Stochastic Bouncy Particle Sampler

Ari Pakman[1*] Dar Gilboa[1*] David Carlson[2] Liam Paninski[1]

[1]Dep. of Statistics and Grossman Center for the Statistics of Mind, Columbia U.

[2]Depts. of Civil and Environmental Engineering, and Biostatistics and Bioinformatics, Duke U.

[*] Equal contribution.

Abstract

We introduce a stochastic version of the non-reversible, rejection-free Bouncy Particle Sampler (BPS), a Markov process whose sample trajectories are piecewise linear, to efficiently sample Bayesian posteriors in big datasets. We prove that in the BPS no bias is introduced by noisy evaluations of the log-likelihood gradient. On the other hand, we argue that efficiency considerations favor a small, controllable bias, in exchange for faster mixing. We introduce a simple method that controls this trade-off. We illustrate these ideas in several examples which outperform previous approaches.

Bouncy Particle Sampler [Bouchard-Côté et.al., 2015]

Let $p(\mathbf{w}) \propto e^{-U(\mathbf{w})}$ be the distribution we want to sample from.

BPS Algorithm

- Initialize $\mathbf{w}_0 \in \mathbb{R}^D$ and velocity $\mathbf{v} \in \mathbb{S}^{D-1}$
- While desired
  - Sample Poisson events t_r, t_b with rates λ_r and $\lambda(t) = [\mathbf{v} \cdot \nabla U(\mathbf{w}_{i-1} + \mathbf{v}t)]_+$
  - Let $t_i = \min(t_b, t_r)$ and move $\mathbf{w}_i = \mathbf{w}_{i-1} + \mathbf{v}t_i$,
  - If $t_b < t_r$, reflect $\mathbf{v} \leftarrow \mathbf{v} - 2\frac{(\mathbf{v}\cdot\nabla U(\mathbf{w}_i))\nabla U(\mathbf{w}_i)}{||\nabla U(\mathbf{w}_i)||^2}$ else, refresh: $\mathbf{v} \sim p(\mathbf{v}) = \text{Unif}[\mathbb{S}^{D-1}]$
- Return piecewise linear trajectory $\{\mathbf{w}_i, t_i\}$

Key Properties

- Preserves $p(\mathbf{w})$ and is non-reversible.
- Empirically faster than HMC when Poisson events can be sampled easily.
- Bounces only when $\mathbf{v}$ points downhill in $U(\mathbf{w})$.

Noise Invariance

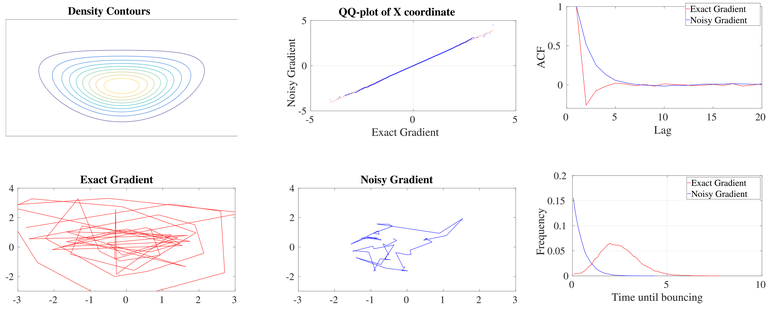
In Bayesian posterior distributions, using a random subset of the data at each gradient evaluation can be represented as

$$\nabla\tilde{U}(\mathbf{w}) = \nabla U(\mathbf{w}) + \mathbf{n}_\mathbf{w}\,, \qquad \mathbf{n}_\mathbf{w} \sim p(\mathbf{n}_\mathbf{w}|\mathbf{w})\,, \tag{1}$$

where $\mathbf{n}_\mathbf{w} \in \mathbb{R}^D$ and $p(\mathbf{n}_\mathbf{w}|\mathbf{w})$ has zero mean.

Theorem: *The invariance of $p(\mathbf{w})p(\mathbf{v})$ under the BPS algorithm is unaffected by the noise (1) if $\mathbf{n}_{\mathbf{w}_1}$ and $\mathbf{n}_{\mathbf{w}_2}$ are independent for $\mathbf{w}_1 \neq \mathbf{w}_2$.*

Price of noise: slower mixing



Stochastic Bouncy Particle Sampler

Challenge in noisy BPS: How to sample events from a noisy Poisson intensity with unknown upper bound?

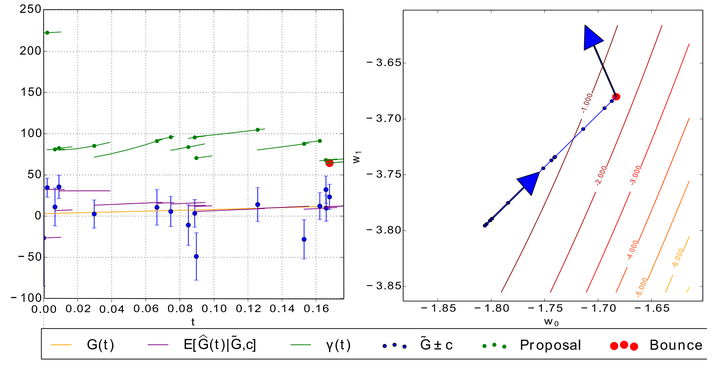
Our solution:

- Fit values of $\tilde{G}(t) = \mathbf{v} \cdot \nabla\tilde{U}(\mathbf{w} + t\mathbf{v})$ to a linear regression model
$$\hat{G}(t) = \hat{\beta}_1 t + \hat{\beta}_0 + \varepsilon_t$$
- Define upper bound $\lambda(t)$ for the noisy Poisson intensity.
$$\lambda(t) = [\hat{\beta}_1 t + \hat{\beta}_0 + k\rho(t)]_+$$
where $\rho^2(t) = (1\,t)\,\Sigma(1\,t)^T$, $k > 0$, and Σ estimates the covariance of $\hat{\beta}_0, \hat{\beta}_1$.
- Sample proposal event with intensity $\lambda(t)$ and accept with probability $\min(1, [\tilde{G}(t)]_+/\lambda(t))$ (thinning method).

Stochastic BPS Algorithm

- Initialize particle position $\mathbf{w}_0 \in \mathbb{R}^D$ and velocity $\mathbf{v} \in \mathbb{S}^{D-1}$
- While desired
  - Sample t_i from Poisson intensity $\lambda(t)$ and move $\mathbf{w}_i = \mathbf{w}_{i-1} + t_i\mathbf{v}$.
  - Make noisy observation $\tilde{G}(t_i) = \mathbf{v} \cdot \nabla\tilde{U}(\mathbf{w}_{i-1} + \mathbf{v}t_i)$.
  - With probability $\min(1, [\tilde{G}(t_i)]_+/\lambda(t_i))$, reflect velocity using noisy gradient $\nabla\tilde{U}$.
  - Else: use $\tilde{G}(t_i)$ to update regression parameters.
- Return piecewise linear trajectory $\{\mathbf{w}_i, t_i\}$.

Sampling bounce times in SBPS



Left panel: linear regression used to estimate $G(t) = \mathbf{v} \cdot \nabla U(\mathbf{w} + t\mathbf{v})$ from noisy observations. *Right panel:* corresponding piecewise linear trajectory of $\mathbf{w}$.

Advantages:

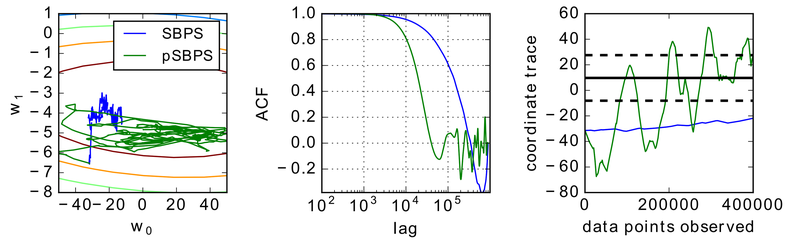
- Less sensitive to hyperparameter choice than SGLD.
- Mixing speed advantage from the use of a non-reversible kernel.
- Faster mixing compared to other stochastic gradient MCMC methods.
- Reflection with noisy gradient makes velocity refreshments empirically unnecessary.

Disadvantage:

- Small bias from events with $[\tilde{G}(t)]_+/\lambda(t) > 1$, controllable by adjusting value of k. The bias can be bound using Stein's method and mixing information in settings where the Laplace approximation of the target density holds.

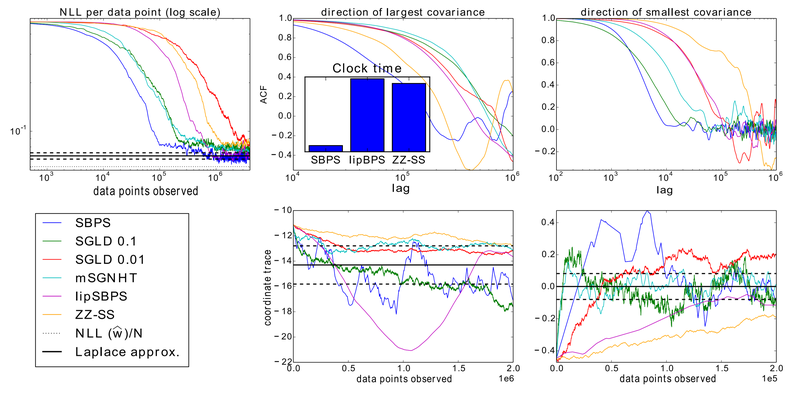
Preconditioned SBPS

- In distributions with non-homogeneous curvature, we can replace the intensity by $[G = \mathbf{v} \cdot \mathbf{A}\nabla U(\mathbf{w})]_+$ and velocity reflection by $\mathbf{v}_r = \mathbf{v} - 2\frac{(\mathbf{v}\cdot\mathbf{A}\nabla U(\mathbf{w}))\mathbf{A}\nabla U(\mathbf{w})}{||\mathbf{A}\nabla U(\mathbf{w})||^2}$.
- The preconditioner $\mathbf{A}$ is learned adaptively, such as a rank-1 approximation of the Hessian using gradient information only.

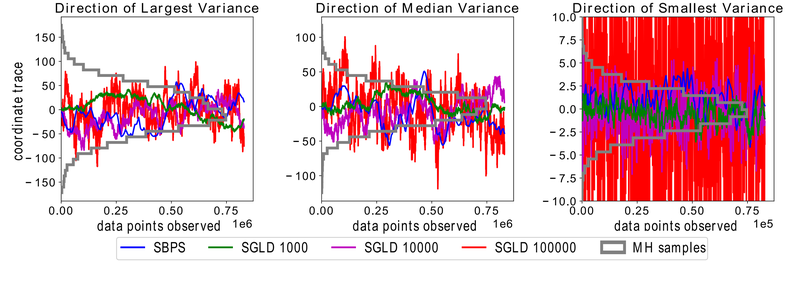


Experimental Results

Logistic regression



Multinomial regression on MNIST



Conclusion: Piecewise deterministic samplers are a novel and promising approach to MCMC sampling and Stochastic BPS is a state-of-the-art Big Data multi-purpose MCMC algorithm.