Speeding Up MCMC by Efficient Data Subsampling

Matias Quiroz, Linköping University and Sveriges Riksbank, Sweden
Jointly with Mattias Villani, Robert Kohn and Minh-Ngoc Tran

Summary

We propose a Pseudo-Marginal Metropolis-Hastings (PMMH) framework where the likelihood function for n observations is estimated from a random subset of m observations. The key features of our approach are (i) efficient control variates for variance reduction of the log-likelihood estimator, (ii) variance reduction of the estimated Metropolis-Hastings ratio via a correlated pseudo-marginal approach and (iii) an accurate approximate bias-correction that theoretically guarantees that the error is small.

PMMH for data subsampling

Let $\hat{p}_{m,n}(y|\theta,u)$ denote an estimate, based on m observations selected by the random vector u, of the likelihood

$$p(y|\theta) = \prod_{k=1}^{n} p(y_k|\theta) = \exp(l(\theta)) \tag{1}$$

where $l(\theta) = \sum l_k(\theta)$ and $l_k(\theta) = \log p(y_k|\theta)$.

Standard PMMH

The pseudo-marginal approach (Andrieu and Roberts, 2009) generates a Markov chain on the augmented space (θ, u) by proposing to move

$$(\theta_c, u_c) \to (\theta_p, u_p) \text{ from the joint proposal } q(\theta|\cdot)p(u)$$

and accept with probability

$$\alpha_{\text{PMMH}} = \min\left(1, \frac{\hat{p}_{m,n}(y|\theta_p,u_p)p(\theta_p)/q(\theta_p|\theta_c)}{\hat{p}_{m,n}(y|\theta_c,u_c)p(\theta_c)/q(\theta_c|\theta_p)}\right). \tag{2}$$

The algorithm targets $\tilde{\pi}(\theta,u) \propto \hat{p}_{m,n}(y|\theta,u)p(\theta)p(u)$ and, if $\hat{p}_{m,n}$ is (a.s.) positive and unbiased

$$p(y|\theta) = \int \hat{p}_{m,n}(y|\theta,u)p(u)du, \tag{3}$$

then $\int_u \tilde{\pi}(\theta,u)du = \pi(\theta) = p(y|\theta)p(\theta)/p(y)$. The efficiency of the simulation is crucially dependent on the variance of $\hat{p}_{m,n}$, where $\sigma^2 = V[\log \hat{p}_{m,n}]$ around 1-3 has been shown to be optimal (Doucet et al., 2015).

Correlated PMMH

One drawback of the standard PMMH algorithm is that m is required to be large to obtain a σ^2 around the optimal value. Deligiannis et al. (2015) propose to correlate the particles v in the standard PMMH algorithm using a Gaussian autoregressive kernel $K(v_c, v_p)$ with a transition defined by

$$v_p = \phi v_c + \sqrt{1-\phi^2}\varepsilon, \quad v_c, \varepsilon \overset{\text{ind.}}{\sim} \mathcal{N}(0, I).$$

We use the v_c and v_p through a Gaussian copula to induce correlation between each pair of selection indicators

$$u_c^{(i)} = \mathbb{1}\left(\Phi(v_c^{(i)}) \leq \frac{m^\star}{n}\right) \text{ and } u_p^{(i)} = \mathbb{1}\left(\Phi(v_p^{(i)}) \leq \frac{m^\star}{n}\right),$$

where $m^\star = E[\sum u_i]$, $E[u_i] = m^\star/n$ and Φ is the standard Gaussian cdf. Setting ϕ close to 1 induces a strong correlation ρ between the logarithms of the estimators in (2), which makes it possible to target a larger σ^2 for a given efficiency, thereby reducing the sample size m.

An alternative and more direct way to control for the correlation ρ is proposed in Tran et. al. (2016). Here u is instead a vector of observation indices

$$u = (u_1, \ldots, u_m)',$$

which is divided into G blocks. The method updates one block at a time (jointly with θ). Furthermore, it can be shown that $\rho = 1 - 1/G$.

Estimating the likelihood

Efficient log-likelihood estimation

Let q_k be a control variate for the kth observation such that $l_k \approx q_k$. Rewrite $l(\theta)$ in Eq. as (1)

$$l = q + d, \; q = \sum_{k=1}^{n} q_k, \; d = \sum_{k=1}^{n} d_k, \text{ with } d_k = l_k - q_k.$$

The *difference estimator* estimates only d,

$$\hat{l}_{m,n} = q + \hat{d}_{m,n}, \; \hat{d}_{m,n} = \frac{1}{m}\sum_{i=1}^{m}\frac{d_{u_i}}{p_{u_i}} \text{ with } p_k = \Pr(u_i = k).$$

This is equivalent to Importance Sampling (IS) (integral w.r.t the counting measure) with importance function p_k and control variates q_k. It is well known that $p_k \propto |d_k|$ gives an efficient IS estimate. Unfortunately, this requires knowledge of all d_k, defeating the purpose of subsampling. We focus on $p_k = 1/n$ and rely on q_k to homogenize the l_k.

Computationally efficient control variates

Define the Computational Cost (CC) for the standard MH that evaluates $l = \sum_{k=1}^{n} l_k$ as $\text{CC}[l] = O_l(n)$, where the subscript indicates the cost of evaluating a single l_k. Similarly,

$$\text{CC}\left[\hat{l}_{m,n}\right] = O_q(n) + O_d(m).$$

We now turn to two particular control variates that reduce $O_q(n)$ significantly.

1. Cluster the data space $z_k = (y_k, x_k)$ into N_C number of clusters. Within each cluster, let q_k be a second order Taylor approximation of l_k around the centroid. $\text{CC}\left[\hat{l}_{m,n}\right] = O_q(N_C) + O_d(m)$.
2. A Taylor approximation around $\theta^\star$ (Bardenet et al., 2015). $\text{CC}\left[\hat{l}_{m,n}\right] = O(1) + O_d(m)$.

Note that $q_k(n)$ improves as n increase. For 2. we have

$$d_k = l_k - q_k(n) = O^\pi(n^{-\alpha}), \text{ where } \alpha = 3/2.$$

Asymptotic properties

We consider two asymptotic cases

(i) $m \to \infty$ for a fixed n,
(ii) $n \to \infty$ and $m \to \infty$, with $m = O(n^\gamma)$.

The variance of $\hat{l}_{m,n}$ and its unbiased estimate are

$$\sigma^2(m,n) = \frac{n^2 \text{V}[d_k]}{m} \text{ and } \hat{\sigma}^2_{m,n} = \frac{n^2 S^2}{m},$$

where $E[S^2] = \text{V}[d_k]$ and $\text{V}[\hat{\sigma}^2_{m,n}] = \Omega(m,n) = \frac{n^4}{m^2}\text{V}[S^2]$.

Lemma. *The following m and n-asymptotics hold*:

$$\sigma^2(m,n) = \begin{cases} O(m^{-1}) \\ O(n^{2(1-\alpha)-\gamma}), \end{cases} \quad \Omega(m,n) = \begin{cases} O(m^{-3}) \\ O(n^{4(1-\alpha)-3\gamma}). \end{cases}$$

Bias-correction

Motivated by $\hat{l}_{m,n} \sim \mathcal{N}(l, \sigma^2)$ (m and n-asymptotically) we propose

$$\hat{p}_{m,n}(y|\theta,u) = \exp\left(\hat{l}_{m,n} - \hat{\sigma}^2_{m,n}/2\right).$$

If we instead correct with the true σ^2, the unbiasedness condition in Eq. (3) holds. However, σ^2 is not feasible because it requires the full data set. Our PMMH therefore targets a perturbed posterior

$$\pi_{m,n}(\theta) \propto \left(\int \hat{p}_{m,n}(y|\theta,u)p(u)du\right)p(\theta) = p_{m,n}(y|\theta)p(\theta)$$

Theorem. *The following m and n-asymptotics hold*:

$$\frac{|\pi_{m,n}(\theta) - \pi(\theta)|}{\pi(\theta)} \leq \begin{cases} O(m^{-1}) \\ O(n^{-a}), \end{cases} \quad a = \min \begin{cases} 3(\alpha-1)+2\gamma \\ 2(\alpha-1)+\gamma \\ 4(\alpha-1)+3\gamma. \end{cases}$$

Application: Comparison to other subsampling approaches

Models

We consider two AR(1) processes with Student-t iid errors $\epsilon_t \sim t(\nu = 5)$, $\nu = 5$.

$$y_t = \begin{cases} \beta_0 + \beta_1 y_{t-1} + \epsilon_t & ,[\text{M}_1, \theta = (\beta_0 = 0.3, \beta_1 = 0.6)] \\ \mu + \rho(y_{t-1} - \mu) + \epsilon_t & ,[\text{M}_2, \theta = (\mu = 0.3, \rho = 0.99)] \end{cases}$$

and priors

$$p(\beta_0, \beta_1) \overset{\text{ind.}}{=} \mathcal{U}(-5,5)\cdot\mathcal{U}(0,1), \quad p(\mu, \rho) \overset{\text{ind.}}{=} \mathcal{U}(-5,5)\cdot\mathcal{U}(0,1).$$

We sample $n = 100,000$ observations from the DGPs and run all algorithms for $N = 55,000$ iterations.

Comparison

We compare our method to Austerity MH (Korattikaria et al., 2014), Firefly Monte Carlo (Maclaurin and Adams, 2014), the confidence sampler and the confidence sampler with proxies (Bardenet et al., 2014, 2015). We measure the performance for algorithm $\mathcal{A}$ relative to standard MH with the Relative Effective Draws (RED)

$$\text{RED} = \left(\frac{\text{ESS}_\mathcal{A}}{\text{CC}_\mathcal{A}}\right) \Big/ \left(\frac{\text{ESS}_\text{MH}}{n}\right)$$

with

$$\text{ESS} = N/\text{IF}, \; \text{IF} = 1 + 2\sum_{l=1}^{\infty}\rho_l,$$

where ρ_k is the auto-correlation at lag k.

We find that the correlated PMMH algorithms outperform both standard MH and other subsampling approaches recently proposed in the literature.

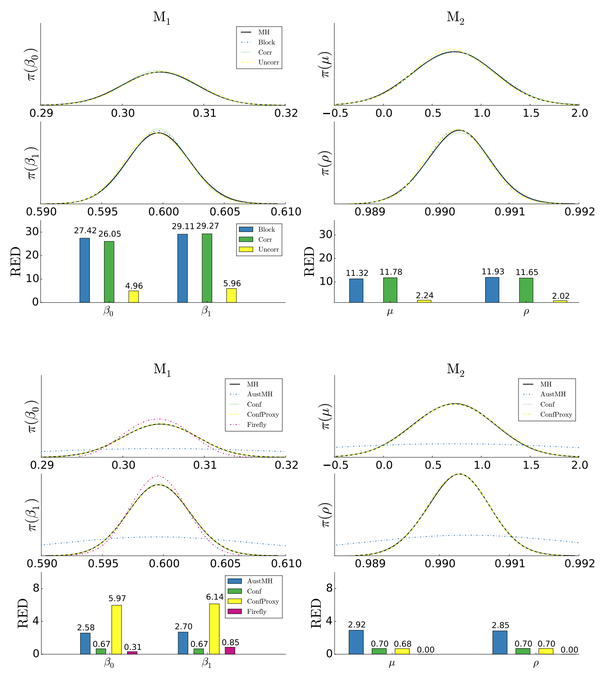
Table 1: Mean of sampling fraction for MH and PMMH

	MH	Uncorr	Corr	Block
M_1	1.000	0.055	0.023	0.023
M_2	1.000	0.159	0.059	0.059

Table 2: Mean of sampling fraction for other approaches

	AustMH	Conf	ConfProxy	Firefly
M_1	0.197	1.493	0.160	0.100
M_2	0.192	1.489	1.497	0.137

Posterior approximation and RED



Footnotes

Quiroz: (quiroz.matias@gmail.com) *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University* and *Research Division, Department of Monetary Policy, Sveriges Riksbank.*
Villani: *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University.*
Kohn: *Australian School of Business, University of New South Wales.*
Tran: *Discipline of Business Analytics, University of Sydney.*

An earlier preprint of this paper is available at http://arxiv.org/abs/1404.4178v3.

Disclaimer: Results are preliminary.

References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697-725.
Bardenet, R., Doucet, A., and Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of The 31st International Conference on Machine Learning*, pages 405-413.
Bardenet, R., Doucet, A., and Holmes, C. (2015). On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*.
Deligiannidis, G., Doucet, A., and Pitt, M. K. (2015). The correlated pseudo-marginal method. *arXiv preprint arXiv:1511.04992v2*.
Doucet, A., Pitt, M.K., Deligiannidis, G. and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* , p.asu075.
Korattikara, A., Chen, Y., and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of The 31st International Conference on Machine Learning*, pages 181-189.
Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. *arXiv preprint arXiv:1403.5693*.
Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016). Block-wise pseudo-marginal Metropolis-Hastings. *arXiv preprint arXiv:1603.02485v2*.