Optimal Tuning of Subsampling Hamiltonian Monte Carlo

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

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

Overview

- Hamiltonian Monte Carlo (HMC)
- HMC-ECS: HMC with Energy Conserving Subsampling
- Block-Poisson estimator of the likelihood
- Signed HMC-ECS
- Optimal tuning of HMC-ECS
- Empirical results

- **Bayesian inference**

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

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

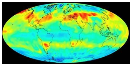
1. Sample $\theta_p \sim q\left(\cdot|\theta^{(i-1)}\right)$ (the **proposal distribution**)
2. Compute the **acceptance probability**

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

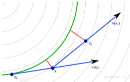
3. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Problem #1 - Likelihood evaluations can be costly.

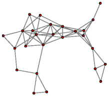
- **High-dimensional spatio-temporal problems** (GMRFs)



- 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.

- Hard to make **good proposals** in **high dimensional parameter spaces**.
- **Random walk Metropolis**

$$\theta_p|\theta^{(i-1)} \sim N(\theta^{(i-1)}, c \cdot \Sigma)$$

  traverses the parameter space slowly in high dimensions.
- Small c - high acceptance probability, but small steps.
- Large c - large steps, but low acceptance probability.

- Augment the posterior with **momentum** variables $\mathbf{r} \in \mathbb{R}^d$.
- **Extended target**:

$$\bar{\pi}(\theta, \mathbf{r}) \propto \exp\left(-\mathcal{H}(\theta, \mathbf{r})\right), \quad \mathcal{H}(\theta, \mathbf{r}) = \mathcal{U}(\theta) + \mathcal{K}(\mathbf{r})$$

$$\mathcal{U}(\theta) = -\log L(\theta) - \log p(\theta) \quad \text{and } \mathcal{K}(\mathbf{r}) = \frac{1}{2}\mathbf{r}^T\mathbf{M}^{-1}\mathbf{r}.$$

- Proposal for $(\mathbf{r}, \theta)$ is generated by drawing $\mathbf{r} \sim N(0, \mathbf{M}^{-1})$ and follow the **Hamiltonian dynamics**:

$$\frac{d\theta}{dt} = \frac{\partial \mathcal{H}(\theta, \mathbf{r})}{\partial \mathbf{r}} = \mathbf{M}^{-1}\mathbf{r}$$
$$\frac{d\mathbf{r}}{dt} = -\frac{\partial \mathcal{H}(\theta, \mathbf{r})}{\partial \theta} = -\frac{\partial \mathcal{U}(\theta)}{\partial \theta}$$

- **Properties** of **Hamiltonian dynamics**:
  1. **Reversible** (one-to-one) [leaves target distribution invariant]
  2. **Volume preserving** [Jacobian is one]
  3. **Energy conserving** [MH acceptance probability is 1]

Hamiltonian Monte Carlo (HMC) - discrete time

- **Leap-frog discretization** of the Hamiltonian dynamics.
- Take L **steps** of **size** ϵ.
- MH acceptance probabilty is no longer 1.
- Problem: each of the L steps require computing posterior gradient $\frac{\partial \log \pi(\theta)}{\partial \theta}$ for each proposal draw. **Costly**!
- Naive solution: simulate the Hamiltonian dynamics on a **subset of the data**.
- Problem: **Energy is not conserved**. MH acceptance probability quickly drops with dimension.
- Our solution: use ideas from MCMC with estimates of the target posterior.

- **Likelihood estimator**, $\hat{L}_m(\theta, \mathbf{u})$, based on m auxilliary random variables $\mathbf{u} = (u_1, ..., u_m)$. **Unbiased**:

$$\mathbb{E}_u\left[\hat{L}_m(\theta, \mathbf{u})\right] = L(\theta).$$

- **Subsampling**:
  - $\mathbf{u} = (u_1, ..., u_m)$ index sampled observations
  - unbiased **log-likelihood estimator** $[\ell(\theta) \equiv \log L(\theta)]$

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

  - **bias-correction**
  - **control variates** to reduce variance
- For (much) more info, see our papers:
  - Speeding Up MCMC by Efficient Data Subsampling, *JASA*.[1]
  - Subsampling MCMC - an Intro for the Survey Statistician, *Sankhya A*.

Pseudo-marginal MH

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

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

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

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

- Targets a joint distribution $\bar{\pi}_m(\theta, \mathbf{u})$ with marginal $\pi(\theta)$ [2]
- .. but **low** $\mathbb{V}\left(\hat{L}\left(\theta, \mathbf{u}\right)\right)$ crucial for **efficient sampling**.

HMC with Energy Conserving Subsampling

- Pseudo-marginal **extended target**

$$\bar{\pi}_m(\theta, \mathbf{r}, \mathbf{u}) \propto \exp\left(-\hat{\mathcal{H}}(\theta, \mathbf{r})\right) p_U(\mathbf{u}), \quad \hat{\mathcal{H}}(\theta, \mathbf{r}) = \hat{\mathcal{U}}(\theta) + \mathcal{K}(\mathbf{r})$$

$$\hat{\mathcal{U}}(\theta) = -\log \hat{L}(\theta) - \log p(\theta) \quad \text{and } \mathcal{K}(\mathbf{r}) = \frac{1}{2}\mathbf{r}^T \mathbf{M}^{-1}\mathbf{r}.$$

- Marginal of θ is still $\pi(\theta)$, if $\hat{L}(\theta)$ is unbiased.
- **HMC-within-Gibbs**:
  1. $\theta, \mathbf{r} \,|\mathbf{u}$ - HMC with energy $\hat{\mathcal{H}}$ given the subsample $\mathbf{u}$
  2. $\mathbf{u}|\theta, \mathbf{r}$ - Metropolis-Hastings update of subsample
- **Crucial**: **the same $\hat{L}(\theta)$** is used in:
  - The Hamiltonian dynamics and
  - The MH acceptance probability
- **HMC-ECS**. Subsampling conserves energy.
- **Distant proposals** with **high MH acceptance probability**.

The Block-Poisson estimator

- The **Block-Poisson estimator** of the likelihood $L(\theta)$:
  - Draw $\mathcal{X}_1, ..., \mathcal{X}_\lambda \overset{iid}{\sim} \mathrm{Pois}(1)$.
  - For $l = 1, ..., \lambda$, draw $\mathcal{X}_l$ mini-batches of data of size m.
  - Compute unbiased mini-batch estimators

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

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

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

- What really matters for MH is the variance of

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

- Product form of $\hat{L}_B(\theta)$: can use **Block Pseudo Marginal** (**BPM**).

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 $\theta \in \Theta$.
  Proof parts:
  - $\mathbb{E}\left(\hat{L}_B(\theta)\right) = \mathbb{E}_{\mathcal{X}_{1:\lambda}} \mathbb{E}_{\mathbf{u}|\mathcal{X}_{1:\lambda}}\left(\hat{L}_B(\theta)\right)$
  - Poisson expectation becomes a power series in ℓ.
  - Taylor series expansion of $L(\theta) = \exp(\ell(\theta))$.
- **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 $a = \ell - \lambda$.

Signed HMC-ECS

- 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:** $\Pr(\hat{\ell}_m^{(h,l)} \geq a)$ close to one. More efficient, but $\hat{L}_B(\theta) < 0$ possible.
- **Signed HMC-ECS** [3]
  - **Run HMC-ECS on absolute value** $\left|\hat{L}_B(\theta)\right| p(\theta)$
  - **Correct for sign** $s = \mathrm{Sign}\left(\hat{L}_B(\theta)\right)$ 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 HMC-ECS

- Standard (No U-turn) HMC tuning of:
  - number of **leapfrog** L
  - **step size** ϵ
  - **mass matrix M**.

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

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

- Optimal λ and m **balances**

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

Optimal tuning of Signed HMC-ECS

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

- Need to assume the **distribution of** $\hat{\ell}_m^{(h,l)}$
  - Normal (or CLT)
  - Mixture of Normals (universal approximator)

- The **derivation of IF** is an extension of the theory in [4] to blocked signed PMMH.

- **Guidelines** based on **idealized assumptions**. But accurate in experiments.

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

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

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

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

- $\hat{L}_B(\theta) > 0$ whenever an even number of ξ_l are negative.
- $\xi_l > 0$ whenever even number of $\frac{\hat{\ell}_m^{(h,l)} - \ell}{\lambda} + 1$ are negative.
- Applying a result from Feller's first book twice:

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

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

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

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

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

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

- $\hat{\ell}_m^{(h,l)} \sim \mathrm{Normal} \Rightarrow \sigma^2_{\log|\hat{L}_B|}(\lambda, m)$ 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 [5]
- Moments of log central χ^2 are known from [6]
- Law of total variance

Optimal tuning - normal case

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

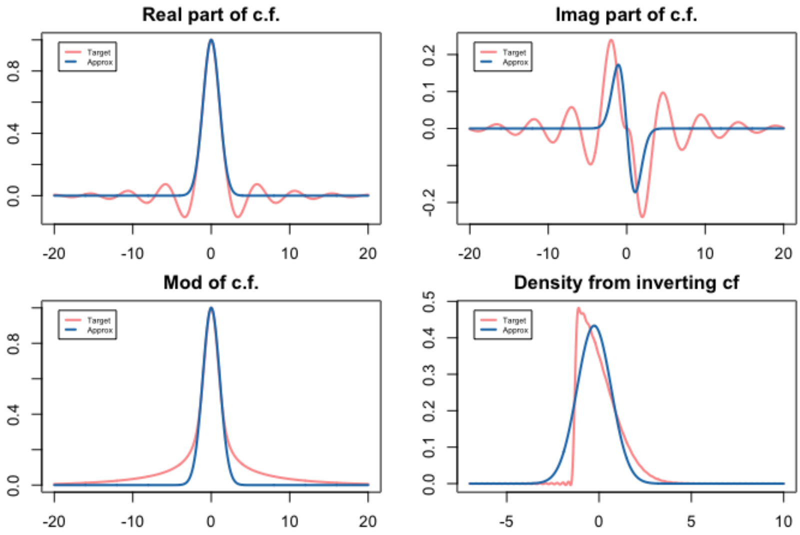
$$\mathbb{V}(\hat{\ell}_m^{(h,l)}(\theta)) = \frac{n^2}{m}\sigma^2_{\ell_i}(\theta)$$

- 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.
- However, numerical experiments tell us that $m = 1$ is optimal.

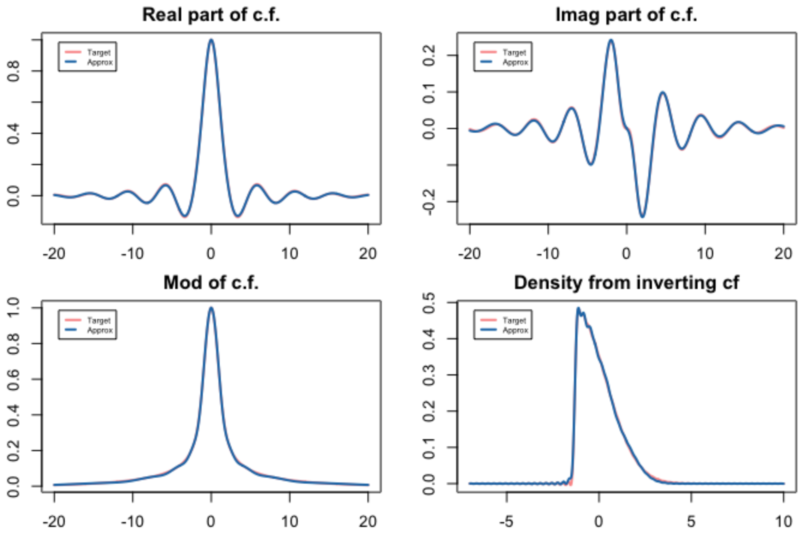
- Assume that $\hat{\ell}_m^{(h,l)}$ follows a **mixture of normals.**
- Mixture of normals are **universal approximators**.
- Both $\Pr(\hat{L}_B \geq 0)$ and $\sigma^2_{\log|\hat{L}_B|}$ are still **tractable**.
- ... but estimating $\sigma^2_{\ell_i}(\theta)$ is not enough anymore.
- How to fit a 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

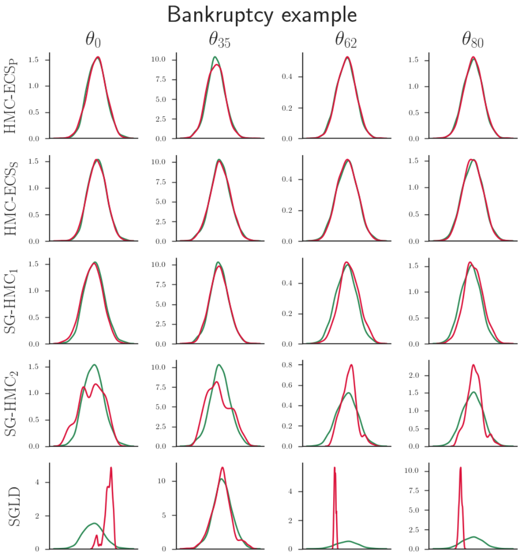


- Performance measure:

$$RCT = \frac{\text{CT of algorithm}}{\text{CT of Perturbed HMC-ECS}}$$

RCT	HMC	HMC-ECS$_S$	SG-HMC$_1$	SG-HMC$_2$	SGLD
min	358.5	1	7.0	48.6	53.9
median	466.8	2.2	9.5	100.2	230.1
max	683.6	7.2	538.7	246.4	2784.2

Posterior perturbation (bias)



- **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.
- Can be used to optimally tune Signed HMC in **doubly intractable problems**.

References

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.

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

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.

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.