# Stochastic Gradient MCMC: the Stochastic Barker Proposal

Lorenzo Mauri

Advisor: Prof. Giacomo Zanella

MSc in Data Science and Business Analytics
Bocconi University

October 16, 2021

# Main Contribution

- We introduce the **Stochastic Barker Proposal** algorithm by applying the stochastic gradient Markov Chain Monte Carlo framework to the Barker Proposal algorithm.
- We explore its performance through numerical studies.

# Outline

1. Introduction

2. Stochastic Gradient MCMC

3. The Barker Proposal

4. Simulation Study

5. Conclusion

# Introduction

Markov Chain Monte Carlo (MCMC) algorithms are used to approximate intractable integrals in Bayesian inference.

Gradient based MCMC:

- explore the space more efficiently;
- can be computationally expensive.

## Stochastic gradient MCMC

Stochastic gradient MCMC have recently gained great attention as they combine scalability to big size datasets with an efficient exploration of the space.

# Analytical setting

Let us consider a target distribution of the form

$$\pi(\theta) \propto f(\theta) \prod_{i=1}^{N} f(y_i|x_i,\theta) = \exp\left(-U(\theta)\right) \quad \theta \in \mathbb{R}^d \tag{1}$$

where $f(\theta)$ is the prior distribution and $f(y_i|x_i,\theta)$ is the likelihood of the $i$-th observation.

# Gradient based MCMC

Gradient based MCMC are derived from the discretization of a stochastic differential equation (SDE) that preserves $\pi$, as the Langevin diffusion[1]:

$$d\theta(t) = -\frac{1}{2}\nabla U(\theta(t)) + dB_t \tag{2}$$

where $dB_t$ is a $d-$dimensional Brownian motion.

- The one step Euler discretization is used as proposal distribution.
- The finite time discretization bias can be corrected with a Metropolis Hastings accept reject step.

[1] Roberts and Tweedie, 1996

# Stochastic Gradient MCMC

- The Potential function, $U$, can be written as the sum of $N$ data points components:

$$U(\theta) = \sum_{i=1}^{N} U_i(\theta) \tag{3}$$

where $U_i(\theta) = -\frac{1}{N}\log(f(\theta)) - \log(f(y_i|x_i,\theta))$.

→ computing the gradient results in a $\Theta(N)$ cost per iteration.

## Stochastic Gradient MCMC

Stochastic Gradient MCMC replace the gradient with a computationally cheaper unbiased estimate:

$$\hat{\nabla} U(\theta)^{(n)} = \frac{N}{n} \sum_{i \in \mathcal{S}_n} \nabla U_i(\theta) \tag{4}$$

where $n << N$ and $\mathcal{S}_n$ is a subsample with size $n$.

# Stochastic Gradient Langevin Dynamics Algorithm

- Sampling from the discretization of (2) after replacing the gradient with (4) gives the stochastic gradient Langevin dynamics (SGLD) algorithm[2].

**Algorithm 1:** Stochastic Gradient Langevin Dynamics

```
Input: $\theta^{(0)}, \{h_1, \dots, h_K\}$
for $t$ =1,..., $T$ do
    Draw $\mathcal{S}_n \subset \{1, \dots, N\}$;
    Estimate $\hat{\nabla} U(\theta^{(t-1)})^{(n)}$ using (4);
    Draw $\epsilon_k \sim N(0, h_k I)$;
    Update $\theta^{(t)} \leftarrow \theta^{(t-1)} - \frac{h_k}{2} \hat{\nabla} U(\theta^{(t-1)})^{(n)} + \epsilon_k$;
end
```

[2] Welling and Teh, 2011

# Stochastic Gradient Langevin Dynamics Algorithm

- By appropriately reducing the step size, the algorithm is consistent and a central limit theorem exits[3].
- When the step-size is kept fixed, SGLD invariant distribution might not coincide with the target distribution.

### Theorem (Brosse et al., 2018)

In the case of a Bayesian linear regression, under proper assumptions,

$$W_2^2(\pi_{SGLD}, \pi) = \Omega(1) \tag{5}$$

where $W_2^2(\mu, \nu)$ is the Wasserstein distance of order 2 between $\mu$ and $\nu$ and $\pi_{SGLD}$ is the invariant distribution of SGLD.

- It is essential to develop algorithms that are robust to the stochastic gradient noise and easy to tune.

[3] Teh et al., 2016

# The Barker Proposal (I)

The Barker Proposal[4] is a novel MCMC that outperforms other gradient based algorithms in terms of robustness to hyperparameter tuning.

- It uses the jump kernel of a Markov jump process that approximately preserves $\pi$ as proposal distribution.
- The gradient is used to inject skewness in the proposal.

[4] Livingstone and Zanella, 2020

# The Barker Proposal (II)

**Algorithm 2:** Barker Proposal in $\mathbb{R}$

```
Input: $\theta^{(0)}, \sigma$
for $t$ =1,..., $T$ do
    Draw $z \sim \mu_\sigma(\cdot)$;
    Define $p(\theta^{(t-1)}, z) = \frac{1}{1+\exp\left(-z\nabla\log\pi(\theta^{(t-1)})\right)}$;
    Set $b = 1$ with probability $p(\theta^{(t-1)}, z)$ otherwise set $b = -1$;
    Propose $\theta' \leftarrow \theta^{(t-1)} + z \times b$;
end
```

- If the signs of $z$ and $\nabla \log \pi(\theta)$ agree, then $p(\theta, z) > \frac{1}{2}$ (as $z\nabla \log \pi(\theta) \uparrow \infty$, $p(\theta, z) \uparrow 1$).

# The Barker Proposal (III)

- The resulting proposal is a skew-symmetric distribution[5].

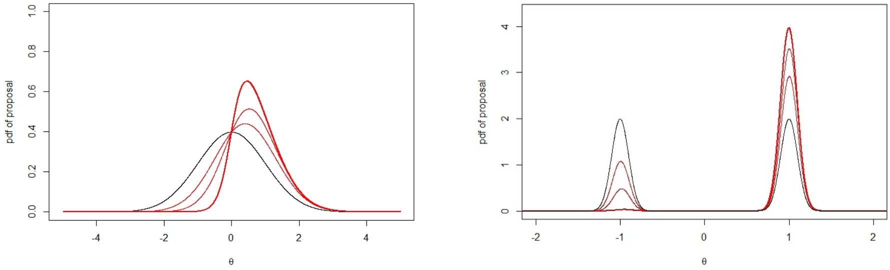

**Fig. 1.** Barker base kernels $\mu_\sigma$ (black), a zero mean Normal distribution (left) and a mixture of two Normal distributions (right), and Barker proposal distributions for increasing values of the gradient (red).

[5] Livingstone and Zanella, 2020

# The Stochastic Barker Proposal (I)

We develop the Stochastic Barker Proposal by substituting the gradient with (4) inside the Barker Proposal. The resulting algorithm:

- enjoys the scalability to large datasets of stochastic gradient MCMC;
- inherits the robustness to hyperparameter tuning from the Barker Proposal.

# The Stochastic Barker Proposal (II)

**Algorithm 3:** The Stochastic Barker Proposal in $\mathbb{R}$

```
Input: $\theta^{(0)}, \sigma$
for $t$ =1,..., T do
    Draw $z \sim \mu_\sigma(\cdot)$;
    Draw $\mathcal{S}_n \subset \{1, \ldots, N\}$;
    Estimate $\hat{\nabla} U(\theta^{(t-1)})^{(n)}$ using (4);
    Define $\hat{p}(\theta^{(t-1)}, z) = \frac{1}{1+\exp\left(z \hat{\nabla} U(\theta)^{(n)}\right)}$;
    Set $b = 1$ with probability $p(\theta^{(t-1)}, z)$ otherwise set $b = -1$;
    Update $\theta^{(t)} \leftarrow \theta^{(t-1)} + z \times b$;
end
```

# Simulations

The Stochastic Barker Proposal is compared to SGLD with respect to the following characteristics:

- the trade-off between mixing and sampling accuracy (measured with mean and variance bias) when hyperparameters are varied;
- predictive accuracy on unseen data.

# Toy Example: Bayesian Normal Model

- We apply the Bayesian Normal model to simulated data.

**Main Result**

With a regular target distribution, SGLD outperforms the SBP in terms of mixing-accuracy trade-off.

# Bayesian Logistic Regression (I)

## Bayesian Logistic Regression Model

Consider the Bayesian Logistic Regression model:

$$\begin{gathered} y_i|\mathbf{x}_i, \theta \sim \textit{Bernoulli}\left(\frac{1}{1 - \exp\left(-\theta^t \mathbf{x}_i\right)}\right) \quad i = 1, \dots, N \\ \theta \sim N_d(0, \tau^2 I). \end{gathered} \tag{6}$$

- We apply (6) to the Arrhythmia dataset[6], keeping the first 100 covariates.
- Ground truth posterior quantities are obtained with the STAN[7] implementation of the No-U-Turn sampler[8].
- We evaluate how sampling accuracy decreases as mixing is increased keeping the number of iterations fixed.

[6] https://archive.ics.uci.edu/ml/datasets/Arrhythmia

[7] Carpenter et al., 2017

[8] Homan and Gelman, 2014

# Bayesian Logistic Regression: Mixing vs Accuracy

## Main Result

The SBP with bimodal noise is more robust to hyperparameter tuning.

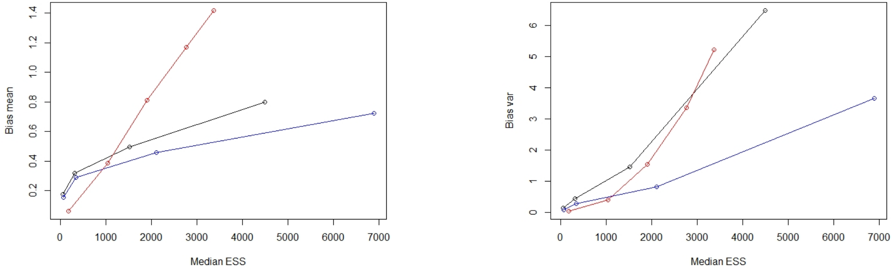

**Fig. 2.** Bias of mean (left) and variance (right) vs median ESS. Red refers to SGLD, black to the SBP with Normal noise and blue to the SBP with bimodal noise.

# Bayesian Probabilistic Matrix Factorization (I)

- To study the predictive performance of the algorithms, we consider a content recommendation task, which is addressed with a probabilistic matrix factorization technique.
- We apply the Bayesian Probabilistic Matrix Factorization model[9] to the MovieLens dataset[10].
- We study the predictive accuracy in terms of root mean squared error (RMSE) on a test set.

[9] Salakhutdinov and Mnih (2008)

[10] https://files.grouplens.org/datasets/movielens/ml-100k/

# Bayesian Probabilistic Matrix Factorization (II)

## Bayesian Probabilistic Matrix Factorization Model

Let us consider a ratings matrix $\mathbf{R} \in \mathbb{R}^{N\times M}$, we introduce two lower dimensional matrices $\mathbf{U} \in \mathbb{R}^{N\times d}, \mathbf{V} \in \mathbb{R}^{M\times d}$.
The Bayesian Probabilistic Matrix Factorization (BPMF) model is defined as follows:

$$\mathbf{R}_{ij}|\mathbf{U}, \mathbf{V}, \alpha \sim N(\mathbf{U}_i^t \mathbf{V}_j, \alpha^{-1}) \text{ for all } i, j \text{ if } I_{ij} = 1 \tag{7}$$

where $I_{ij}$ is the indicator function which is equal to $1$ if user $i$ rated item $j$. Columns of $\mathbf{U}$ and $\mathbf{V}$ are assigned a Normal distribution with common mean $\mu_{\mathbf{U}}$ and $\mu_{\mathbf{V}}$. $\mu_{\mathbf{U}}$ and $\mu_{\mathbf{V}}$ are assigned a Normal prior. Precisions are assigned a Gamma prior.

# BPMF: Predictive Accuracy

## Main Result

The SBP produces more accurate estimates on unseen data.

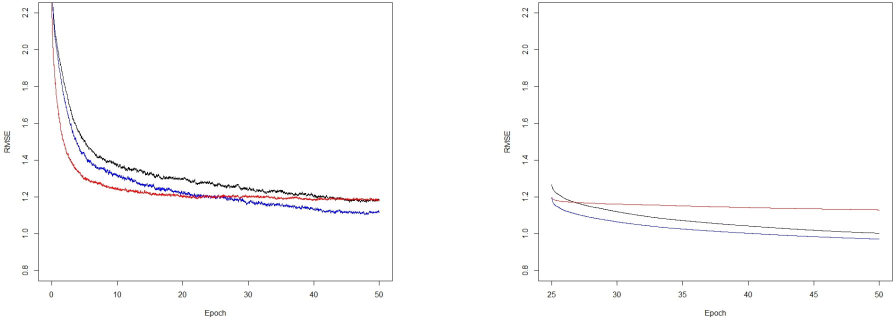

**Fig. 3.** BPMF RMSE with mini-batch size=800, the root mean squared error produced by each sample predictions (left) and by the MCMC estimates (right). Red refers to SGLD, black to the SBP with Normal noise and blue to the SBP with bimodal noise.

# Conclusion

- The SBP seems to outperform SGLD in terms of robustness to hyperpameter tuning with irregular posterior distributions.
- The SBP shows good predictive accuracy on unseen data.
- Interesting extensions include understanding the theoretical properties of the SBP, implementing variance reduction techniques, adding momentum to the dynamics and developing the stochastic optimization variant of the SBP.

# References I

Brosse, Nicolas, Alain Durmus, and Eric Moulines (2018). “The promises and pitfalls of Stochastic Gradient Langevin Dynamics”. In: *Advances in Neural Information Processing Systems* 31.

Carpenter, Bob et al. (2017). “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software* 76.1.

Homan, Matthew D. and Andrew Gelman (2014). “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. In: *J. Mach. Learn. Res.* 15.1, 1593–1623.

Livingstone, Samuel and Giacomo Zanella (2020). *The Barker proposal: combining robustness and efficiency in gradient-based MCMC*. arXiv: `1908.11812 [stat.CO]`.

Roberts, Gareth O. and Richard L. Tweedie (1996). “Exponential convergence of Langevin distributions and their discrete approximations”. In: *Bernoulli* 2.4, pp. 341 –363.

# References II

Salakhutdinov, R. and A. Mnih (2008). “Bayesian probabilistic matrix factorization using Markov chain Monte Carlo”. In: *ICML '08.*

Teh, Yee Whye, Alexandre H. Thiery, and Sebastian J. Vollmer (2016). “Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics”. In: *J. Mach. Learn. Res.* 17.1, 193–225.

Welling, Max and Yee Teh (2011). “Bayesian Learning via Stochastic Gradient Langevin Dynamics”. In: *ICML'11*, pp. 681–688.