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

Charles C. Margossian¹, Matthew D. Hoffman², Pavel Sountsov³, Lionel Riou-Durand⁴, Aki Vehtari⁴ and Andrew Gelman⁵
¹Center for Computational Mathematics, Flatiron Institute, USA; ²Google Research, USA; ³Department of Statistics, University of Warwick, UK; ⁴Department of Computer Science, Aalto University, Finland; ⁵Department of Statistics and Political Science, Columbia University, USA.

“The human spirit is made up by two things: the will to wait and not rely too much on 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-\(\hat{R}\): a generalization of \(\hat{R}\) to diagnose convergence in the many-short-chains regime.
• An asymptotic analysis of \(\hat{R}\) and \(n\hat{R}\) for non-stationary chains.

The many-short-chains regime

MCMC runs in two phases:
1. A warmup phase, made of \(W\) iterations, to reduce the bias.
2. A sampling phase, made of \(N\) iterations, to reduce the variance.

GPU-friendly samples:
• 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 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.

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 \(n\hat{R}\)-conditioned Gaussian using ChEES-HMC \([1]\) over 1,000 iterations. With 2 chains, a target bias below \(\hat{R}^{-1}\) cannot be achieved in 1,000 iterations.

Figure 3. \(\hat{R}\) and \(n\hat{R}\) 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 \(\hat{R}\) to decay to 1.01. With sufficient subchains, \(n\hat{R}\) immediately diagnoses convergence.

Asymptotic analysis.
• An important question is to understand which quantity \(\hat{R}\) (and \(n\hat{R}\)) 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,

\[
\hat{\text{Var}}(\theta^\text{MC}) = \hat{\text{Var}}(\theta^\text{single-chain}) + \frac{1}{K}\hat{\text{Var}}(\hat{\theta}^\text{MC} - \bar{\theta}^\text{MC})
\]

Decays as the chains converge
Decays as the number of subchains increases

Monitored by \(n\hat{R}\)

Figure 4. Expected squared error decomposition for a superchain Monte Carlo estimator.
• We argue the main utility of \(n\hat{R}\) 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 \(\propto e^{-t}\).
• The persistence variance can be linked to ESS and is a nuisance variance. This nuisance term is \(O(1/Kd)\) with our nesting strategy.

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 \(n\hat{R}\).
• Examples of how our nesting strategy can be applied to other convergence diagnostics, notably for multivariate chains \([e.g. 6, 7] \).

References
