Accelerating MCMC with an approximation

Importance sampling versus delayed acceptance

Jouni Helske, Matti Vihola, and Jordan Franks

University of Jyväskylä, Department of Mathematics and Statistics

jouni.helske@iki.fi, mvihola@iki.fi, jordan.j.franks@jyu.fi

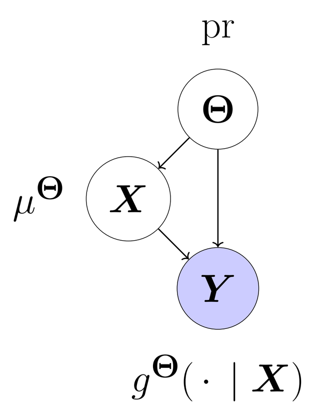
UNIVERSITY OF JYVÄSKYLÄ

Introduction

We consider an importance sampling (IS) type estimator based on Markov chain Monte Carlo (MCMC) which targets an approximate marginal distribution. The IS approach provides a natural alternative to delayed acceptance (DA) pseudo-marginal MCMC, and enjoys many benefits against DA, including a straightforward parallelisation and additional flexibility in MCMC implementation. We compare the computational efficiency of IS and DA approaches in a geometric Brownian motion setting where the IS approach provides substantial efficiency improvements over DA.

Bayesian latent variable models

Variable	Name	Conditional density
Θ	(hyper)parameters	$\Theta \sim \mathrm{pr}(\,\cdot\,)$
$\boldsymbol{X}$	latent variables	$\boldsymbol{X} \mid \Theta \sim \mu^{\Theta}(\,\cdot\,)$
$\boldsymbol{Y}$	observations	$\boldsymbol{Y} \mid (\boldsymbol{X}, \Theta) \sim g^{\Theta}(\,\cdot \mid \boldsymbol{X})$



We are interested in the full posterior with observed $\boldsymbol{Y} = \boldsymbol{y}$:

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = \mathrm{pr}(\boldsymbol{\theta})\mu^{\boldsymbol{\theta}}(\boldsymbol{x})g^{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}).$$

Typical scenario in a latent variable model:

- The hyperparameters Θ are low-dimensional
- The latent variables $\boldsymbol{X}$ are high-dimensional

Approaches and challenges for inference

Standard 'out-of-the-box' inference (e.g. using BUGS, Stan, ...)

- Approach: Simulate MCMC chain $\boldsymbol{Z}_k = (\Theta_k, \boldsymbol{X}_k)$, targeting π
- Problem: High overall dimension & high correlations $\Longrightarrow$ often **inefficient**, even useless

Factorization of the posterior

- Approach: Consider the following factorization of the posterior:

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}) = \pi_m(\boldsymbol{\theta})r(\boldsymbol{x} \mid \boldsymbol{\theta}),$$

where the marginal posterior density and the corresponding conditional are given as

$$\pi_m(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta}, \boldsymbol{x})\mathrm{d}\boldsymbol{x} \propto \mathrm{pr}(\boldsymbol{\theta})L(\boldsymbol{\theta})$$

$$r(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{p^{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})}{L(\boldsymbol{\theta})} = \frac{\mu^{\boldsymbol{\theta}}(\boldsymbol{x})g^{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x})}{\int p^{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{x}}$$

- Problem: $L(\boldsymbol{\theta})$ and $r(\boldsymbol{x} \mid \boldsymbol{\theta})$ are often intractable.
- Possible solutions: approximate with $\hat{L}(\boldsymbol{\theta})$ and $\hat{r}(\boldsymbol{x} \mid \boldsymbol{\theta})$ obtained either
  - deterministically, e.g. Gaussian approximation [3, 7], EKF, INLA [6], variational Bayes
    $\Longrightarrow$ **Bias** is hopefully negligible, but hard to assess in practice
  - stochastically, e.g. sequential Monte Carlo (SMC) [1]
    $\Longrightarrow$ Provides unbiased estimates of $L(\boldsymbol{\theta})$ and $r(\boldsymbol{x} \mid \boldsymbol{\theta})$ but often **computationally demanding**
  - SMC with m particles generates $(U, V^{(i)}, \boldsymbol{X}^{(i)})$ satisfying

$$\mathbf{E}[U] = L(\boldsymbol{\theta}), \quad \mathbf{E}\big[U \sum_{i=1}^{m} V^{(i)} f(\boldsymbol{X}^{(i)})\big] = \int p^{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

Two-stage exact MCMC algorithms for faster inference

Delayed acceptance (DA): Combine particle MCMC [1] with delayed acceptance [2]

- Draw a proposal $\tilde{\Theta}_k \sim q(\Theta_{k-1}, \cdot\,)$
- Stage 1: With probability

$$\min\left\{1, \frac{\mathrm{pr}(\tilde{\Theta}_k)\hat{L}(\tilde{\Theta}_k)q(\tilde{\Theta}_k, \Theta_{k-1})}{\mathrm{pr}(\Theta_{k-1})\hat{L}(\Theta_{k-1})q(\Theta_{k-1}, \tilde{\Theta}_k)}\right\}$$

  continue to the next step, otherwise reject.
- Stage 2: Generate unbiased estimate of $L(\tilde{\Theta}_k)$ and $(\tilde{U}_k, \tilde{V}_k^{(i)}, \tilde{\boldsymbol{X}}_k^{(i)})$ using a particle filter. Set $\tilde{W}_k := \tilde{U}_k / \hat{L}(\tilde{\Theta}_k)$ and with probability $\min\{1, \frac{\tilde{W}_k}{W_{k-1}}\}$ accept, otherwise reject.

Then, form the DA estimator:

$$E_n^{\mathrm{DA}} := \frac{\sum_{k=1}^{n}\sum_{i=1}^{m} V_k^{(i)} f(\boldsymbol{\Theta_k}, \boldsymbol{X}_k^{(i)})}{n}$$

Importance sampling type correction (IS): Correction [9] of an approximate marginal (particle) MCMC targeting $\pi_a(\boldsymbol{\theta}) \propto \mathrm{pr}(\boldsymbol{\theta})\hat{L}(\boldsymbol{\theta})$

- Draw a new proposal $\tilde{\Theta}_k \sim q(\Theta_{k-1}, \cdot\,)$
- Stage 1: With probability

$$\min\left\{1, \frac{\mathrm{pr}(\tilde{\Theta}_k)\hat{L}(\tilde{\Theta}_k)q(\tilde{\Theta}_k, \Theta_{k-1})}{\mathrm{pr}(\Theta_{k-1})\hat{L}(\Theta_{k-1})q(\Theta_{k-1}, \tilde{\Theta}_k)}\right\}$$

  accept $\Theta_k := \tilde{\Theta}_k$, otherwise reject.
- Stage 2: Generate $(U_k, V_k^{(i)}, \boldsymbol{X}_k^{(i)})$ as above. Set $W_k := U_k / \hat{L}(\Theta_k)$.

Then, form the IS estimator

$$E_n^{\mathrm{IS}} := \frac{\sum_{k=1}^{n} W_k \sum_{i=1}^{m} V_k^{(i)} f(\Theta_k, \boldsymbol{X}_k^{(i)})}{\sum_{j=1}^{n} W_j}.$$

Why IS might be better than DA?

- Stage 2 corrections entirely independent
  $\Longrightarrow$ parallelisable $\Longrightarrow$ scalable
- Allows for calculating the correction only for accepted states ('jump chain')
  $\Longrightarrow$ less expensive than DA
- Correction only for subsampled chain
  $\Longrightarrow$ statistically efficient *thinning*
- The approximate marginal MCMC (Θ_k) need not rely on estimators
  $\Longrightarrow$ safer & easier to implement efficiently (e.g. adaptive MCMC...)
- The MCMC (Θ_k) need not be reversible
  $\Longrightarrow$ non-reversible samplers applicable
- Non-negativity of the estimator W_k not required
  $\Longrightarrow$ Allows for 'debiasing' tricks (or 'randomized multi-level Monte Carlo') [5, 8]

Consistency & CLT

Let π_a be an approximation of $\pi_m \ll \pi_a$, $\quad w_u(\boldsymbol{\theta}) = c_w \frac{\pi_m(\boldsymbol{\theta})}{\pi_a(\boldsymbol{\theta})}$, $\quad c_w > 0$, $\quad \xi_k(f) = \sum_{i=1}^{m} V_k^{(i)} f(\Theta_k, \boldsymbol{X}_k^{(i)})$

With mild assumptions [9]:

- Consistency:

$$E_n^{\mathrm{IS}} = \frac{\sum_{k=1}^{n} W_k \xi_k}{\sum_{j=1}^{n} W_j} \xrightarrow[\text{a.s.}]{n\to\infty} \pi(f) = \int f(\boldsymbol{\theta}, \boldsymbol{x})\pi(\boldsymbol{\theta}, \boldsymbol{x})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{x}$$

- CLT:

$$\sqrt{n}\big[E_n - \pi(f)\big] \xrightarrow[\text{d}]{n\to\infty} N\Bigg(0, \overbrace{\frac{\mathrm{Var}(w_u \bar{f}^*, P)}{c_w^2}}^{\text{MCMC}} + \overbrace{\frac{\pi_a(v)}{c_w^2}}^{\text{IS corr}}\Bigg),$$

where $v(\boldsymbol{\theta}) = \mathrm{Var}\big(W_k \xi_k(\bar{f}) \mid \Theta_k = \boldsymbol{\theta}\big)$, $\bar{f}(\boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{\theta}, \boldsymbol{x}) - \pi(f)$, and $\bar{f}^*(\boldsymbol{\theta}, \boldsymbol{x}) = \int \bar{f}(\boldsymbol{\theta}, \boldsymbol{x}')r(\boldsymbol{x}' \mid \boldsymbol{\theta})\mathrm{d}\boldsymbol{x}'$

- Theoretical results [4] in terms of asymptotic variance:
  - If $W_k \le C$ for all $k \ge 1$ a.s., then

$$\mathrm{Var(IS)} \le c_w^{-1}\big[\,C\,\mathrm{Var(DA)} + \bar{\pi}(\xi^2[C - w])\big]$$

  - If $w_u(\boldsymbol{\theta_k}) \le C$ for all $k \ge 1$ a.s., then

$$\mathrm{Var(IS)} \le c_w^{-1}\big[\,C\,\mathrm{Var(DA)} + \bar{\pi}(\xi^2[C + w])\big]$$

  where $\bar{\pi}$ is the stationary probability of the DA chain.

Example: Geometric Brownian motion

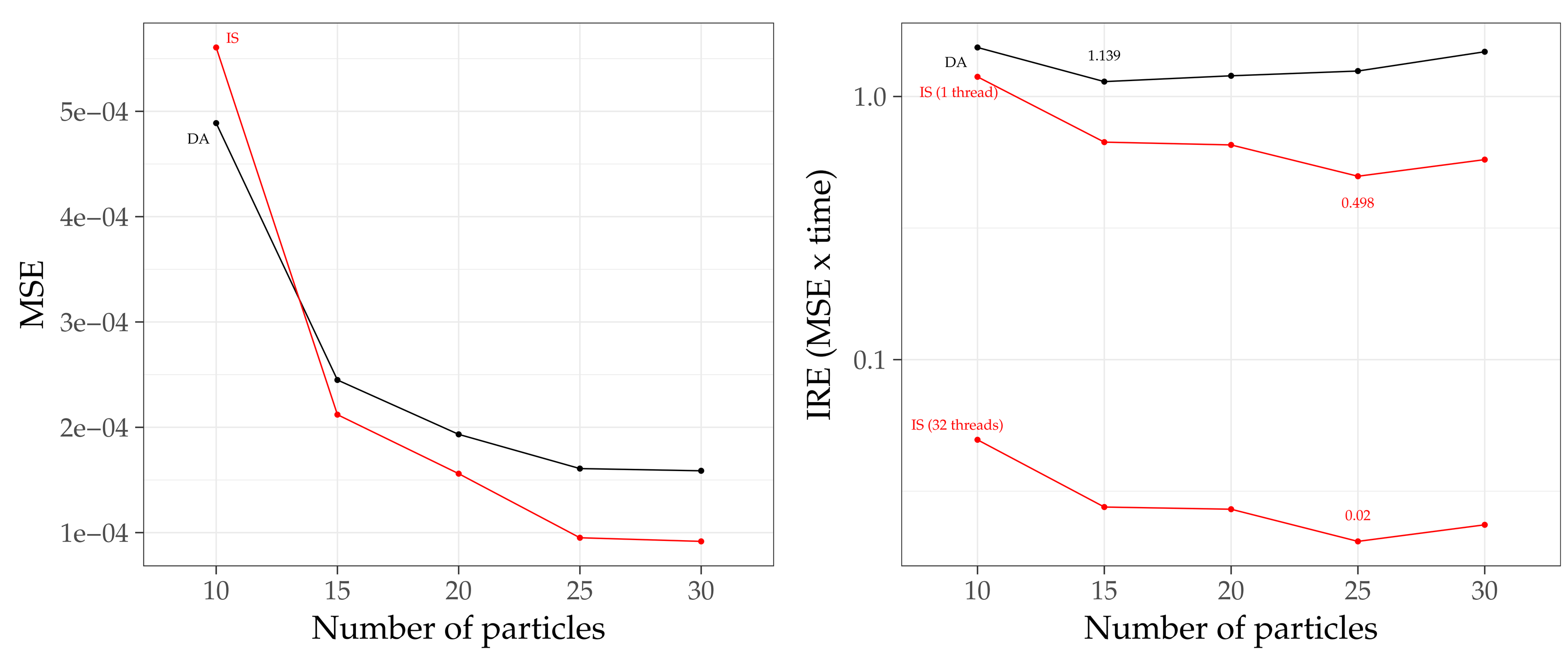
- State process is a geometric Brownian motion:

$$\mathrm{d}X_t = \nu X_t \mathrm{d}t + \sigma_x X_t \mathrm{d}B_t, \qquad X_0 \equiv 1,$$

  where $(B_t)_{t\ge1}$ is a standard Brownian motion.
- Conditionally independent observations $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(T)})$ at integer times:

$$g^{\boldsymbol{\Theta}}(y_t \mid X_t = x_t) = N(\log(x_t), \sigma_y^2).$$

- Here we consider SMC based on a discretisation with Milstein scheme using uniform meshes of size $2^{L_C} = 2^2$ and $2^{L_F} = 2^{12}$ for $\mu^{\boldsymbol{\Theta}}(x_t \mid x_{t-1})$.
  - The approximation is based on the coarse level L_C, and we assume that the fine level L_F provides sufficiently accurate results for practical purposes.
  - This differs from examples in [9] where we used deterministic approximations.
- In our experiment, we simulated one realization using $\boldsymbol{\theta} = (\nu, \sigma_x, \sigma_y) = (0.05, 0.2, 1)$, and $T = 50$.
- We compare the mean square error (MSE) and the inverse efficiency (IRE), defined as the MSE multiplied by the average computation time from 50 independent MCMC runs with 75,000 MCMC iterations with first 25,000 discarded as burn-in. Both MSE and IRE are averaged over the parameters $(\nu, \sigma_x, \sigma_y, x_1, x_2, \ldots, x_T)$.



Average MSE and IRE for varying number of particles in bootstrap particle filter. DA is shown in black, jump chain IS in red.

References

[1] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010. (with discussion).

[2] J. Andrés Christen and Colin Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4), 2005.

[3] James Durbin and Siem Jan Koopman. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684, 1997. doi: 10.1093/biomet/84.3.669.

[4] Jordan Franks and Matti Vihola. Importance sampling and delayed acceptance via a Peskun type ordering. Preprint arXiv:1706.09873, 2017. **Theoretical comparison of IS and DA in terms of asymptotic variances**.

[5] Chang-Han Rhee and Peter Glynn. Unbiased estimation with square root convergence for SDE models. *Oper. Res.*, 63(5):1026–1043, 2015.

[6] Havard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2): 319–392, 2009.

[7] Neil Shephard and Michael K. Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667, 1997. ISSN 00063444.

[8] Matti Vihola. Unbiased estimators and multilevel Monte Carlo. Preprint arXiv:1512.01022v1, 2015.

[9] Matti Vihola, Jouni Helske, and Jordan Franks. Importance sampling type correction of Markov chain Monte Carlo and exact approximations. Preprint arXiv:1609.02541, 2016. **Review, consistency and CLT, with more illustrations with Poisson and stochastic volatility models (where the approximate chain does not rely on SMC)**.

Acknowledgments

This project has been supported by Academy of Finland research grants 274740 and 284513.