Comparing apples to oranges: a universal effective sample size

Dootika Vats
Indian Institute of Technology Kanpur, India

Monte Carlo Methods and Applications, Paris

June 30, 2023

Estimation Problem

Consider an expectation estimation problem, typically found in Bayesian inference. Let π be the density of a distribution on $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$. We are interested in

$$\theta := \int_{\mathcal{X}} f(x)\pi(x)dx < \infty.$$

We will eventually consider multivariate functions as well.

Estimation Problem

Typically π is complex enough that a sampling procedure is used:

1. Vanilla Monte Carlo: $X_1, \ldots, X_m \overset{iid}{\sim} \pi$

$$\hat{\theta}_{\text{VMC}} := \frac{1}{m} \sum_{t=1}^{m} f(X_t) \overset{a.s.}{\to} \theta \qquad \text{as } m \to \infty\,.$$

Assuming $\lambda^2 := \text{Var}_\pi[f(X_1)] < \infty$,

$$\sqrt{m}\left(\hat{\theta}_{\text{VMC}} - \theta\right) \overset{d}{\to} N(0, \lambda^2) \qquad \text{as } m \to \infty.$$

Estimation Problem

Typically π is complex enough that a sampling procedure is used:

1. Vanilla Monte Carlo: $X_1, \ldots, X_m \overset{iid}{\sim} \pi$

$$\hat{\theta}_{\text{VMC}} := \frac{1}{m} \sum_{t=1}^{m} f(X_t) \overset{a.s.}{\to} \theta \qquad \text{as } m \to \infty\,.$$

Assuming $\lambda^2 := \text{Var}_\pi[f(X_1)] < \infty$,

$$\sqrt{m}\left(\hat{\theta}_{\text{VMC}} - \theta\right) \overset{d}{\to} N(0, \lambda^2) \qquad \text{as } m \to \infty.$$

2. Markov chain Monte Carlo
3. Importance Sampling

Markov chain Monte Carlo

Let $\{X_t\}_{t\geq 1}$ be a π-ergodic Markov chain. Then:

$$\hat{\theta}_{\mathrm{MCMC}} := \frac{1}{n}\sum_{t=1}^{n} f(X_t) \overset{a.s.}{\to} \theta \qquad \text{as } n \to \infty\,.$$

Markov chain Monte Carlo

Let $\{X_t\}_{t\geq 1}$ be a π-ergodic Markov chain. Then:

$$\hat{\theta}_{\text{MCMC}} := \frac{1}{n}\sum_{t=1}^{n} f(X_t) \overset{a.s.}{\to} \theta \qquad \text{as } n \to \infty\,.$$

Further, if a Markov chain CLT holds, then as $n \to \infty$

$$\sqrt{n}\left(\hat{\theta}_{\text{MCMC}} - \theta\right) \overset{d}{\to} N(0, \sigma^2)\,,$$

where

$$\sigma^2 = \lambda^2 + 2\sum_{k=1}^{\infty} \text{Cov}_\pi(f(X_1), f(X_{1+k}))\,.$$

Effective sample size (ESS) in MCMC

If MCMC samples of size n

$$\mathsf{Var}_\pi\left(\hat{\theta}_{\mathsf{MCMC}}\right) \approx \frac{\sigma^2}{n}$$

Effective sample size (ESS) in MCMC

If MCMC samples of size n

$$\mathrm{Var}_{\pi}\left(\hat{\theta}_{\mathrm{MCMC}}\right) \approx \frac{\sigma^2}{n}$$

If iid samples of size m

$$\mathrm{Var}_{\pi}\left(\hat{\theta}_{\mathrm{VMC}}\right) = \frac{\lambda^2}{m}$$

Effective sample size (ESS) in MCMC

If MCMC samples of size n

$$\mathrm{Var}_\pi\left(\hat{\theta}_{\mathrm{MCMC}}\right) \approx \frac{\sigma^2}{n}$$

If iid samples of size m

$$\mathrm{Var}_\pi\left(\hat{\theta}_{\mathrm{VMC}}\right) = \frac{\lambda^2}{m}$$

Question: Can we compare the MCMC estimation quality to the estimation quality from iid samples?

ESS in MCMC

Answer: For what m, is $\text{Var}_\pi\left(\hat{\theta}_{\text{MCMC}}\right) = \text{Var}_\pi\left(\hat{\theta}_{\text{VMC}}\right)$?
That m, is the effective sample size (ESS).

$$m = ESS := n\frac{\lambda^2}{\sigma^2}\,.$$

ESS in MCMC

Answer: For what m, is $\text{Var}_\pi\left(\hat{\theta}_{\text{MCMC}}\right) = \text{Var}_\pi\left(\hat{\theta}_{\text{VMC}}\right)$?
That m, is the effective sample size (ESS).

$$m = ESS := n\frac{\lambda^2}{\sigma^2}\,.$$

Interpretation: In order to estimate θ, this MCMC sample is equivalent to ESS amount of iid samples from π.

ESS in MCMC

Answer: For what m, is $\mathrm{Var}_\pi\left(\hat{\theta}_{\mathrm{MCMC}}\right) = \mathrm{Var}_\pi\left(\hat{\theta}_{\mathrm{VMC}}\right)$?
That m, is the effective sample size (ESS).

$$m = ESS := n\frac{\lambda^2}{\sigma^2}\,.$$

Interpretation: In order to estimate θ, this MCMC sample is equivalent to ESS amount of iid samples from π.

Estimating λ^2 and the more complicated σ^2 well is challenging and important. See Flegal and Jones (2010).

Multivariate ESS in MCMC

Following the same principles for $f : \mathcal{X} \to \mathbb{R}^p$:

$$\Lambda = \mathrm{Var}_\pi[f(X_1)]$$

$$\Sigma = \Lambda + \sum_{k=1}^{\infty} \left(\mathrm{Cov}_\pi(f(X_1), f(X_{k+1})) + \mathrm{Cov}_\pi(f(X_1), f(X_{k+1}))^T \right)$$

then, Vats et al. (2019) define a multivariate ESS:

$$ESS = n \left(\frac{\det(\Lambda)}{\det(\Sigma)} \right)^{1/p} .$$

Importance Sampling

Let $X_1, \dots, X_n \stackrel{iid}{\sim} q$, where q is an importance density.

Define weights

$$w(X_t) = \frac{\pi(X_t)}{q(X_t)}\,.$$

The *self-normalized importance sampling* estimator of θ is

Importance Sampling

Let $X_1, \ldots, X_n \overset{iid}{\sim} q$, where q is an importance density.

Define weights

$$w(X_t) = \frac{\pi(X_t)}{q(X_t)}\,.$$

The *self-normalized importance sampling* estimator of θ is

$$\hat{\theta}_{SNIS} := \frac{\sum_{t=1}^{n} f(X_t) w(X_t)}{\sum_{t=1}^{n} w(X_t)} \overset{a.s.}{\to} \theta \qquad \text{as } n \to \infty.$$

Importance Sampling

Let $X_1, \ldots, X_n \stackrel{iid}{\sim} q$, where q is an importance density.

Define weights

$$w(X_t) = \frac{\pi(X_t)}{q(X_t)} \,.$$

The *self-normalized importance sampling* estimator of θ is

$$\hat{\theta}_{SNIS} := \frac{\sum_{t=1}^{n} f(X_t) w(X_t)}{\sum_{t=1}^{n} w(X_t)} \stackrel{a.s.}{\to} \theta \qquad \text{as } n \to \infty.$$

Quality of estimation depends critically on q and thus the weights w.

Importance Sampling Variance

Assume

$$\tau^2 := \lim_{n\to\infty} n\mathrm{Var}_q(\hat{\theta}_{SNIS}) = \frac{\mathrm{E}_q(w(X_1)^2(f(X_1)-\theta)^2)}{\mathrm{E}_q(w(X_1))} < \infty$$

then asymptotic normality of the SNIS estimator holds:

$$\sqrt{n}(\hat{\theta}_{SNIS} - \theta) \xrightarrow{d} N(0, \tau^2)$$

τ^2 can be estimated using weighted samples from q. For the purposes of this talk, we will not discuss estimation.

Kong's ESS in Importance Sampling

A popular measure of the quality of importance sampling procedure is the *ESS* of Kong (1992). Let

$$\tilde{w}(X_t) = \frac{w(X_t)}{\sum_{i=1}^{n} w(X_i)}$$

$$ESS = \frac{1}{\sum_{t=1}^{n} \tilde{w}(X_t)^2}$$

Kong's ESS in Importance Sampling

A popular measure of the quality of importance sampling procedure is the *ESS* of Kong (1992). Let

$$\tilde{w}(X_t) = \frac{w(X_t)}{\sum_{i=1}^{n} w(X_i)}$$

$$ESS = \frac{1}{\sum_{t=1}^{n} \tilde{w}(X_t)^2}$$

- useful to assess quality of weights

Kong's ESS in Importance Sampling

A popular measure of the quality of importance sampling procedure is the *ESS* of Kong (1992). Let

$$\tilde{w}(X_t) = \frac{w(X_t)}{\sum_{i=1}^{n} w(X_i)}$$

$$ESS = \frac{1}{\sum_{t=1}^{n} \tilde{w}(X_t)^2}$$

- useful to assess quality of weights
- not interpretable as an effective sample size
- no dependence on f

A closer look

A closer look at Kong (1992) reveals the evolution of how this ESS came about. Elvira et al. (2018) study this in good detail.

The original definition of ESS in Kong (1992) is

$$ESS = n \frac{\mathrm{Var}_{\pi}(\hat{\theta}_{VMC})}{\mathrm{Var}_{q}(\hat{\theta}_{SNIS})}$$

A closer look

A closer look at Kong (1992) reveals the evolution of how this ESS came about. Elvira et al. (2018) study this in good detail.

The original definition of ESS in Kong (1992) is

$$ESS = n\frac{\mathrm{Var}_{\pi}(\hat{\theta}_{VMC})}{\mathrm{Var}_{q}(\hat{\theta}_{SNIS})}$$

Through a series of approximations, Kong (1992) arrives at the popular approximation of the ESS used today.

This first definition is similar to ESS in MCMC.

A modification of ESS

In Agarwal et al. (2022), we make a slight modification in the definition of ESS

$$ESS = n \frac{n \mathrm{Var}_{\pi}(\hat{\theta}_{VMC})}{\lim_{n \to \infty} n \mathrm{Var}_{q}(\hat{\theta}_{SNIS})} = n \frac{\lambda^2}{\tau^2} \, .$$

This definition allows for:

- A clear interpretation of ESS as *effective sample size*
- dependency on f (as it should)
- a stopping criterion based on $ESS \geq$ pre-determined lower bound.

Universal ESS: a roadmap

$$\text{Default:}\sqrt{n}\left(\hat{\theta}_{\text{VMC}} - \theta\right) \xrightarrow{d} N(0, \lambda^2)$$

Universal ESS: a roadmap

$$\text{Default:}\sqrt{n}\left(\hat{\theta}_{\text{VMC}}-\theta\right)\xrightarrow{d}N(0,\lambda^2)$$

MCMC	SNIS
$\sqrt{n}\left(\hat{\theta}_{\text{MCMC}}-\theta\right)\xrightarrow{d}N(0,\sigma^2)$	$\sqrt{n}\left(\hat{\theta}_{SNIS}-\theta\right)\xrightarrow{d}N(0,\tau^2)$
$ESS = n\dfrac{\lambda^2}{\sigma^2}$	$ESS = n\dfrac{\lambda^2}{\tau^2}$
$ESS = n\left(\dfrac{\det(\Lambda)}{\det(\Sigma)}\right)^{1/p}$	$ESS = n\left(\dfrac{\det(\Lambda)}{\det(T)}\right)^{1/p}$

where T is the asymptotic covariance matrix for $\hat{\theta}_{SNIS}$.

Universal ESS: a roadmap

$$\text{Default:}\sqrt{n}\left(\hat{\theta}_{\mathrm{VMC}} - \theta\right) \xrightarrow{d} N(0, \lambda^2)$$

MCMC

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{MCMC}} - \theta\right) \xrightarrow{d} N(0, \sigma^2)$$

$$ESS = n\frac{\lambda^2}{\sigma^2}$$

$$ESS = n\left(\frac{\det(\Lambda)}{\det(\Sigma)}\right)^{1/p}$$

SNIS

$$\sqrt{n}\left(\hat{\theta}_{SNIS} - \theta\right) \xrightarrow{d} N(0, \tau^2)$$

$$ESS = n\frac{\lambda^2}{\tau^2}$$

$$ESS = n\left(\frac{\det(\Lambda)}{\det(T)}\right)^{1/p}$$

where T is the asymptotic covariance matrix for $\hat{\theta}_{SNIS}$.

This recipe may be followed generally!

Stopping rules for simulations

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N_p(0, \Upsilon)\,, \qquad \text{and} \qquad \hat{\Upsilon}$$

stop simulation when:

Stopping rules for simulations

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N_p(0, \Upsilon), \qquad \text{and} \qquad \hat{\Upsilon}$$

stop simulation when:

$$\text{Volume of Confidence Ellipsoid} + \text{vanishing } n \text{ term} \leq \epsilon \det(\hat{\Lambda})^{1/2}$$

Stopping rules for simulations

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N_p(0, \Upsilon)\,, \qquad \text{and} \qquad \hat{\Upsilon}$$

stop simulation when:

$$\text{Volume of Confidence Ellipsoid} + \text{vanishing } n \text{ term} \leq \epsilon \det(\hat{\Lambda})^{1/2}$$

Theorem

Let estimators $\hat{\Upsilon}$ and $\hat{\Lambda}$ be strongly consistent. Let ϵ be a desired tolerance level for quality of estimation, and α be a required confidence level. Then stopping at the random time T^* when

$$\widehat{ESS} := T^* \left(\frac{det(\hat{\Lambda})}{det(\hat{\Upsilon})}\right)^{1/p} \geq \frac{2^{2/p}\pi\chi^2_{1-\alpha,p}}{(p\Gamma(p/2))^{2/p}} \frac{1}{\epsilon^2}$$

yields asymptotically valid confidence region for $\hat{\theta}$ as $\epsilon \to 0$.

Stopping rules for simulations

Stop simulation when

$$\widehat{\text{ESS}} \geq \frac{2^{2/p}\pi\chi^2_{1-\alpha,p}}{(p\Gamma(p/2))^{2/p}} \frac{1}{\epsilon^2}$$

- ϵ: is chosen by user
- α: is chosen by user – default .95
- p: dimension of estimation
- lower bound available before simulation begins

Example: Gaussian target

Consider SNIS estimator of mean of

$$\pi = N\left(0, \quad \Lambda = \begin{pmatrix} 2 & .5\sqrt{2} \\ .5\sqrt{2} & 1 \end{pmatrix}\right) \qquad \text{with}$$

$$q = N\left(0, \quad \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$$

Example: Gaussian target

Consider SNIS estimator of mean of

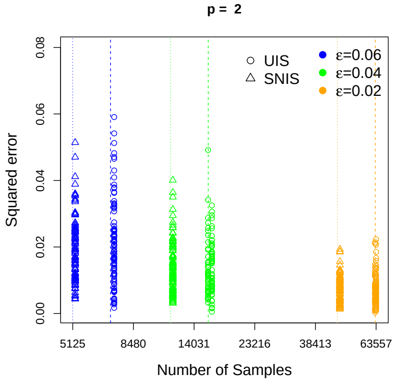
$$\pi = N\left(0, \quad \Lambda = \begin{pmatrix} 2 & .5\sqrt{2} \\ .5\sqrt{2} & 1 \end{pmatrix}\right) \qquad \text{with}$$

$$q = N\left(0, \quad \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$$

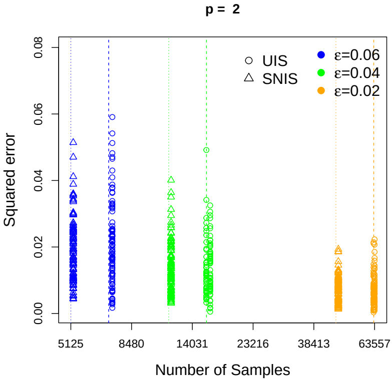
In 100 repetitions

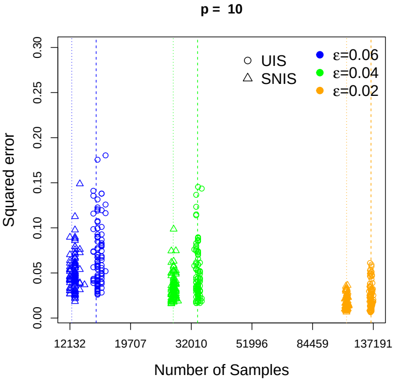
- We set $\epsilon = .02, .04, .06$
- For each ϵ, determine when simulation stops according to ESS criterion
- Plot $\|\hat{\theta} - \theta\|^2$ vs Monte Carlo sample size.

Example: Stopping rule



Example: Stopping rule





Example: Apples to oranges

Consider estimating mean of

$$\pi = N\left(0, \quad \Lambda = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}\right)$$

We will visualize the asymptotic covariance from

- Vanilla Monte Carlo (Λ)
- SNIS with

$$q = N\left(0, \quad \begin{pmatrix} 1.2 & .5 \\ .5 & 1.2 \end{pmatrix}\right)$$

- Gibbs sampler

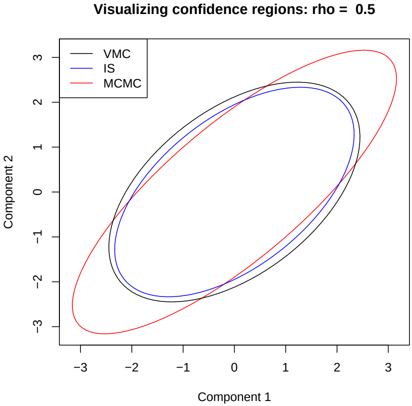
The form of Λ, T, and Σ are known in closed form

Example: Apples to oranges

The relative volumes of the confidence regions is essentially the ESS.

Example: Apples to oranges

The relative volumes of the confidence regions is essentially the ESS.



$$\mathrm{ESS}_{\mathrm{MCMC}} = .436n \qquad \text{and} \qquad \mathrm{ESS}_{\mathrm{SNIS}} = 1.002n$$

Conclusion

- We re-define ESS in importance sampling for improved interpretability
- General framework for comparing different kinds of estimators
- Of course, now one can also do ESS/time
- More interesting examples in the paper

Conclusion

- We re-define ESS in importance sampling for improved interpretability
- General framework for comparing different kinds of estimators
- Of course, now one can also do ESS/time
- More interesting examples in the paper

Paper: Agarwal, M., Vats, D., and Elvira, V. (2021). A principled stopping rule for importance sampling, *Electronic Journal of Statistics*, 2022

Thank you

Reference I

Agarwal, M., Vats, D., and Elvira, V. (2022). A principled stopping rule for importance sampling. *Electronic Journal of Statistics*, 16:5570–5590.

Elvira, V., Martino, L., and Robert, C. P. (2018). Rethinking the effective sample size. *International Statistical Review*.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38:1034–1070.

Kong, A. (1992). A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337.