Hamiltonian ABC

Meeds, Leenders, Welling UAI 2015

Heiko

Gatsby MLJC

June 8, 2014

Hamiltonian Approximate Bayesian Computation

Antagonistic?

- ABC is intended for complex and mostly intractable likelihoods
- HMC requires a lot from the target: gradients and Hessians

Ideas

Motivation:

- Overcome random walk on ABC

High level:

- Construct (parametric) synthetic likelihood
- Stochastic gradients
- Hamiltonian dynamics

Computational tricks:

- Synthetic likelihood is Gaussian
- Stochastic finite differences for differentiation
- Variance reduction via sticky random numbers

ABC

- Bayesian posterior

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)\pi(\mathbf{y}|\theta)$$

  with $\mathbf{y} \in \mathbb{R}^J$ summary statistics of raw observations
- ABC: Likelihood is intractable
- Have simulator given for $\mathbf{x} \in \mathbb{R}^J$ given $\theta \in \mathbb{R}^D$
- Idea to estimate $\pi(\mathbf{y}|\theta)$
  - Simulate $\mathbf{x}^{(s)} \sim \pi(\mathbf{x}|\theta)$. In practice $\mathbf{x}^{(s)} = f(\theta, \omega)$ with seed ω
  - Compare to observed data $\mathbf{y}$ via an ϵ-kernel $\pi_\epsilon(\mathbf{y}|\mathbf{x})$

$$\pi_\epsilon(\mathbf{y}|\theta) = \int \pi_\epsilon(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\theta)d\mathbf{x} \approx \frac{1}{S}\sum_{s=1}^{S} \pi_\epsilon(\mathbf{y}|\mathbf{x}^{(s)})$$

  - Examples: ϵ-ball, Gaussian, etc.

ABC-MCMC

- Targets approximate posterior:

$$\pi_\epsilon(\theta|\mathbf{y}) \propto \pi(\theta)\pi_\epsilon(\mathbf{y}|\theta)$$

- Proposal: θ',$\mathbf{x}^{(1)'}, \ldots, \mathbf{x}^{(S)'}$ from

$$q(\theta'|\theta)\prod_s \pi(x^{(s)'}|\theta')$$

- Acceptance probability:

$$\min\left(\frac{\pi(\theta')}{\pi(\theta)} \times \frac{\frac{1}{S}\sum_{s=1}^{S}\pi_\epsilon(\mathbf{y}|\mathbf{x}^{(s)'})}{\frac{1}{S}\sum_{s=1}^{S}\pi_\epsilon(\mathbf{y}|\mathbf{x}^{(s)})} \times \frac{q(\theta|\theta')}{q(\theta'|\theta)}\right)$$

- Pseudo-Marginal MCMC, Marginal MCMC
- Under conditions: $\pi_\epsilon(\theta|\mathbf{y}) \to \pi(\theta|\mathbf{y})$ as $\epsilon \to 0$

Synthetic likelihoods

- Conditional model for $\pi(\mathbf{x}|\theta)$
- Can be Gaussian (Wood, 2010)

$$\pi(\mathbf{x}|\theta) = \mathcal{N}(\mathbf{x}|\mu_\theta, \sigma_\theta^2)$$

  with $\mu_\theta, \sigma_\theta^2$ estimated from $\mathbf{x}^{(s)} \sim \pi(\mathbf{x}|\theta)$
- Can also be KDE or GP (Meeds, Welling, 2014)
- If the ϵ-kernel and $\pi(\mathbf{x}|\theta)$ are Gaussian

$$\pi_\epsilon(\mathbf{y}|\theta) = \int \mathcal{N}(\mathbf{y}|\mathbf{x}, \epsilon^2)\mathcal{N}(\mathbf{x}|\mu_\theta, \sigma_\theta^2)d\mathbf{x}$$
$$= \mathcal{N}(\mathbf{y}|\mu_\theta, \sigma_\theta^2 + \epsilon^2)$$

- Paper claims: More robust to small ϵ
- Xian's Og: Doesn't make sense as ϵ is estimated from $\mathbf{x}^{(s)}$ too

Gradients?

- Recall model

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta)\pi(\mathbf{y}|\theta) \approx \pi(\theta)\pi_\epsilon(\mathbf{y}|\theta)$$

- Gradient-based posterior inference on θ needs $\nabla_\theta \pi_\epsilon(\mathbf{y}|\theta)$
- Here, that is

$$\nabla_\theta \mathcal{N}(\mathbf{y}|\mu_\theta, \sigma_\theta^2 + \epsilon^2)$$

  where e.g.

$$\mu_\theta = \frac{1}{S}\sum_{s=1}^{S} \mathbf{x}^{(s)} \qquad \text{and} \qquad \sigma_\theta^2 = \frac{1}{S-1}\sum_{s=1}^{S} \mathbf{x}^{(s)}\left(\mathbf{x}^{(s)}\right)^\top$$

- Unfortunately $\nabla_\theta \mathbf{x}^{(s)} = \nabla_\theta f(\theta, \omega)$ depends on simulator

Stochastic gradients

- Finite differenc quotient for dimension d

$$\frac{\partial}{\partial \theta_d} \pi_\epsilon(\mathbf{y}|\theta) \approx \frac{\pi_\epsilon(\mathbf{y}|\theta_d + d_\theta) - \pi_\epsilon(\mathbf{y}|\theta_d)}{d_\theta}$$

- Too expensive, pick random directions
- Simultaneous perturbation stochastic approximation (SPSA)

$$\pi_\epsilon(\mathbf{y}|\theta) \approx \frac{\pi_\epsilon(\mathbf{y}|\theta + d_\theta \Delta) - \pi_\epsilon(\mathbf{y}|\theta - d_\theta \Delta)}{2 d_\theta} [\Delta_1^{-1}, \ldots, \Delta_D^{-1}]$$

  with random perturbation mask $\Delta_d \in \{-1, 1\}$
- Unbiased gradient estimator using $2D$ simulations

SGLD reminder

- Stochastic gradient Langevin (Welling & Teh 2011)
- Gradient descent + noise
- Proposal

$$\theta_{t+1} = \theta_t + \eta_t \mathcal{N}(0, M) - \frac{1}{2}\eta_t^2 \nabla \hat{U}(\theta)$$

- Correct as $\sum_t \eta_t = \infty$ and $\sum_t \eta^2 < 0$
- Local!

HMC reminder

- MCMC using Hamiltonian dynamics (Neal, 2011)
- Define joint log-density on (θ, ρ), the Hamiltonian

$$H(\theta, \rho) = U(\theta) + K(\rho)$$

where

$$U(\theta) = -\log \pi(\theta|\mathbf{y}) \qquad \text{and} \qquad K(\rho) = -\frac{1}{2}\rho^\top M^{-1} \rho$$

- Dynamics parametrised in $t \in \mathbb{R}$ on contours of H

$$d\theta = M^{-1} \rho dt \qquad \text{and} \qquad d\rho = -\nabla_\theta U(\theta) dt$$

- HMC is MCMC on (θ, ρ)-space
  - Re-sample ρ'
  - Simulate numerically $(\theta, \rho') \mapsto (\theta^*, \rho^*)$ using $dt = \eta$
  - Accept/reject

Stochastic gradient HMC

- Stochastic gradient HMC (Chen 2014)
- Stochastic gradient thermostats (Ding, 2014)
- The fundamental incompatibility of HMC of sub-sampling (Betancourt 2015)

- Replace $\nabla U(\theta)$ with noisy version $\nabla \hat{U}(\theta)$
- Mini-batches (Big Data), stochastic finite differences, etc
- Problem: noise form? CLT 'model':

$$\nabla \hat{U}(\theta) = \nabla U(\theta) + \mathcal{N}(\theta|\mathbf{0}, \eta^2 V(\theta))$$

- Dynamics become

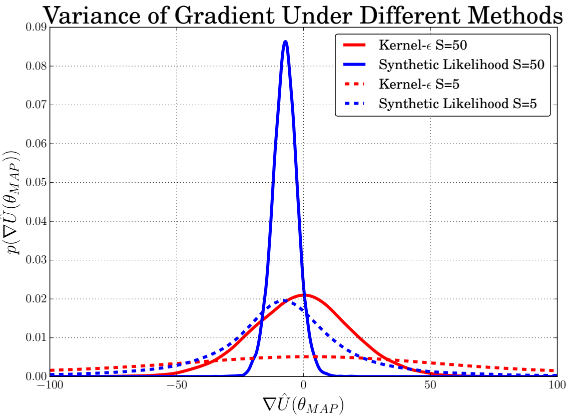
$$d\theta = M^{-1}\rho dt \qquad \text{and} \qquad d\rho = -\nabla_\theta U(\theta)dt + \mathcal{N}(0, \eta^2 V(\theta))dt$$

- Problem: H not invariant under those dynamics
- To correct: accept/reject (?) or add friction $-\eta^2 V(\theta) M^{-1} \rho dt$

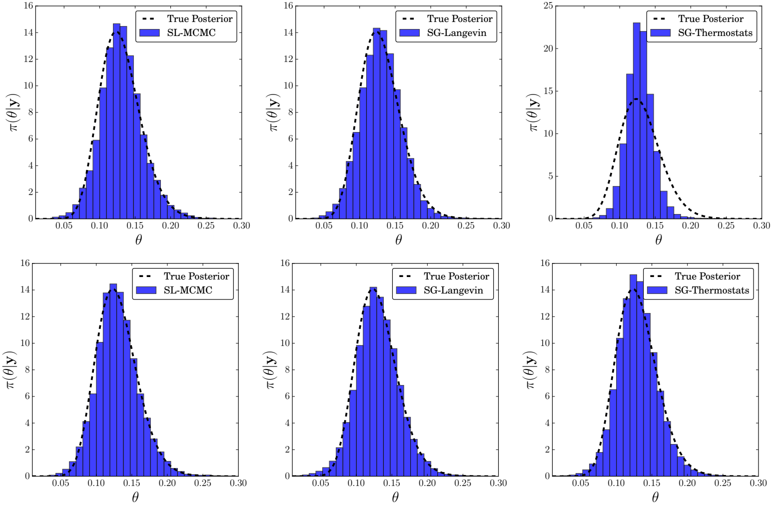
Bias vs. variance: synthetic likelihoods

Recall:

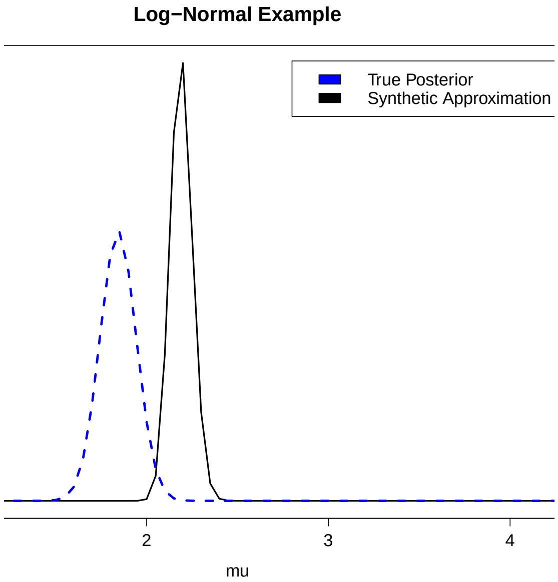
- Synthetic likelihood: $\pi_\epsilon(\mathbf{y}|\theta) = \mathcal{N}(\mathbf{y}|\mu_\theta, \sigma_\theta^2 + \epsilon^2)$
- Gaussian ϵ-kernel: $\pi_\epsilon(\mathbf{y}|\theta) = \frac{1}{S}\sum_{s=1}^{S} \mathcal{N}(\mathbf{x}^{(s)}|\mathbf{y}, \epsilon^2 I)$



Impact on posterior inference



Well ...

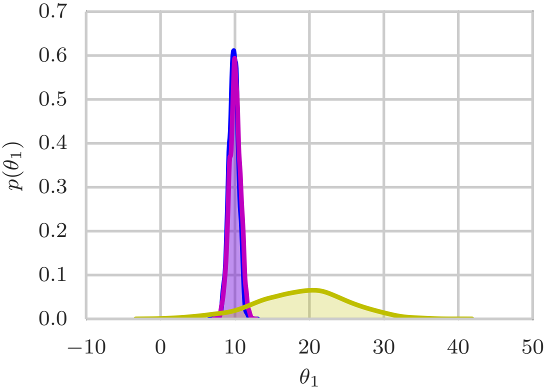


Log−Normal Example
True Posterior
Synthetic Approximation
2
3
4
mu

Impact on posterior inference

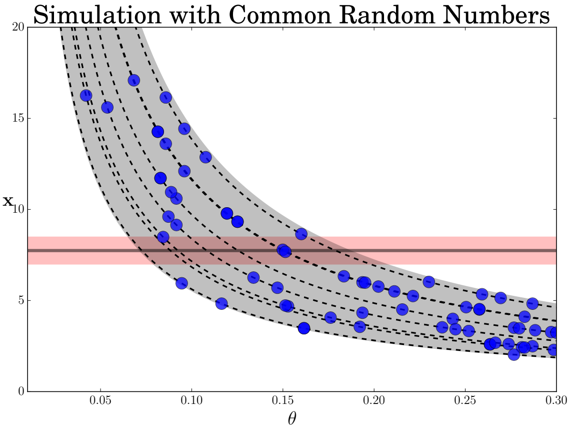
- Skew normal:

$$p(y|\theta) = \mathcal{N}(y|\mu = 10, 1)\,\Phi(10y)$$



Reduce noise 'for free' – sticky random numbers

- Recall $\nabla_\theta U(\theta) = -\nabla_\theta \pi(\theta)\pi_\epsilon(\mathbf{y}|\theta)$
- Numerical integration of HMC dynamics requires to evaluate $\nabla_\theta U(\theta)$ at each point of trajectory
- Assume $\nabla_\theta \pi_\epsilon(\mathbf{y}|\theta)$ is smooth in θ, use CRNs
- Deterministic simulation $\mathbf{x}^{(s)} = f(\theta, \omega)$ with seed ω



Reading suggestions

- MCMC using Hamiltonian dynamics (Neal, 2011)
- Stat. inference for noise nonlinear ecological dynamical systems (Wood, 2010)
- Stochastic gradient HMC (Chen, Fox, Guestrin, 2014)
- Stochastic gradient thermostats (Ding et al 2014)
- The fundamental incompatibility of HMC of sub-sampling (Betancourt 2015)
- Gaussian Process Surrograte ABC (Meeds, Welling, 2014)