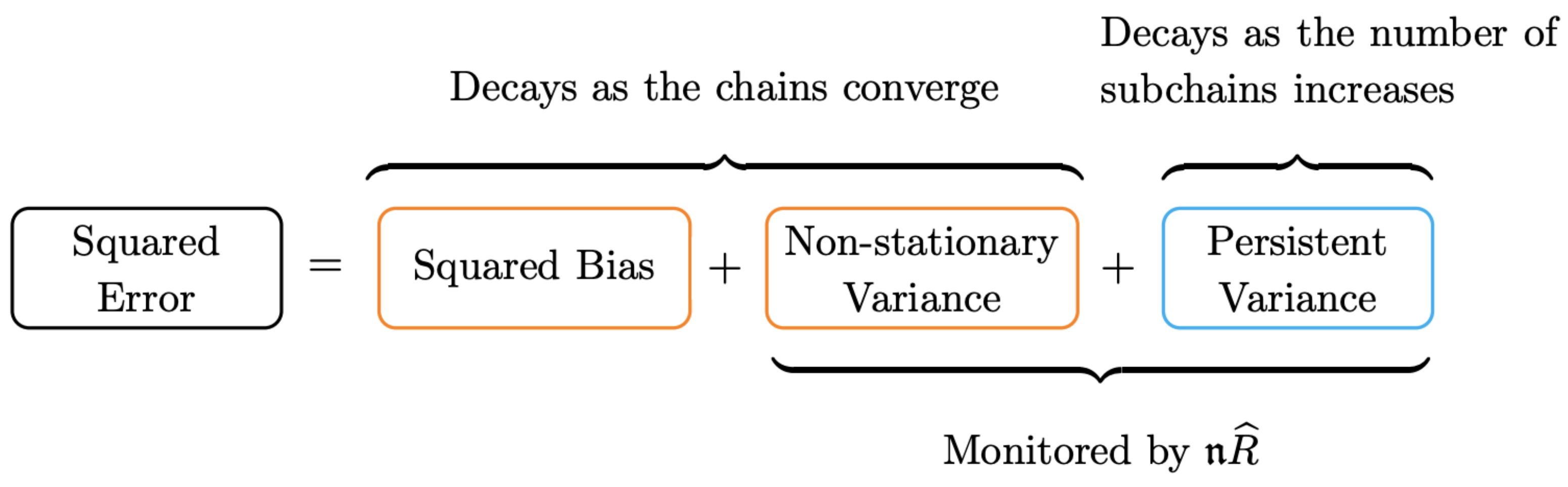
Figure 3. $\widehat{R}$ *and* $\mathfrak{n}\widehat{R}$ *on Banana problem. After warmup, a single iteration suffices to achieve an ESS of 100, when running 512 chains. However, more than 1,000 iterations are required for* $\widehat{R}$ *to decay to 1.01. With sufficient subchains,* $\mathfrak{n}\widehat{R}$ *immediately diagnoses convergence.*

Asymptotic analysis.

- An important question is to understand which quantity $\widehat{R}$ (and $\mathfrak{n}\widehat{R}$) measure [7, 5].
- To analyze non-stationary chains we assume the chains have finite length, but consider the asymptotic limit along the number of superchains, $K \to \infty$.

Theorem. The variance of the superchain Monte Carlo estimator observes the following decomposition,

$$\lim_{K\to\infty} \mathfrak{n}\widehat{B} = \underbrace{\text{Var}\mathbb{E}\left(\bar{\theta}^{(\cdot\cdot k)} \mid \theta_0^k\right)}_{\text{Non-stationary variance}} + \underbrace{\frac{1}{M}\mathbb{E}\text{Var}\left(\bar{\theta}^{(\cdot mk)} \mid \theta_0^k\right)}_{\text{Persistent variance}}.$$



Decays as the chains converge

Decays as the number of subchains increases

Squared Error = Squared Bias + Non-stationary Variance + Persistent Variance

Monitored by $\mathfrak{n}\widehat{R}$

Figure 4. *Expected squared error decomposition for a superchain Monte Carlo estimator.*

- We argue the main utility of $\mathfrak{n}\widehat{R}$ is to measure the **non-stationary variance** which can be used as a "proxy clock" for the **squared bias** (Figure 4).
- In the Langevin diffusion approximation (Gaussian initialization and target), the non-stationary variance and the squared bias both decay at a rate $\propto e^{-2T}$.
- The **persistence variance** can be linked to ESS and is a nuisance term. This nuisance term is $\mathcal{O}(1/M)$ with our nesting strategy.

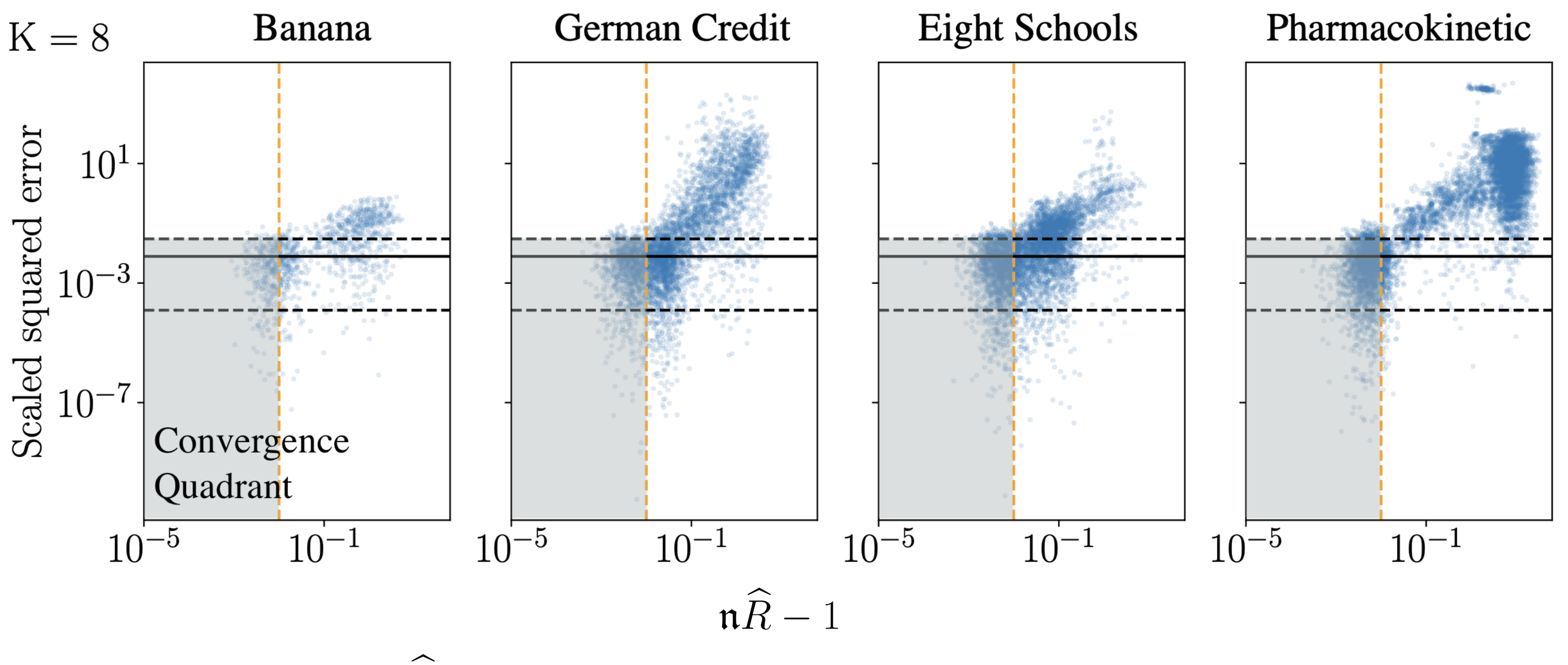


Figure 5. *Application of* $\mathfrak{n}\widehat{R}$ *to a diversity of Bayesian models using* $KM = 128$ *chains, partitioned into* $K = 8$ *superchains. Passed a certain threshold (*$\mathfrak{n}\widehat{R} \leq 1.01$*, yellow vertical line), the squared error behaves as we would expect from stationary chains (0.9 coverage for stationary error distribution given by horizontal dashed lines) and falls in the convergence quadrant.*

Discussion. Our preprint [4] provides:

- Discussion on how to partition the Markov chains (i.e. choice of K).
- Preliminary experiments on early stopping of the warmup using $\mathfrak{n}\widehat{R}$.
- Examples of how our nesting strategy can be applied to other convergence diagnostics, notably for multivariate chains [e.g 7, 5, 2].

References

[1] M. D. Hoffman, A. Radul, and P. Sountsov. An adaptive MCMC scheme for setting trajectory lengths in Hamiltonian Monte Carlo. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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

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

[4] C. C. Margossian, M. D. Hoffman, P. Sountsov, L. Riou-Durand, A. Vehtari, and A. Gelman. Nested $\widehat{R}$: Assessing the convergence of markov chain monte carlo when running many short chains. *arXiv:2110.13017.*, December 2022.

[5] T. Moins, J. Arbel, A. Dutfoy, and S. Girard. On the use of a local $\widehat{R}$ to improve MCMC convergence diagnostic. *arXiv:2205.06694*, 2022.

[6] L. Riou-Durand, P. Sountsov, J. Vogrinc, C. C. Margossian, and S. Power. Adaptive tuning for metropolis adjusted langevin trajectories. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Avril 2023.

[7] D. Vats and D. Knudson. Revisiting the Gelman-Rubin diagnostic. *Statistical Science*, 36:518–529, 2021.

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

BayesComp, March 2023, Levi, Finland; preprint: https://arxiv.org/abs/2110.13017