Recent Developments in Subsampling for Large-Scale Bayesian Inference

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Overview

- Subsampling MCMC/HMC
- Optimal Tuning of Subsampling MCMC
- Grouped Control Variates
- Subsampling for Stationary Time Series

- Slides: http://mattiasvillani.com/news

Large-scale project: many papers and researchers

- **Robert Kohn**, UNSW Sydney and **Matias Quiroz**, UTS Sydney
- **Minh-Ngoc Tran**, University of Sydney
- **Khue-Dung Dang**, UNSW Sydney
- **Robert Salomone**, UNSW Sydney

The Metropolis-Hastings (MH) algorithm

- **Bayesian inference**

$$\pi(\theta) \propto L(\theta)p(\theta)$$

- Initialize $\theta^{(0)}$ and iterate for $k = 1, 2, ..., N$

  1. Sample $\theta_p \sim q\left(\cdot|\theta^{(k-1)}\right)$ (the **proposal distribution**)
  2. Accept θ_p with **acceptance probability**

$$\alpha = \min\left(1, \frac{L(\theta_p)p(\theta_p)}{L(\theta^{(k-1)})p(\theta^{(k-1)})}\frac{q\left(\theta^{(k-1)}|\theta_p\right)}{q\left(\theta_p|\theta^{(k-1)}\right)}\right)$$

- **Costly** to evaluate $L(\theta_p)$ when n is large. **Big data**.

Naive Subsampling MH

- Estimate log-likelihood $\ell(\theta)$ from subsample of size $m \ll n$

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

- Unbiased: $\mathbb{E}_{\mathbf{u}}[\hat{\ell}(\theta, \mathbf{u})] = \ell(\theta)$.
- Run Pseudo-marginal MH with $\hat{L}(\theta, \mathbf{u}) = \exp\left(\hat{\ell}(\theta, \mathbf{u})\right)$.

- Initialize $\left(\theta^{(0)}, \mathbf{u}^{(0)}\right)$ and iterate for $k = 1, 2, ..., N$
  1. Sample $\theta_p \sim q\left(\cdot | \theta^{(k-1)}\right)$ and subsample $\mathbf{u}_p \sim p(\mathbf{u})$
  2. Accept $(\theta_p, \mathbf{u}_p)$ with acceptance probability

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

Isses with Naive Subsampling MH

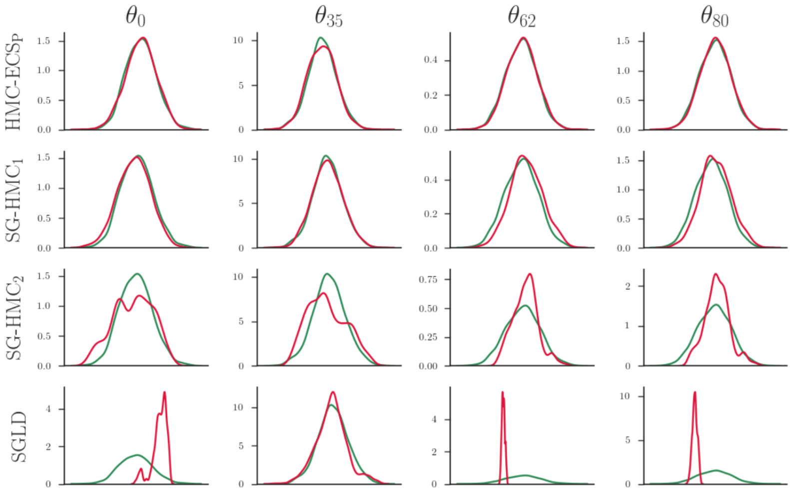
- PMMH samples from $\pi(\theta)$ if $\hat{L}$ **is unbiased** [1]
  - **Approximate bias correction of** $\exp\left(\hat{\ell}(\theta, \mathbf{u})\right)$ [2]
    Theorem: $O(m^{-2}n^{-1})$ posterior perturbation in TV-norm. [3]
  - **Unbiased Block-Poisson estimator** + **Signed PMMH**. [4]
- **Low** $\mathbb{V}\left(\hat{L}\left(\theta, \mathbf{u}\right)\right)$ crucial for **efficient sampling**. Stuck.
  - Difference estimator and **control variates** [3, 5]
  - **Optimal tuning** of m [4]
  - **Block Pseudo-marginal**: only refresh part of the subsample. [6, 7]
- **High-dim** case: **Energy Conserving Subsampling HMC**. Estimate likelihood and Hamiltonian dynamics from **same** subsample. [8]

Logistic spline regression, 81 parameters

- Firm bankruptcy data. $n = 4,748,089$ firm-year obs.
- Subsample size: $m = 1000$.
- **Computational Time (CT)**:
  - Computing time to obtain the equivalent of an iid draw.
  - Balances **computational cost** and **MCMC inefficiency**.
  - **Relative CT (RCT)**

	# evaluations	RCT	IF
HMC	110601×10^6	7691.8	2.20
HMC-ECS_P	14.02×10^6	1	2.20
SG-HMC_1	120×10^6	9.49	2.42
SG-HMC_2	14×10^6	100.29	226.75
SGLD	11×10^6	230	649.0

Bias - Logistic spline regression, 81 parameters



The Block-Poisson estimator

- The Block-Poisson estimator of the likelihood $L(\theta)$: [4, 9]
  - For $l = 1, ..., \lambda$
    - draw $\mathcal{X}_l \sim \mathrm{Pois}(1)$
    - draw $\mathcal{X}_l$ mini-batches of data of size m.
    - Compute unbiased mini-batch estimators of $\ell(\theta)$

$$\hat{\ell}_m^{(h,l)}, \quad \text{for } h = 1, ..., \mathcal{X}_l$$

  - Construct likelihood estimate for some constant $a \in \mathbb{R}$

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

- Product form of $\hat{L}_B(\theta)$: use Block Pseudo Marginal.
- Unbiased: $\mathbb{E}\left(\hat{L}_B(\theta)\right) = L(\theta)$ for all $\theta \in \Theta$.
- Positive: $\hat{L}_B(\theta) > 0$ only if $\hat{\ell}_m^{(h,l)} > a$ for all h and l.

Signed HMC-ECS

- For a given λ, $\mathbb{V}\left(\hat{L}_B(\theta)\right)$ is minimized for $a = \ell - \lambda$.
- Forcing a to be a **lower bound** for all $\hat{\ell}_m^{(h,l)}$ is impractical:
  - Usually need to know ℓ_i for all data points.
  - $a = \ell - \lambda$ implies that λ will be large. Costly!
- **Soft lower bound:** Set a so $\Pr(\hat{\ell}_m^{(h,l)} \geq a) \approx 1$.
  More efficient, but $\hat{L}_B(\theta) < 0$ possible.

- **Signed HMC-ECS** [10]
  - Run **PMMH** on $\left|\hat{L}_B(\theta)\right| p(\theta)$ and store $\mathrm{s} = \mathrm{Sign}\left(\hat{L}_B(\theta)\right)$.
  - **Correct for sign** using importance sampling

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

  where $\psi(\theta)$ is a function of the parameters.

Optimal tuning of Signed HMC-ECS

- **Optimal** λ and m minimizes **Computational Time (CT)**:

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

- Optimal λ and m **balances**
  1. The **cost** of computing $\hat{L}_B$, which is $O(m\lambda)$ on average
  2. **MH inefficiency**, IF
  3. Probability of a **positive sign** $\tau(\lambda, m) \equiv \Pr(\hat{L}_B \geq 0)$.

Optimal tuning of Signed HMC-ECS

- We derive **analytical** expressions for all parts of $\mathrm{CT}(\lambda, m)$:
  - IF
  - $\sigma^2_{\log|\hat{L}_B|}(\lambda, m)$
  - $\tau(\lambda, m)$
- Need to assume a **distribution for $\hat{\ell}_m^{(h,l)}$**.
- Approach 1: Normal $\hat{\ell}_m^{(h,l)}$ by CLT when $m > 20$.
- Approach 2: Universal approximator by Mixture of normals.

Optimal tuning - normal case

- Set $m = 20$ and assume $\hat{\ell}_m^{(h,l)} \sim \mathrm{Normal}$ by CLT. Optimize λ.
- Both $\Pr(\hat{L}_B \geq 0)$ and $\sigma^2_{\log|\hat{L}_B|}(\lambda, m)$ are functions of

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

- Estimate $\sigma^2_{\ell_i}(\theta)$ from a subsample for some selected θ.
- However, numerical experiments tell us that $m = 1$ is optimal.
- Alternative: Approx $\hat{\ell}_m^{(h,l)}$ by mixture by matching characteristic functions. [4]

Grouped control variates

- **Difference estimator** with **control variates** $q_j(\theta)$

$$\hat{\ell}(\theta, \mathbf{u}) = \sum_{j=1}^{n} q_j(\theta) + \frac{n}{m} \sum_{i \in \mathbf{u}} \left(\log p(y_i|\theta) - q_i(\theta) \right)$$

- $q_j(\theta)$ by **quadratic expansion** of $\log p(y_i|\theta)$ around $\theta^\star$.
- Problematic when $\log p(y_i|\theta)$ is far from quadratic.

- **Grouped control variates** based on grouping of data points

$$\ell(\theta) = \underbrace{\ell_1(\theta) + \ldots + \ell_{|G_1|}(\theta)}_{\ell_{G_1}(\theta)} + \underbrace{\ell_{|G_1|+1}(\theta) + \ldots + \ell_{|G_1|+|G_2|}(\theta)}_{\ell_{G_2}(\theta)} + \ldots$$

- **Subsample groups**, not individual observations.
- Bernstein-von Mises: $\ell_{G_k}(\theta)$ approach quadratic as $|G_k| \to \infty$.
- **Grouped difference estimator** [11]

$$\hat{\ell}_{\mathrm{gr}}(\theta) = \sum_{k=1}^{|\mathcal{G}|} q_{G_k}(\theta) + \frac{|\mathcal{G}|}{m} \sum_{i=1}^{m} \left(\ell_{G_{u_i}}(\theta) - q_{G_{u_i}}(\theta) \right)$$

Subsampling MCMC for stationary time series

- **Covariance function** $\gamma_{\theta}(\tau)$, $\tau = 0, 1, \ldots$ and **spectral density**

$$f_{\theta}(\omega) \equiv \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_{\theta}(\tau) \exp(-\mathrm{i}\omega\tau) \text{ for } \omega \in (-\pi, \pi].$$

- **Discrete Fourier Transform** (DFT) of the time series

$$J(\omega_k) \equiv \frac{1}{\sqrt{2\pi}} \sum_{t=1}^{n} X_t \exp(-\mathrm{i}\omega_k t)$$

at $\omega_k \in \{2\pi k/n \text{ for } k = -\lceil n/2 \rceil + 1, \ldots, \lfloor n/2 \rfloor\}$.

- The **periodogram**

$$\mathcal{I}(\omega_k) = n^{-1} \left| J(\omega_k) \right|^2.$$

- **Asympotically independent** periodogram ordinates

$$\mathcal{I}(\omega_k) \overset{indep}{\sim} \mathrm{Exp}(f_{\theta}(\omega_k)), \quad k = 1, \ldots, n$$

Subsampling MCMC for stationary time series

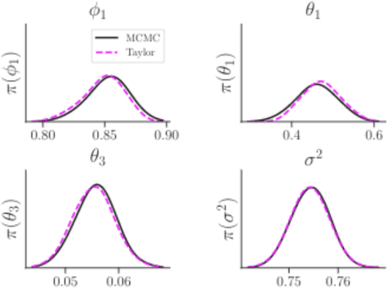
- **Whittle log-likelihood is a sum**

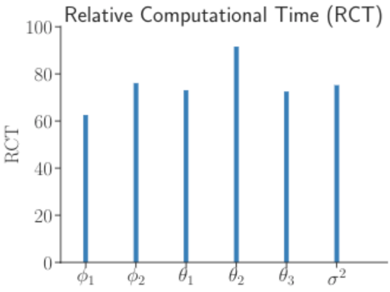
$$\ell_W(\boldsymbol{\theta}) \equiv -\sum_{\omega_k \in \Omega} \left(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{\mathcal{I}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} \right)$$

- Whittle may be **biased for small n**.
- But **subsampling** is only relevant for **large n**.
- **Subsampling** for stationary **time series** [11]
  - **Compute periodogram** before MCMC at cost $O(n \log n)$.
  - Estimate $\ell_W(\boldsymbol{\theta})$ by systematic **subsampling of frequencies**.
- Extensions:
  - **Tapering**
  - **Debiased Whittle**
  - Multidimensional FFT for **spatial data**.

ARMA(2,3) for temperature time series

- Temperature on $n = 44001$ days in Vancouver.





- Also ARFIMA example in [11]

Conclusions

- **Subsampling** to speed up MCMC and HMC.
- **Block-Poisson** is an **unbiased** and **efficient** estimator of the likelihood.
- **Optimal tuning of Signed HMC-ECS** with Block-Poisson estimator.
- **Very large speed-ups** compared to regular HMC and state-of-the-art subsampling algorithms.
- **Grouped control variates**
- Time series extension: **subsample periodogram** frequencies.

References

- C. Andrieu and G. O. Roberts, "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, pp. 697–725, 2009.
- D. Ceperley and M. Dewing, "The penalty method for random walks with uncertain energies," *The Journal of chemical physics*, vol. 110, no. 20, pp. 9812–9820, 1999.
- 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.
- 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.
- 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.

- M.-N. Tran, R. Kohn, M. Quiroz, and M. Villani, "Block-wise pseudo-marginal metropolis-hastings," *arXiv preprint arXiv:1603.02485*, 2016.
- G. Deligiannidis, A. Doucet, and M. K. Pitt, "The correlated pseudo-marginal method," *arXiv preprint arXiv:1511.04992*, 2015.
- 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.
- O. Papaspiliopoulos, "A methodological framework for monte carlo probabilistic inference for diffusion processes," 2009.
- 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.

R. Salomone, M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran, "Spectral subsampling mcmc for stationary time series," *arXiv preprint arXiv:1910.13627*, 2019.