

# Behaviour of informed ESS estimators

Alexandre Bouchard-Côté

2025-02-04

## Introduction

After reviewing the notion of an *informed ESS estimator* we show some numerical experiments to assess the behaviour of informed ESS estimators in a range of regimes including efficient as well as very inefficient MCMC algorithms.

## Background

Consider a setup where we are benchmarking an MCMC method. To do so, we often pick a test function with known mean  $\mu$  and variance  $\sigma^2$  under the target distribution  $\pi$ . Here we review the construction of an *informed ESS estimator* based on these known parameters.

**Markov chain CLT:** Fix a Markov Kernel and a test function satisfying a central limit theorem for Markov chains, which motivates approximations of the form:

$$\sqrt{k}(\hat{I}_k - \mu) \approx \mathcal{N}(0, \sigma_a^2),$$

where  $\hat{I}_k = \frac{1}{k} \sum_{i=1}^k g(X_i)$  and  $\mu = \mathbb{E}[g(X)]$  for  $X \sim \pi$ , and  $\sigma_a^2$  is the *asymptotic variance*, a constant that depends on  $g$ ,  $\pi$  and the mixing of the Markov chain.

Now from the CLT for Markov chains it follows that if we have a Monte Carlo average  $I_k$  based on a MCMC chain of length  $k$ , then

$$k\text{Var}(\hat{I}_k) \approx \sigma_a^2. \tag{1}$$

**Independent MCMC chains:** Suppose first we had  $a_n$  independent copies of MCMC (we will relax this shortly), each of length  $b_n$ . Let  $\hat{I}^{(1)}, \dots, \hat{I}^{(a_n)}$  denote  $a_n$  independent estimators, the first one based on the first copy, second on second copy, etc. Since the  $I^{(i)}$  are independent and identically distributed,

$$\text{Var}(I^{(1)}) \approx \frac{1}{a_n} \sum_{i=1}^{a_n} (I^{(i)} - \mu)^2. \tag{2}$$

Combining Equation 1 and Equation 2, we obtain:

$$\frac{b_n}{a_n} \sum_{i=1}^{a_n} (I^{(i)} - \mu)^2 \approx \sigma_a^2.$$

**Batch mean trick:** view a trace of length  $n$  as  $a_n$  subsequent batches of length  $b_n$ . A popular choice is  $a_n = b_n = \sqrt{n}$ .

**Effective sample size:** recall the effective sample size (ESS) is defined as  $\text{ESS} = n(\sigma^2/\sigma_a^2)$ . This is the quantity we seek to estimate.

Applying the batch mean trick with  $a_n = b_n = \sqrt{n}$ , we obtain:

$$\text{ESS} = n \frac{\sigma^2}{\sigma_a^2} \approx n \left( \frac{1}{\sigma^2} \frac{b_n}{a_n} \sum_{i=1}^{a_n} (I^{(i)} - \mu)^2 \right)^{-1} = \sqrt{n} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{\sqrt{n}} \left( \frac{I^{(i)} - \mu}{\sigma} \right)^2 \right)^{-1}. \quad (3)$$

The right hand side of this equation is the *informed ESS estimator*.

We can also generalize this to the size of the batch given by  $b_n = n^\theta$  for some parameter  $\theta \in ((1 + \delta/2)^{-1}, 1)$  where we assume  $4 + \delta$  moments for the test function of interest (see Jones et al. 2006, Remark 6). For example, with  $\theta = 1/3$ , assume 8 moments; for  $\theta = 1/2$ , 6 moments. This yields:

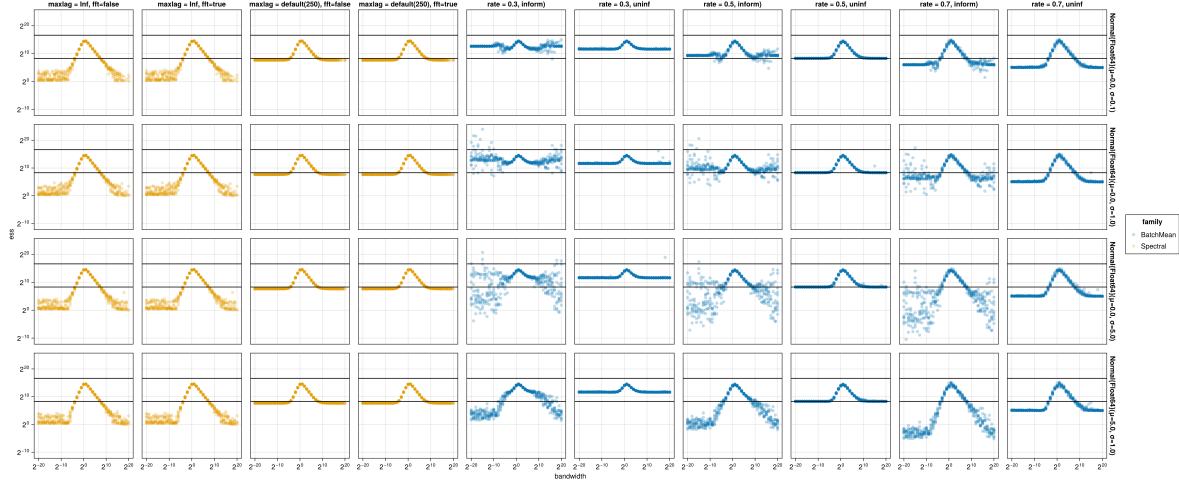
$$\text{ESS} \approx a_n \left( \frac{1}{a_n} \sum_{i=1}^{a_n} \left( \frac{I^{(i)} - \mu}{\sigma} \right)^2 \right)^{-1}.$$

## Numerical investigation

We replicate and expand a numerical experiment reported by Trevor Campbell (personal communication, January 2025). The setup is the following:

- Target distribution is  $\mathcal{N}(0, 1)$  in all experiment.
- We consider Metropolis-Hastings algorithms with normal proposals. We vary the proposal bandwidth (x-axis, log-scale).
- Facet rows: different initial distributions for the MCMC.
- Facet columns: different ESS estimators:
  - *Batch mean ESS (informed)*: the estimator reviewed above.
  - *Batch mean ESS*: the same as Equation 3 but where  $\mu$  and  $\sigma$  are replaced by the sample mean and standard deviation (based on the full trace; this is equivalent to the classical batch mean ESS estimator).
  - *TuringESS basic and bulk*: ESS estimators based on truncated spectral estimation, default algorithms in the Turing.

- For each initial distribution and proposal bandwidth, we ran 10 independent chains and estimate 10 ESS from each chain separately.
- Each chain contains the samples from 100,000 iteration.
- The top solid black line denotes an idealized effective sample size of 100,000, the bottom solid black line denotes the square root of that.
- The test function used here is  $g(x) = x^2$ , so the reference distribution is a  $\chi^2$  with one degree of freedom. The values of  $\mu$  and  $\sigma$  are computed from that distribution.



## Observations:

- As bandwidth goes to zero or infinity, we expect the effective sample size to go to zero.
- The only setup achieving this is the informed ESS with an initialization far from the target.
- However, for the other initial distributions considered, some of the informed ESS estimates are the worst observed, sometimes taking non-sense values higher than the number of MCMC samples, which is not possible in reversible chains.

## Next steps

- Maybe time to move away from batch mean methods
  - Check this [page benchmarking various ESS estimators](#).
    - \* They have a nice benchmark for detecting non-convergence due to multimodality. Covers both well mixing case and multi-chains detecting poor mixing. Batch means is not the winner. Geyer's method comes up better.
  - Pierre Jacob's handbook highlights the slow convergence of batch mean methods (slower than MC).
- Might still be interesting to look into the variance bias square decomposition

- Optimal burn-in
- Detecting when ESS estimation is meaningless.
- Debunking some misinformation (use high dim normal initialized at zero). [1](#) [2](#) [3](#)

Update: started working on informed spectral, then realized it would probably not be able either to detect ESS values much lower than  $\sqrt{n}$  either. Added that lines to plot and indeed all methods seem to fail around there. So maybe a pragmatic solution here is to just have a threshold at  $\sqrt{n}$  and give NA for these.

**TODO:** discuss and package-up a square-root check in our nextflow scripts.

- Still helpful to do jack-knife bias estimation
- Would be interesting to add a multimodal example?

## References

Jones, Galin L, Murali Haran, Brian S Caffo, and Ronald Neath. 2006. “Fixed-Width Output Analysis for Markov Chain Monte Carlo.” *Journal of the American Statistical Association* 101 (476): 1537–47. <https://doi.org/10.1198/016214506000000492>.