A Piecewise Deterministic Markov Process via (r, θ) swaps in hyperspherical coordinates

Alexander Terenin
Imperial College London

Joint work with Daniel Thorngren

MaxEnt '18
July 5th, 2018

http://avt.im/

arXiv:1807.00420

Imperial College London

Bayesian methods at scale: big data

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left[\prod_{i=1}^{N} f(y_i \mid \boldsymbol{\theta}) \right] \pi(\boldsymbol{\theta})$$

MCMC methods generally require evaluating the full data at every iteration $\rightarrow$ $\mathcal{O}(N)$

Scalable methods

Stochastic Gradient Descent $$\boldsymbol{\theta}' = \boldsymbol{\theta} + \varepsilon \nabla \ln f(\boldsymbol{\theta} \mid y_i)$$

Stochastic Variational Inference $$D_{\mathrm{KL}}(q \,||\, f) = \mathbb{E}_q \left[\ln \frac{q(\boldsymbol{\theta})}{f(\boldsymbol{\theta} \mid y_i)} \right]$$

Logs turn products into sums that are amenable to unbiased estimation $\rightarrow$ $\mathcal{O}(1)$

Introduce control variates $\rightarrow$ $\mathcal{O}(1)$ iterative incl. mixing

Piecewise Deterministic Markov Processes

Example: Bouncy Particle Sampler

- Deterministic Dynamics

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v} \qquad \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \mathbf{0}$$

- Switching Rate

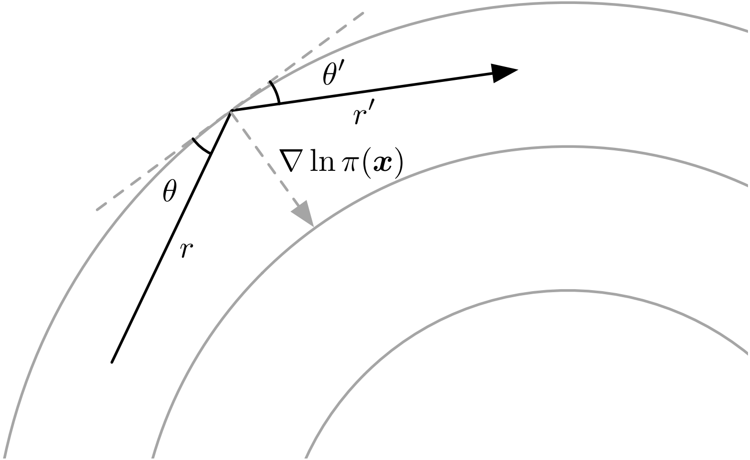
$$\lambda(\boldsymbol{x}, \boldsymbol{v}) = \max\{0, -\boldsymbol{v} \cdot \nabla \ln \pi(\boldsymbol{x})]\}$$

- Transition Function

$$F_{\boldsymbol{x}}(\boldsymbol{v}) = \boldsymbol{v} - 2\frac{\boldsymbol{v} \cdot \nabla \ln \pi(\boldsymbol{x})}{||\nabla \ln \pi(\boldsymbol{x})||^2} \nabla \ln \pi(\boldsymbol{x})$$

Invariant measure: $\pi(\boldsymbol{x}, \boldsymbol{v}) = \pi(\boldsymbol{x})\pi(\boldsymbol{v})$ with $\pi(\boldsymbol{v})$ Gaussian

Bouncy Particle Sampler



Does the transition function need to be linear?

- Deterministic Dynamics

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v} \qquad\qquad \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \boldsymbol{0}$$

- Switching Rate

$$\lambda(\boldsymbol{x}, \boldsymbol{v}) = \max\{0, -\boldsymbol{v} \cdot \nabla \ln \pi(\boldsymbol{x})]\}$$

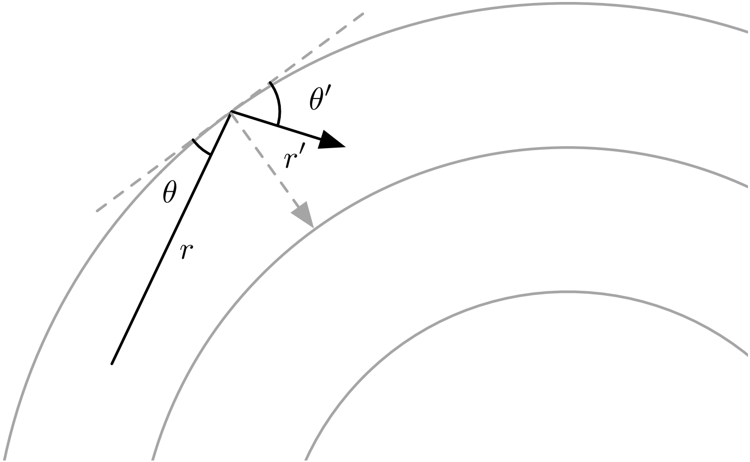
- Transition Function

$F_{\boldsymbol{x}}(\boldsymbol{v})$ s.t. $F_{\boldsymbol{x}}^{-1}(\boldsymbol{v})$ exists and $\pi(\boldsymbol{x}, \boldsymbol{v})$ is stationary

Does a non-BPS process satisfying these conditions exist?
Does it perform better with dimension?

Potential benefit of nonlinear transitions

Idea: swap magnitude for direction in hyperspherical coordinates



Solving the Fokker-Planck equation

$$\lambda(\boldsymbol{x},\boldsymbol{v})\pi(\boldsymbol{v}) - \lambda(\boldsymbol{x}, F_{\boldsymbol{x}}^{-1}(\boldsymbol{v}))\pi(F_{\boldsymbol{x}}^{-1}(\boldsymbol{v}) \mid \boldsymbol{x}) \left| \frac{\partial F_{\boldsymbol{x}}^{-1}(\boldsymbol{v})}{\partial \boldsymbol{v}} \right| = -v \cdot (\nabla \ln \pi(\boldsymbol{x}))\, \pi(\boldsymbol{v})$$

Solution

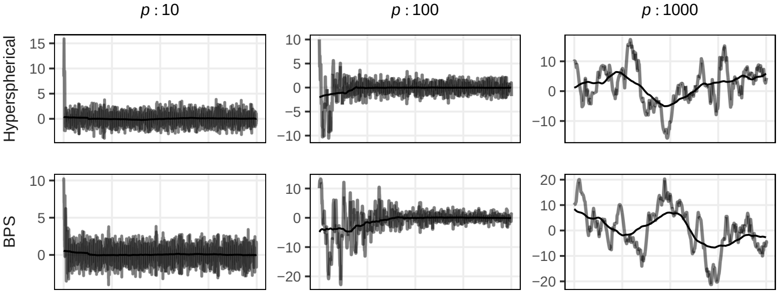
$$\frac{\mathrm{d}\theta'(r)}{\mathrm{d}r} = \frac{k r^p \exp\left\{\frac{r^2}{-2}\right\}}{\cos\left[\theta'(r)\right] \sin^{p-2}\left[\theta'(r)\right]}$$

For large p

$$\theta'(r) \approx \frac{\pi}{2} - \frac{1}{2}\sqrt{\frac{-8}{p-2} \ln \Phi\left[\sqrt{2}\,(r - \sqrt{p})\right]}$$

Does it work better?

Target: isotropic Gaussian



Breaks in sufficiently high dimension

Conclusions

Still not good enough in high dimension

- Nowhere near performance of SGD or SVI
- Need more than non-magnitude-preserving transitions

Fokker-Planck equation has many solutions

- Which one should be used?
- Are there other tractable transition functions?
- Are other switching rates possible?

A. Terenin, D. Thorngren. A Piecewise Deterministic Markov Process via (r, θ) swaps in hyperspherical coordinates. arXiv:1807.00420.