Stochastic Gradient MCMC for Nonlinear State Space Models

Srshti Putcha[1] Christopher Aicher[3] Christopher Nemeth[2] Paul Fearnhead[2] Emily Fox[3,4]

[1]STOR-i Centre for Doctoral Training, Lancaster University [2]Department of Mathematics and Statistics, Lancaster University
[3]Department of Statistics, University of Washington [4]Paul G. Allen School of Computer Science and Engineering, University of Washington

Problem Motivation

The aim of our work [1] is to conduct scalable Bayesian inference for long sequences of time series data. To do this, we can make use of **state space models** (SSMs).

SSMs provide a flexible framework for capturing the behaviour of complex time series via a latent stochastic process. Nonlinear, non-Gaussian SSMs are widely applied in many scientific domains for modelling time series data and sequential data. These applications include:

- **engineering** *e.g. target tracking,*
- **epidemiology** *e.g. compartmental disease models,* and
- **financial data**, *e.g stochastic volatility models.*

State Space Modelling

State space models are a type of discrete-time bivariate stochastic process, consisting of:

- a **hidden** or **latent** process, $\{X_t \in \mathbb{R}^{d_x}\}_{t=1}^T$, and
- a second **observed** process, $\{Y_t \in \mathbb{R}^{d_y}\}_{t=1}^T$.

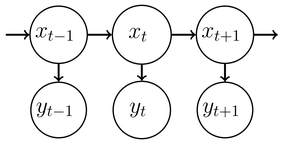

Figure 1. Graphical representation of an SSM.

Given the initial prior $X_0 \sim \nu(x_0|\theta)$ and parameters $\theta \in \Theta$, the generative model is,

$$X_t \,|\, (X_{t-1} = x_{t-1},\, \theta) \sim p(x_t \,|\, x_{t-1}, \theta), \qquad (1)$$
$$Y_t \,|\, (X_t = x_t,\, \theta) \sim p(y_t \,|\, x_t, \theta).$$

Fisher's Identity

A quantity of interest for inference is the **score function**, $\nabla_\theta \log p(y_{1:T}|\theta)$, as it can be used to maximise the log-likelihood.

If the **latent state posterior** $p(x_{1:T}|y_{1:T}, \theta)$ can be written analytically, we can calculate the score exactly using the following identity [2]:

$$\nabla_\theta \log p(y_{1:T}|\theta) = \mathbb{E}_{X|Y,\theta}[\nabla_\theta \log p(X_{1:T}, y_{1:T}|\theta)] = \sum_{t=1}^{T} \mathbb{E}_{X|Y,\theta}[\nabla_\theta \log p(X_t, y_t|x_{t-1}, \theta)].$$

However, if $p(x_{1:T}|y_{1:T}, \theta)$ is not available in closed-form, we need to approximate expectations with respect to the latent state posterior with **particle filtering** methods.

Sequential Importance Resampling (SIR)

Particle filtering algorithms can be used to create empirical approximations of the expectation of $H(X_{1:T})$ with respect to the posterior $p(x_{1:T}|y_{1:T}, \theta)$.

This is done by generating a cloud of N **particles**, $\{x_t^{(i)}\}_{i=1}^N$ and calculating their associated **importance weights**, $\{w_t^{(i)}\}_{i=1}^N$, recursively. We can update the particles and weights using SIR as follows.

1. **Resample** ancestor indices $\{a_1, \ldots, a_N\}$, i.e. $a_i \sim \text{Categorical}(w_{t-1}^{(i)})$.
2. **Propagate** $x_t^{(i)} \sim q(\cdot|x_{t-1}^{(a_i)}, y_t, \theta)$, using a proposal distribution $q(\cdot|\cdot)$.
3. **Update** and normalize weights such that

$$w_t^{(i)} \propto \frac{p(y_t|x_t^{(i)}, \theta)p(x_t^{(i)}|x_{t-1}^{(a_i)}, \theta)}{q(x_t^{(i)}|x_{t-1}^{(a_i)}, y_t, \theta)}, \quad \sum_i w_t^{(i)} = 1.$$

When our target function decomposes into a pairwise sum $H(x_{1:T}) = \sum_{t=1}^T h_t(x_t, x_{t-1})$, then we only need to keep track of the partial sum $H_t = \sum_{s=1}^t h_s(x_s, x_{s-1})$ during SIR. In particular, we know that $h_t = \nabla_\theta \log p(y_t, x_t|x_{t-1}, \theta)$ for the Fisher's identity.

Stochastic Gradient MCMC

One popular method to conduct scalable Bayesian inference for large data sets is **stochastic gradient Markov chain Monte Carlo** (SGMCMC).

To draw a sample θ from the posterior $p(\theta \,|\, y) \;\propto\; p(y \,|\, \theta)p(\theta)$, typical gradient-based MCMC algorithms simulate a stochastic differential equation based on the gradient of the log-likelihood, $g_\theta = \nabla_\theta \log p(y \,|\, \theta)$. SGMCMC methods replace g_θ with an unbiased stochastic version, $\widehat{g}_\theta$, using subsamples of the data to avoid costly computations.

A fundamental SGMCMC method is the **stochastic gradient Langevin dynamics** (SGLD) algorithm [3],

$$\theta^{(k+1)} \leftarrow \theta^{(k)} + \epsilon^{(k)} \cdot \left(\widehat{g}_\theta + \nabla \log p(\theta)\right) + \mathcal{N}\left(0, 2\epsilon^{(k)}\right), \qquad (2)$$

where $\{\epsilon^{(k)}\}$ is a decreasing step-size schedule.

To apply SGMCMC to SSMs, we propose to tackle the temporal dependence between observations using a **buffered stochastic gradient**. For a contiguous subsequence of length S, $\mathcal{S} = \{s+1, \ldots, s+S\}$, the modified gradient is given by,

$$\widehat{g}_\theta(S, B) = \sum_{t \in \mathcal{S}} \frac{\mathbb{E}_{x|y_{\mathcal{S}^*},\theta}[\nabla_\theta \log p(X_t, y_t|X_{t-1}, \theta)]}{\Pr(t \in \mathcal{S})}, \qquad (3)$$

where B is the buffer length and $\mathcal{S}^* = \{s+1-B, \ldots, s+S+B\}$. If the SSM and its gradient satisfy a Lipschitz condition, the bias decays geometrically in B.

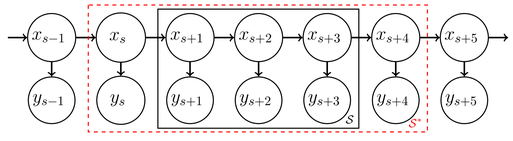

Figure 2. Graphical model of $\mathcal{S}^*$ with $S = 3$ and $B = 2$.

With the help of Fisher's identity, we construct a particle approximation of (3), $g_\theta^{PF}(S, B, N)$, that is suitable for nonlinear SSMs.

SGLD on Synthetic Data

We first fitted SGLD to synthetic Linear Gaussian SSM (LGSSM) data, where gradients can be calculated exactly using the Kalman Filter. The LGSSM is given by,

$$X_t \,|\, (X_{t-1} = x_{t-1},\, \theta) \sim \mathcal{N}(x_t \,|\, \phi x_{t-1}\,,\, \sigma^2), \qquad (4)$$
$$Y_t \,|\, (X_t = x_t,\, \theta) \sim \mathcal{N}(y_t \,|\, x_t\,,\, \tau^2).$$

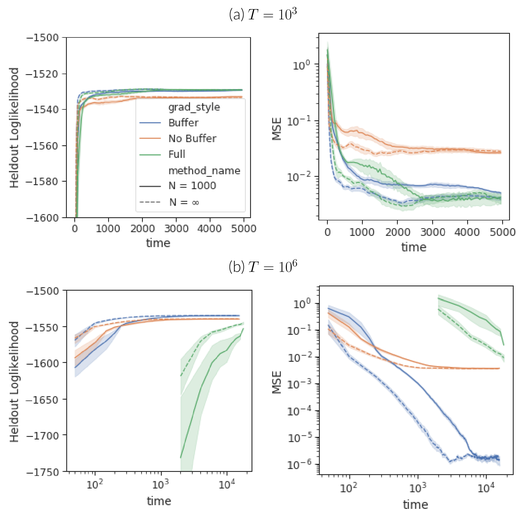

Figure 3. SGLD on synthetic LGSSM data: (top) $T = 10^3$, (bottom) $T = 10^6$; (left) heldout log-likelihood, (right) MSE of estimated posterior mean to true $\phi = 0.9$.

SGLD on Exchange Rate Log Returns

We then fitted the Stochastic Volatility Model (SVM),

$$X_t \,|\, (X_{t-1} = x_{t-1},\, \theta) \sim \mathcal{N}(x_t \,|\, \phi x_{t-1}\,,\, \sigma^2), \qquad (5)$$
$$Y_t \,|\, (X_t = x_t,\, \theta) \sim \mathcal{N}(y_t \,|\, 0\,,\, \exp(x_t)\tau^2).$$

to EUR-US exchange rate data obtained at the minute resolution between November 2017 and October 2018. The data consists of 350,000 observations of demeaned log-returns and was broken into 53 weekly segments of ∼7,000 observations each.

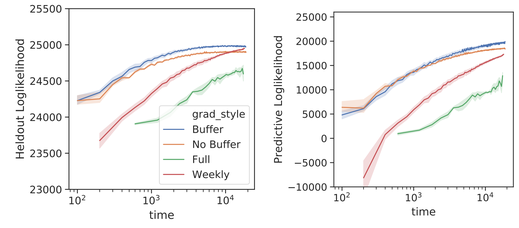

Figure 4. SGLD plots on exchange rate data: (left) heldout log-likelihood, (right) 3-step ahead predictive log-likelihood.

Buffered Stochastic Gradient Estimate Error

Overall, there are three main factors that control the stochastic gradient error:

- the number of particles, N,
- the subsequence length, S, and
- the buffer length, B.

Below, we investigate the impact that changing the buffer length, B, has on the stochastic gradient error for a fixed subsequence length of $S = 16$.

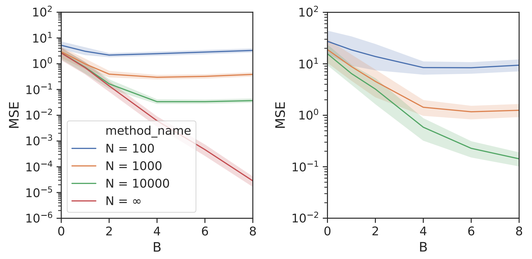

Figure 5. Buffered stochastic gradient estimate error plots: (left) LGSSM ϕ, (right) SVM ϕ

References

[1] Christopher Aicher, Srshti Putcha, Christopher Nemeth, Paul Fearnhead, and Emily B Fox. Stochastic gradient MCMC for Nonlinear State Space Models. *arXiv preprint arXiv:1901.10568*, 2019.

[2] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[3] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Contact

Website: **www.lancs.ac.uk/∼putchas** Email: **s.putcha1@lancaster.ac.uk**