The Block-Poisson Estimator for Optimally Tuned Exact Subsampling MCMC

Matias Quiroz, Minh-Ngoc Tran, **Mattias Villani**
Robert Kohn and Khue-Dung Dang

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview

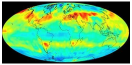
- **Pseudo-Marginal MCMC and Subsampling MCMC**
- **The Block-Poisson likelihood estimator**
- **Optimal subsample size**
- **Empirical results**

Motivation

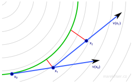
- **MCMC** is still the workhorse for **Bayesian inference**.
- **MCMC is often slow**
  - Many iterations
  - Need to evaluate the likelihood function in each iteration
- **Hamiltonian Monte Carlo** (**HMC**)
  - quickly traverse high-dimensional parameter spaces
  - ... at the cost of a very large number of gradient evaluations.
- **Subsampling MCMC**: **estimate the likelihood** from a subsample in each MCMC iteration. Fewer evaluations. Faster!

Likelihood evaluations are often very expensive

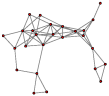
- **High-dimensional spatio-temporal problems** (GMRFs)



- Models where **numerical methods** are needed for evaluating $p(y_i|\theta)$ (ODEs, optimization, etc)



- **Doubly intractable problems** with costly normalization constants (ERGMs)



- So called **Big data** problems with many observations.

The Pseudo-Marginal MH (PMMH) algorithm

- Initialize $\left(\theta^{(0)}, \mathbf{u}^{(0)}\right)$ and iterate for $i = 1, 2, ..., N$

  1. Sample $\theta_p \sim q\left(\cdot|\theta^{(i-1)}\right)$ and $\mathbf{u}_p \sim p(\mathbf{u})$ to obtain the **unbiased likelihood estimate** $\hat{p}(\mathbf{y}|\theta_p, \mathbf{u}_p)$
  2. Compute the **acceptance probability**

  $$\alpha = \min\left(1, \frac{\hat{p}\left(\mathbf{y}|\theta_p, \mathbf{u}_p\right) p(\theta_p)}{\hat{p}\left(\mathbf{y}|\theta^{(i-1)}, \mathbf{u}^{(i-1)}\right) p(\theta^{(i-1)})} \frac{q\left(\theta^{(i-1)}|\theta_p\right)}{q\left(\theta_p|\theta^{(i-1)}\right)}\right)$$

  3. With probability α set $\left(\theta^{(i)}, \mathbf{u}^{(i)}\right) = \left(\theta_p, \mathbf{u}_p\right)$ and $\left(\theta^{(i)}, u^{(i)}\right) = \left(\theta^{(i-1)}, \mathbf{u}^{(i-1)}\right)$ otherwise.

- Targets a joint distribution $\tilde{p}(\theta, \mathbf{u}|\mathbf{y})$ with marginal $p(\theta|\mathbf{y})$ [1].
- True **for any** positive unbiased estimator, but ...
- ... large $\mathbb{V}\left(\hat{p}(\mathbf{y}|\theta, \mathbf{u})\right)$ gives **inefficient sampling**.

Bias-corrected log-likelihood based estimator [3]

- Estimate $L(\theta) = \exp\left(\ell(\mathbf{y}|\theta, \mathbf{u})\right)$ by **bias-correcting** $\exp\left(\hat{\ell}(\mathbf{y}|\theta, \mathbf{u})\right)$. HMC extension [2].
- Subsampling estimate of the log-likelihood for iid data

$$\hat{\ell}(\mathbf{y}|\theta, \mathbf{u}) = \frac{n}{m} \sum_{i \in \mathbf{u}} \ell(y_i|\theta)$$

- **Difference estimator** with **control variates** $q_i(\theta) \approx \ell(y_i|\theta)$ [3]

$$\hat{\ell}(\mathbf{y}|\theta, \mathbf{u}) = \sum_{i=1}^{n} q_i(\theta) + \frac{n}{m} \sum_{i \in \mathbf{u}} \underbrace{\left(\ell(y_i|\theta) - q_i(\theta)\right)}_{d_i(\theta)}$$

- Two types of control variates
  - **Parameter-expanded** [4]
  - **Data-expanded** [3]
- Targets a **perturbed posterior** with TV-norm error of $O(n^{-1}m^{-2})$.

Doubly intractable problems

- Doubly intractable

$$p(\theta|\mathbf{y}) \propto \frac{f(\mathbf{y};\theta)p(\theta)}{Z(\theta)}$$

- Common:
  - **Graph-based models (ERGMs)** $Z(\theta)$ is a sum over all graphs
  - **Spatial models** like Potts model.
  - **Directional statistics** $Z(\theta)$ is an intractable integral over the sphere.

- Exponential augmentation trick: $v \sim \mathrm{Exp}(Z(\theta))$ [5]

$$\tilde{\pi}(\theta, v) \propto \exp\left(-vZ(\theta)\right) f(\mathbf{y};\theta)p(\theta)$$

PMMH with dependent subsamples

- What really matters for MH is the variance of

$$\log \frac{\hat{p}\left(\mathbf{y}|\theta_p, \mathbf{u}_p\right)}{\hat{p}\left(\mathbf{y}|\theta^{(i-1)}, \mathbf{u}^{(i-1)}\right)}$$

- **Correlated Pseudo Marginal** (**CPM**) [6, 7]: correlate the **u** over MH iterations using an autoregressive proposal $\mathbf{u}^{(i)} = \phi\mathbf{u}^{(i-1)} + \epsilon$.

- **Subsampling** context: correlate binary subsampling indicators with Gaussian copula [3].

- **Block Pseudo Marginal** (**BPM**) [8]: partition $\mathbf{u} = (u_1, ..., u_m)$ in blocks and **update a single block** jointly with θ at each iteration.

The Block-Poisson estimator

- The **Block-Poisson estimator** of the likelihood $L(\theta)$:

$$\hat{L}_B(\theta) \equiv \prod_{l=1}^{\lambda} \xi_l$$

$$\xi_l \equiv \exp\left(\frac{a+\lambda}{\lambda}\right) \prod_{h=1}^{\mathcal{X}_l} \left(\frac{\hat{\ell}_m^{(h,l)} - a}{\lambda}\right)$$

  - $\lambda \in \mathbb{N}^+$ and $a \in \mathbb{R}$
  - $\hat{\ell}_m^{(h,l)}$ is an unbiased estimator of ℓ from a batch of m obs
  - $\mathcal{X}_1, ..., \mathcal{X}_\lambda \overset{iid}{\sim} \text{Pois}\,(1)$
- Product form allows us to use **Block Pseudo Marginal** (**BPM**).
- $\hat{L}_B(\theta)$ requires on average λm evaluations of ℓ_i's.

Properties of the Block-Poisson estimator

$$\hat{L}_B(\theta) = \prod_{l=1}^{\lambda} \xi_l, \text{ where } \xi_l = \exp\left(\frac{a+\lambda}{\lambda}\right) \prod_{h=1}^{\mathcal{X}_l} \left(\frac{\hat{\ell}_m^{(h,l)} - a}{\lambda}\right)$$

- **Unbiased**: $\mathbb{E}\left(\hat{L}_B(\theta)\right) = L(\theta)$ for all $\theta \in \Theta$.
- **Positive**: $\hat{L}_B(\theta)$ is almost surely positive only if $\hat{\ell}_m^{(h,l)} \geq a$ almost surely for all h and l.
- For a given λ, $\mathbb{V}\left(\hat{L}_B(\theta)\right)$ is minimized for $a = \ell - \lambda$.
- $\mathbb{V}\left(\hat{L}_B(\theta)\right) = \mathbb{V}\left(\hat{L}_P(\theta)\right)$ where $\hat{L}_P(\theta)$ is the usual Poisson estimator in e.g. [9].

Signed PMMH

- Forcing a to be a **lower bound** for all $\hat{\ell}_m^{(h,l)}$ is not good:
  - Usually need to know ℓ_i for all data points.
  - $a = \ell - \lambda$ implies that λ will be large. Costly!
- **Soft lower bound:**
  - $\Pr(\hat{\ell}_m^{(h,l)} \geq a)$ close to one.
  - More efficient, but $\hat{L}_B(\theta) < 0$ possible.

- **Signed PMMH** [5]
  - **Run PMMH on absolute value** $\left|\hat{L}_B(\theta)\right| p(\theta)$
  - **Correct for the sign** $s = \text{Sign}\left(\hat{L}_B(\theta)\right)$ using importance sampling

$$\widehat{\mathbb{E}\psi(\theta)} = \frac{\sum_{i=1}^N \psi(\theta^{(i)}) s^{(i)}}{\sum_{i=1}^N s^{(i)}}.$$

Optimal tuning of Signed PMMH based on $\hat{L}_B(\theta)$

- **Optimal** subsample size m in regular PMMH?
- Minimize (normalized) asymptotic variance of PMMH estimates of $\mathbb{E}\left[\psi(\theta)\right]$ per unit of computing time

$$\mathrm{CT}(m) \propto m \cdot \mathrm{IACT}(\sigma^2_{\log \hat{L}})$$

- Regular PMMH is optimal when $\sigma^2_{\log \hat{L}} \approx 1$ [10, 11].
- **Optimal** λ and m in **signed PMMH** minimizes

$$\mathrm{CT}(\lambda, m) \propto m\lambda \cdot \frac{\mathrm{IACT}\left[\sigma^2_{\log\left|\hat{L}_B\right|}(\lambda, m)\right]}{\left(2\tau(\lambda, m) - 1\right)^2}$$

- Optimal λ and m balances
  1. The **cost** of computing $\hat{L}_B$, which is $m\lambda$ on average
  2. **MH inefficiency**, IACT
  3. Probability of a **positive sign** $\tau(\lambda, m)$

Optimal tuning of Signed PMMH

- To compute $\mathrm{CT}(\lambda, m)$, we need expressions for:
  - $\mathrm{IACT}(\cdot)$
  - $\sigma^2_{\log|\hat{L}_B|}(\lambda, m)$
  - $\tau(\lambda, m)$

- The **derivation of** IF is an extension of the theory in [10] to blocked signed PMMH.

- Idealized assumptions:
  - Perfect MH proposal for θ
  - $\sigma^2_{\log|\hat{L}_B|}$ is not a function of θ

- **Heuristic guidelines**. But accurate in experiments.

- **Conservative guidelines**: $m\lambda$ is not suggested too small.

$\tau \equiv \Pr(\hat{L}_B \geq 0)$

- Under the minimum variance condition $a = \ell - \lambda$

$$\hat{L}_B(\theta) = \prod_{l=1}^{\lambda} \xi_l, \text{ where } \xi_l = \exp\left(\frac{\ell}{\lambda}\right) \prod_{h=1}^{\mathcal{X}_l} \left(\frac{\hat{\ell}_m^{(h,l)} - \ell}{\lambda} + 1\right)$$

- Applying a result from Feller's first book twice:

$$\Pr(\hat{L}_B \geq 0) = \frac{1}{2}\left[1 + (1 - 2\Psi(m, \lambda))^{\lambda}\right]$$

where

$$\Psi(m, \lambda) \equiv \Pr(\xi_l < 0) = \frac{1}{2}\sum_{j=1}^{\infty}\left[1 - (1 - 2\Pr(A_m < 0))^{j}\right]\Pr(\mathcal{X}_l = j),$$

$\mathcal{X}_l \overset{iid}{\sim} \text{Pois}(1)$ and $A_m = \frac{\hat{\ell}_m - \ell}{\lambda} + 1$.

$\sigma^2_{\log\left|\hat{L}_B\right|}(\lambda, m)$

- Under the condition $a = \ell - \lambda$ we have

$$\begin{aligned}
\log\left|\hat{L}\right| &= \ell + \sum_{l=1}^{\lambda} \sum_{h=1}^{\mathcal{X}_l} \log\left(\left|\frac{\hat{\ell}_m^{(h,l)} - \ell}{\lambda} + 1\right|\right) \\
&= \ell + \frac{1}{2} \sum_{l=1}^{\lambda} \sum_{h=1}^{\mathcal{X}_l} \log\left(\frac{\hat{\ell}_m^{(h,l)} - \ell}{\lambda} + 1\right)^2
\end{aligned}$$

- $\hat{\ell}_m^{(h,l)} \sim \text{Normal} \Rightarrow \sigma^2_{\log\left|\hat{L}_B\right|}(\lambda, m)$ is the variance of a random sum of logs of non-central χ^2 variables.
- Non-central χ^2 is a Poisson mixture of central χ^2 [12]
- Moments of log central χ^2 are known from [13]
- Law of total variance

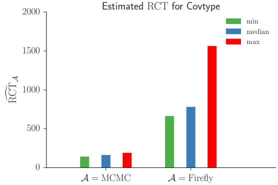
Optimal tuning - normal case

- Assume $\hat{\ell}_m^{(h,l)} \sim \text{Normal}$.
- Both $\Pr(\hat{L}_B \geq 0)$ and $\sigma^2_{\log\left|\hat{L}_B\right|}(\lambda, m)$ are functions of the variance of $\hat{\ell}_m^{(h,l)}$

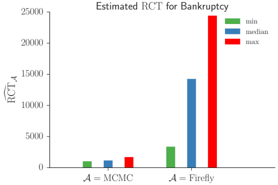
$$\mathbb{V}(\hat{\ell}_m^{(h,l)}(\theta)) = \frac{n^2}{m}\sigma^2_{\ell_i}(\theta)$$

- Optimal tuning therefore depends on $\sigma^2_{\ell_i}(\theta)$.
- Solution: estimate $\sigma^2_{\ell_i}(\theta)$ from a subsample for some selected θ.
- What if $\hat{\ell}_m^{(h,l)}$ are not normal?
- Set $m = 20$ and rely on the CLT. Optimize only λ.

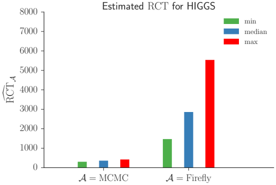
Relative CT - logistic regression on three real datasets



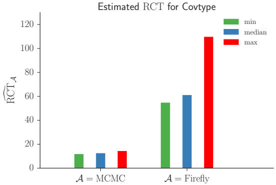
(A) Covtype data.



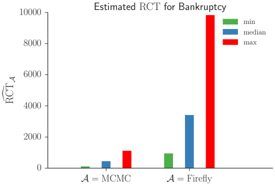
(B) Bankruptcy data.



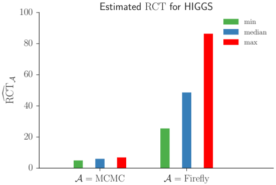
(C) HIGGS data.



(A) Covtype data.

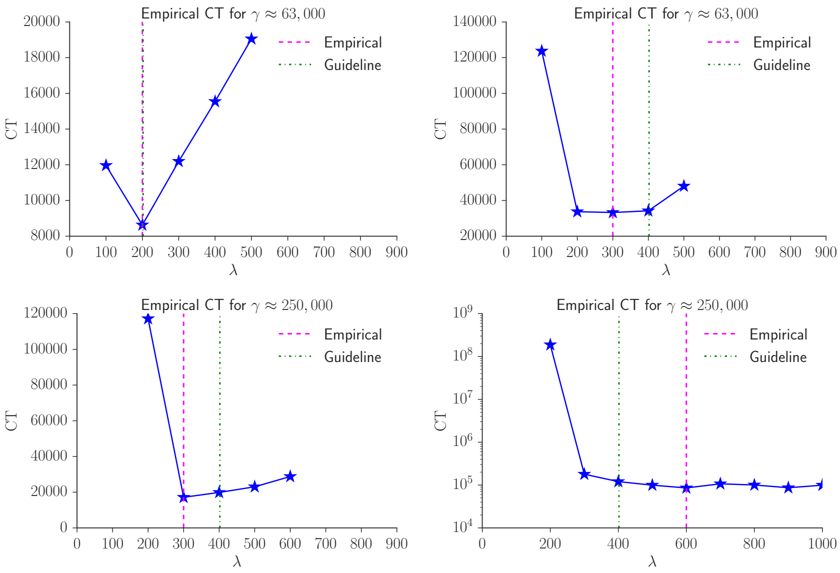


(B) Bankruptcy data.



(C) HIGGS data.

Checking the optimality guidelines

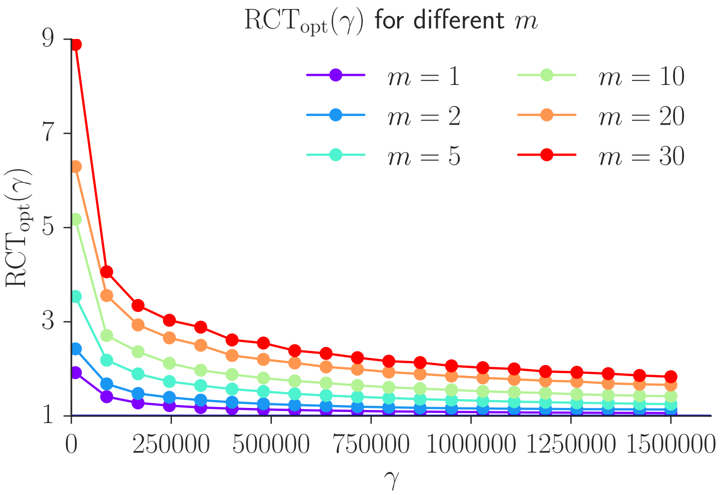


(A) γ does not depend on θ.

(B) γ depends on θ.

Relative CT: Signed PMMH vs Approximate PMMH

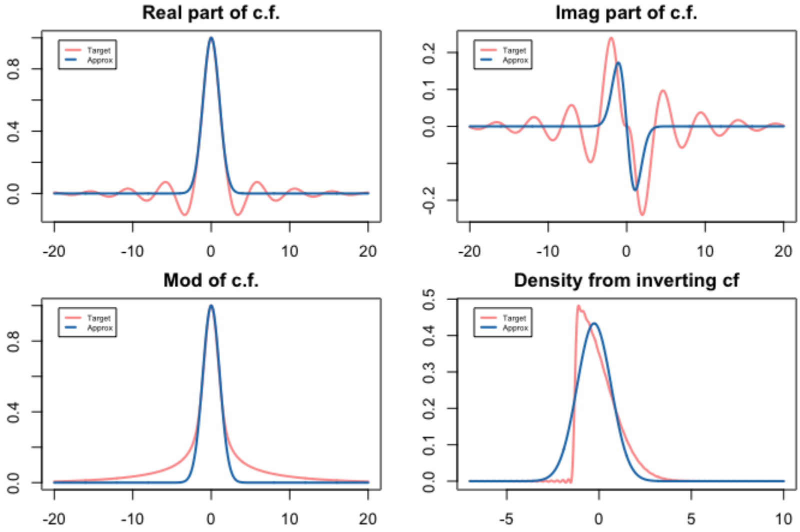
- $\gamma = n^2\sigma^2_{\ell_i}$



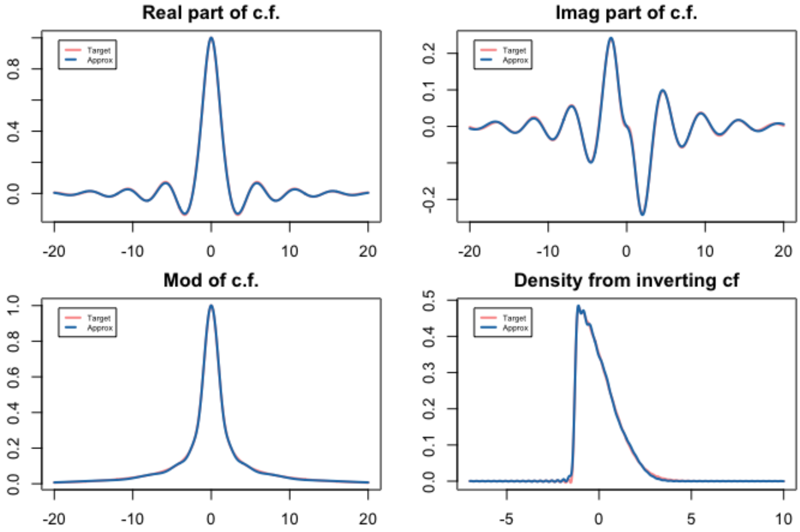
Optimal tuning - mixture of normals case

- We can instead assume that $\hat{\ell}_m^{(h,l)}$ follows a **mixture of normals.**
- Mixture of normals are **universal approximators**.
- Both $\Pr(\hat{L}_B \geq 0)$ and $\sigma^2_{\log|\hat{L}_B|}$ are still **tractable**.
- ... but estimating $\sigma^2_{\ell_i}(\theta)$ is not enough anymore.
- How to fit a mixture of normals to $\hat{\ell}_m^{(h,l)}$?
- **Matching characteristic functions** (c.f.)
  1. Fit any distribution to a subsample of ℓ_i's and get the c.f. $\varphi_\ell(t)$.
  2. Compute the c.f. of $\hat{\ell}_m^{(h,l)}$ as $\varphi_{\hat{\ell}_m}(t) = (\varphi_\ell(t/m))^m$.
  3. Approximate the distribution of $\hat{\ell}_m^{(h,l)}$ by a normal mixture by L2-matching of c.f.'s. Plancherel's theorem.

Matching a 1-component MoN to skewed normal



Matching a 5-component MoN to skewed normal



Conclusions

- **Subsampling** to speed up MCMC and HMC.
- **Control variates** and **slowly evolving subsamples** are important for efficiency.
- **Block-Poisson** is an **unbiased** and **efficient** estimator of the likelihood.
- **Optimal tuning of Signed PMMH** with Block-Poisson estimator.
- **Very large speed-ups** compared to regular MCMC and FireFly MC.
- Can be used to optimally tune Signed PMMH in **doubly intractable problems**.

References

C. Andrieu and G. O. Roberts, "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, pp. 697–725, 2009.

K.-D. Dang, M. Quiroz, R. Kohn, M.-N. Tran, and M. Villani, "Hamiltonian monte carlo with energy conserving subsampling," *arXiv preprint arXiv:1708.00955*, 2017.

M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran, "Speeding up mcmc by efficient data subsampling," *Journal of the American Statistical Association*, no. forthcoming, pp. 1–35, 2018.

R. Bardenet, A. Doucet, and C. Holmes, "On markov chain monte carlo methods for tall data," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1515–1557, 2017.

A.-M. Lyne, M. Girolami, Y. Atchade, H. Strathmann, D. Simpson, *et al.*, "On russian roulette estimates for bayesian inference with doubly-intractable likelihoods," *Statistical science*, vol. 30, no. 4, pp. 443–467, 2015.

G. Deligiannidis, A. Doucet, and M. K. Pitt, "The correlated pseudo-marginal method," *arXiv preprint arXiv:1511.04992*, 2015.

J. Dahlin, F. Lindsten, J. Kronander, and T. B. Schön, "Accelerating pseudo-marginal metropolis-hastings by correlating auxiliary variables," *arXiv preprint arXiv:1511.05483*, 2015.

M. Quiroz, M.-N. Tran, M. Villani, R. Kohn, and K.-D. Dang, "The block-Poisson estimator for optimally tuned exact subsampling MCMC," *arXiv preprint arXiv:1603.08232*, 2018.

O. Papaspiliopoulos, "A methodological framework for monte carlo probabilistic inference for diffusion processes," 2009.

M. K. Pitt, R. d. S. Silva, P. Giordani, and R. Kohn, "On some properties of Markov chain Monte Carlo simulation methods based on the particle filter," *Journal of Econometrics*, vol. 171, no. 2, pp. 134–151, 2012.

- A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn, "Efficient implementation of markov chain monte carlo when using an unbiased likelihood estimator," *Biometrika*, vol. 102, no. 2, pp. 295–313, 2015.
- C. Walck, "Hand-book on statistical distributions for experimentalists," tech. rep., 1996. http://inspirehep.net/record/1389910/files/suf9601.pdf.
- S. E. Pav, "Moments of the log non-central chi-square distribution," *arXiv preprint arXiv:1503.06266*, 2015.