A Review of
Pseudo-Marginal Markov Chain Monte Carlo

Discussed by: Yizhe Zhang

October 21, 2016
Outline

1. Overview
2. Paper review
3. Experiment
4. Conclusion
Motivation & overview

- **Notation:** $\theta$ denotes the parameter of interests. $Y$ denotes the observations.
- Standard task of *Markov Chain Monte Carlo* (MCMC): draw samples from posterior $\pi(\theta)$, where

$$\pi(\theta) \propto p(Y|\theta) p(\theta), \quad (1)$$

and $p(Y|\theta)$ can be estimated **pointwise**.
- **Pseudo-Marginal MCMC:** $p(Y|\theta)$ cannot easily be evaluated pointwise. However, an **unbiased** estimator $\hat{f}(\theta)$, where $\mathbb{E}\hat{f}(\theta) = p(Y|\theta)$ can be achieved for any $\theta$. 
Examples

- **Latent variable models**: \( \{X_n\}_{n=1,...} \in \mathcal{X} \) denote latent variable indexed by \( n \).

  \[
  X_n \sim g_1(\cdot|\theta),
  Y_n \sim g_2(\cdot|X_n, \theta),
  \]

  \[
  p(Y_n|\theta) = \int_{\mathcal{X}} g_1(X_k|\theta)g_2(Y_k|X_n, \theta) dX_k,
  \]

- The above integral can be intractable.

- **Approximate Bayesian computation (ABC)**: \( V \) denotes auxiliary variable. \( K(\cdot, \cdot): \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+ \) denotes a kernel. \( Y' \) are some simulated *Pseudo-observations*.

  \[
  V \sim g_1(\cdot),
  Y' = g_2(V, \theta),
  \]

  \[
  p(Y|\theta) = \int_{\mathcal{Y}} K(Y, Y')p(Y'|\theta) dY' = \int K(Y, g_2(V, \theta))p(V) dV,
  \]

- Again, this likelihood can be intractable.
Examples

- **Doubly-intractable distribution**: We know \( p(Y|\theta) \) up to a constant \( Z(\theta) \), i.e., \( p(Y|\theta) = g(y; \theta)/Z(\theta) \)

- We define an \( f(\theta) \), where \( \pi(\theta) \propto f(\theta) \):

\[
f(\theta) = \frac{g(y; \theta)p(\theta)Z(\hat{\theta})}{Z(\theta)},
\]

where \( \hat{\theta} \) is a fixed parameter, thus \( Z(\hat{\theta}) \) is a constant

- An *unbiased* estimator \( \hat{f}(\theta) \) can be achieved by importance sampling estimates.

\[
\frac{Z(\hat{\theta})}{Z(\theta)} = \frac{\int g(y; \hat{\theta})p(y|\hat{\theta})dy}{\int g(y; \theta)p(y|\theta)dy} \rightarrow \hat{f}(\theta) = \frac{g(y; \theta)p(\theta)g(y^{*}; \hat{\theta})}{g(y^{*}; \theta)}
\]

where \( y^{*} \) is a sample from \( p(y|\theta) \propto g(y; \theta) \),

\[
y^{*} \sim p(y|\theta)
\]
• “Pseudo-marginal” represents exact-approximate of likelihood.

• Representative papers on pseudo-marginal MCMC:
  • Iain Murray and Matthew M Graham (2015). “Pseudo-marginal slice sampling”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)*
Outline

1. Overview
2. Paper review
3. Experiment
4. Conclusion
### Pseudo-marginal with Metropolis-Hastings updates

- Developed upon Møller et al. (2006).

<table>
<thead>
<tr>
<th>Inputs:</th>
<th>current state $\theta$, previous estimate of its unnormalized target probability $\hat{f}$, proposal dist. $q$, unbiased estimator s.t. $\mathbb{E}_{\epsilon(f; \theta)}[\hat{f}] = f(\theta)$ for all $\theta$,</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>new state-estimate pair $(\theta, \hat{f})$.</td>
</tr>
</tbody>
</table>

1. Propose new state and estimate its probability:
   
   $$\theta' \sim q(\cdot; \theta)$$
   $$\hat{f}' \sim \epsilon(\cdot; \theta')$$

2. Metropolis–Hastings style acceptance rule,

   **with probability** $\min \left(1, \frac{\hat{f}'}{\hat{f}} \frac{q(\theta'; \theta')}{q(\theta'; \theta)} \right)$:

   - **Accept:** return $(\theta', \hat{f}')$
   - **else:**
     - **Reject:** return $(\theta, \hat{f})$
Pseudo-marginal with Metropolis-Hastings updates (Cont’d)

• Note that the \( \hat{f} \) is **recycled** each time. *i.e.* The \( \hat{f} \), as auxiliary variables, becomes part of the Markov Chain.
• When \( \hat{f} = f \), the algorithm recovers standard MH.
• Andrieu and Roberts (2009) demonstrated that, *only by doing this*, the convergence to the correct target distribution can be guaranteed.
• **Problem**: the acceptance ratio,

\[
\alpha = \frac{\hat{f}'q(\theta; \theta')}{\hat{f}q(\theta'; \theta)}, \tag{9}
\]

can be low if the \( \hat{f} \) has large variance \( \rightarrow \) **“sticking”** behavior of sampler

How to overcome this? Using conditional updating.
Auxiliary pseudo-marginal (APM)

- Alternate between updating $\theta$ and randomness $u$

**Inputs:** current state: parameters $\theta$, randomness $u$; unbiased estimator s.t. $\mathbb{E}_{q(u')}[\hat{f}(\theta'; u')] = f(\theta')$ for all $\theta'$,

**Output:** new state $(\theta, u)$.

1. Update $u$ leaving invariant its target conditional:
   \[
   \pi(u | \theta) \propto \hat{f}(\theta; u) q(u)
   \]

2. Update $\theta$ leaving invariant its target conditional:
   \[
   \pi(\theta | u) \propto \hat{f}(\theta; u)
   \]

**Figure:** Framework for Auxiliary Pseudo-Marginal (APM) methods.
Thoughts on the acceptance ratio

- They denote all the “randomness” as $u$. (i.e. the results of all calls to a random number generator)
- **Joint update:** $\alpha = \frac{\hat{f}(\theta', u')}{\hat{f}(\theta, u)}$
- Conditional updates for $\theta$: $\alpha = \frac{\hat{f}(\theta', u')q(u')q(u)}{\hat{f}(\theta, u)q(u)q(u')} = \frac{\hat{f}(\theta', u)}{\hat{f}(\theta, u)}$
- Conditional updates for $u$: $\alpha = \frac{\hat{f}(\theta, u')}{\hat{f}(\theta, u)}$

\[ f(\theta) = \mathbb{E}_{q(u)}[\hat{f}(\theta; u)] \]
Thoughts on the acceptance ratio

- The $\theta$ will almost always move, but the $u$ may still not move using MH.

Try other MCMC updates rules.
A quick review of slice sampling (SS)

**Input:** current state $\theta$, unnormalized target distribution $f$, initial search width $w$, whether to do optional part of update step.out.

**Output:** a new state $\theta'$. When $\theta$ is drawn from $\pi(\theta) \propto f(\theta)$, the marginal distribution of $\theta'$ is also $\pi$.

1. Random height under curve:
   \[ u_1 \sim \text{Uniform}[0,1] \]
   \[ h \leftarrow u_1 f(\theta) \]

2. Randomly place interval around the current state:
   \[ u_2 \sim \text{Uniform}[0,w] \]
   \[ [\theta_{\text{min}}, \theta_{\text{max}}] \leftarrow [\theta-u_2, \theta-u_2+w] \]

3. if step.out expand interval (linear step version):
   \[
   \text{while } f(\theta_{\text{min}}) > h : \quad \theta_{\text{min}} \leftarrow \theta_{\text{min}} - w
   \]
   \[
   \text{while } f(\theta_{\text{max}}) > h : \quad \theta_{\text{max}} \leftarrow \theta_{\text{max}} + w
   \]

4. Sample proposal on interval:
   \[ \theta' \sim \text{Uniform}[\theta_{\text{min}}, \theta_{\text{max}}] \]
Pseudo-marginal slice sampling (PM-SS)

- **Noisy slice sampler** [Murray (2007)]: Replacing $f(\theta)$ with $\hat{f}(\theta)$ still leads to valid slice sampler.

- **Elliptical SS** [Murray, Adams, and MacKay (2010)]: efficient when dealing with Gaussian prior.

- **Pseudo-marginal slice sampling**: Use noisy SS or noisy Elliptical SS to sample $u$. 
Outline

1. Overview
2. Paper review
3. experiment
4. conclusion
Demonstration

- 5-dimensional Gaussian

\[ p(\theta) = 1, g(x|\theta) = \exp[-1/2(x - \theta)^T(x - \theta)] \]  

- Here we pretend the normalizer \( Z(\theta) \) is **unknown**.
- Given the single observation \( x = 0 \), which gives the posterior of \( \theta \) will be \( \mathcal{N}(0, I) \).
- The unbiased estimator is by importance sampling as in (8).
- Comparing 5 schemes: PM-MH, APM (MI+MH), APM (MI+SS), APM (SS+MH), APM (SS+SS)
Demonstration (cont’d)

(a) PM-MH

(b) APM MI+MH

(c) $p(\theta_1)$ estimates

(d) Cost-scaled autocorrelation
Experiments

- **Ising model:**

\[
p(y|\theta) = \frac{1}{Z(\theta)} \exp\left( \sum_{i \neq j \in E} \theta_{ij} y_i y_j + \sum_i \theta_h y_i \right)
\]  

(11)
Experiments

(e) Cost-scaled autocorrelations

(f) Cost-scaled autocorrelations, with $K = 35$ AIS steps
Experiments

- **Hierarchical Gaussian process**: for $y \in \{+1, -1\}, X \in \mathbb{R}^d$, the marginal likelihood $p(y|\theta, X)$ cannot be evaluated exactly, as the marginalization does not have close form.

- Sample $\theta = (\sigma, \tau)$ of the exponential covariance function. (variance $\sigma$, length-scale $\tau$)

- Use an importance sampling estimate.

<table>
<thead>
<tr>
<th>Method</th>
<th>$N_{c,\text{op.}}/10^3$</th>
<th>Acc. rate</th>
<th>$\text{ESS}$</th>
<th>$\frac{\text{ESS}}{N_{c,\text{op.}}}/10^{-3}$</th>
<th>$\hat{R}$</th>
<th>$\text{ESS}$</th>
<th>$\frac{\text{ESS}}{N_{c,\text{op.}}}/10^{-3}$</th>
<th>$\hat{R}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM MH</td>
<td>74.0 (0.026)</td>
<td>0.201 (0.0065)</td>
<td>306 (7.8)</td>
<td>4.14 (0.11)</td>
<td>1.00</td>
<td>441 (8.6)</td>
<td>5.96 (0.12)</td>
<td>1.00</td>
</tr>
<tr>
<td>APM MI+MH</td>
<td>74.1 (0.029)</td>
<td>0.219 (0.0034)</td>
<td>357 (8.8)</td>
<td>4.81 (0.12)</td>
<td>1.00</td>
<td>512 (19)</td>
<td>6.92 (0.26)</td>
<td>1.00</td>
</tr>
<tr>
<td>APM SS+MH</td>
<td>74.1 (0.028)</td>
<td>0.204 (0.0046)</td>
<td>370 (7.1)</td>
<td>4.99 (0.097)</td>
<td>1.00</td>
<td>526 (26)</td>
<td>7.11 (0.35)</td>
<td>1.00</td>
</tr>
<tr>
<td>PM MH</td>
<td>97.8 (0.14)</td>
<td>0.180 (0.013)</td>
<td>185 (19)</td>
<td>1.90 (0.20)</td>
<td>1.01</td>
<td>277 (28)</td>
<td>2.83 (0.28)</td>
<td>1.00</td>
</tr>
<tr>
<td>APM MI+MH</td>
<td>98.3 (0.060)</td>
<td>0.208 (0.0046)</td>
<td>533 (5.8)</td>
<td>5.43 (0.059)</td>
<td>1.00</td>
<td>559 (13)</td>
<td>5.69 (0.13)</td>
<td>1.00</td>
</tr>
<tr>
<td>APM SS+MH</td>
<td>98.4 (0.054)</td>
<td>0.206 (0.0044)</td>
<td>519 (9.8)</td>
<td>5.27 (0.099)</td>
<td>1.00</td>
<td>631 (13)</td>
<td>6.41 (0.13)</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Figure**: Convergence and efficiency results
Experiments

**Figure:** Cost-scaled autocorrelations

(a) $\tau$ given Pima data

(b) $\tau$ given Breast data
Outline

1 Overview

2 Paper review

3 experiment

4 conclusion
Wrap-ups

- Do more conventional MCMC on auxiliary distributions.
- Easy to tune (no tuning on MI/MH), often better.
- Easy to implement.
Additional slides for ABC (from wikipedia)
References


