Stochastic Gradient Monomial Gamma Sampler

**Presenter:** Yizhe Zhang

**Joint with:** Changyou Chen, Zhe Gan, Ricardo Henao, Lawrence Carin

Duke University

August 9, 2017
Sampling from $f(\theta) \propto \exp(-U(\theta, X))$  

Bayesian model averaging; uncertainty estimation;  

SG-MCMC replaces $U(\theta, X)$ with an unbiased stochastic likelihood, $	ilde{U}(\theta, x_\tau)$, evaluated from a subset of data, $x_\tau$

$$
\tilde{U}(\theta) = -\frac{N}{N'} \sum_{i=1}^{N'} \log p(x_{\tau_i}|\theta) - \log p(\theta), \quad (1)
$$

where $\{\tau_1, \cdots, \tau_{N'}\}$ are random subsets.
Driven by a continuous-time Markov stochastic process.

\[ d\Gamma = V(\Gamma)dt + D(\Gamma)dW , \]  

where \( \Gamma \) denotes the parameters of the augmented system, e.g., \( p \) and \( \theta \), \( V(\cdot) \) and \( D(\cdot) \) are referred as drift and diffusion vectors, respectively, and \( W \) denotes a standard Wiener process.

To have a stationary distribution \( p(\Gamma) \), Fokker-Planck equation needs to be hold.

\[ \nabla_\Gamma \cdot p(\Gamma)V(\Gamma) = \nabla_\Gamma \nabla_\Gamma^T : [p(\Gamma)D(\Gamma)] \]
SGHMC (stochastic gradient Hamiltonian Monte Carlo) [Chen et. al., 2014] use stochastic gradient $\partial_\theta \tilde{U}(\theta)$, and introduce a friction term $B(\theta)$ to account for stochastic noise. The SDE is given as

$$d\theta = \partial_p K(p) dt$$

$$dp = -\partial_\theta \tilde{U}(\theta) dt - B(\theta) \partial_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt).$$

where $K(p)$ is the kinetics, $K(p) = p^T p / m$.
SGNHT (stochastic gradient Nosé-Hoover thermostat) [Ding et. al, 2015] generalize the SGHMC to use thermostat for estimating the stochastic noise.

\[ d\theta = \partial_p K(p) dt \]  \hspace{1cm} (5)

\[ dp = -\partial_\theta \tilde{U}(\theta) dt - \xi \partial_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt) \]  \hspace{1cm} (6)

\[ d\xi = (p^T p - 1) dt. \]  \hspace{1cm} (7)
Improving over SGMCMC

We propose three techniques for improving efficiency of SGMCMC.

- Use *generalized kinetics* which delivers superior mixing rate.
- Use *additional dynamic* which helps convergence, and has better ergodic properties.
- Use *stochastic resampling* which helps convergence.
More efficient kinetics

- We consider *monomial Gamma* (MG) [Zhang et. al. 2016] kinetics $K(p) = |p|^{1/a}$, where $a \geq 1$.
- 1) Better stationary mixing
- 2) Better exploring multimodal distribution.
- However, directly applying such $K(p)$ will not satisfy FP equation.
- We use a softened version of MG kinetics.

![Graph showing kinetic function values for different cases](image)

(Duke university) August 9, 2017
Hamiltonian system with a generalized form of kinetics and thermostat variable (stochastic noise).

\[ H = K(p) + U(\theta) + F(\xi), \]  

Consider SDE of SGNHT under this generalized form

\[ d\theta = \nabla K(p)dt \]  
\[ dp = -\left(\sigma_p + \gamma \nabla F(\xi) \right) \odot \nabla K(p)dt \]  
\[ \quad - \nabla U(\theta)dt + \sqrt{2\sigma_p}dW, \]  
\[ d\xi = \gamma \left[ \nabla K_c(p) \odot \nabla K(p) - \nabla^2 K(p) \right] dt. \]

With **numerical integrator**, \( \nabla U(\theta_t) \) is large \( \rightarrow p_{t+1} \) is large.

For \( a > 1 \), \( \nabla K(p) \approx |p|^{1/a-1} \). \( p_{t+1} \) is large \( \rightarrow \nabla K(p) \) is small \( \rightarrow \theta \) won’t change.
We consider adding first-order dynamics to $\theta$ and $\xi$

\[
\begin{align*}
    d\theta &= \nabla K_c(p) dt - \sigma_\theta \nabla U(\theta) dt + \sqrt{2\sigma_\theta} dW \\
    dp &= - (\sigma_p + \gamma \nabla F(\xi)) \odot \nabla K_c(p) dt \\
        &\quad - \nabla U(\theta) dt + \sqrt{2\sigma_p} dW, \\
    d\xi &= \gamma \left[ \nabla K_c(p) \odot \nabla K_c(p) - \nabla^2 K_c(p) \right] dt \\
        &\quad - \sigma_\xi \nabla F(\xi) dt + \sqrt{2\sigma_\xi} dW.
\end{align*}
\]

Fortunately, the first order Langevin directly compensate this with large updating signal $\nabla U(\theta_{t+1})$

On the other hand, when $\nabla U(\theta)$ is small, $\nabla K_c(p)$ would be large.

The proposed SDE also has better theoretic guarantee on the existence and convergence of bounded solutions for a particular differential equation.
Stochastic resampling

- Resample $p$ and $\xi$ from their marginal distribution $(\propto \exp[-K(p)]; \exp[-F(\xi)])$ with a fixed frequency.
- Move on a higher energy level is less efficient.
- Make the sampler to **immediately** move to a lower energy level.
- Converge to stationary distribution.

**Figure:** Stochastic resampling.
Theoretical properties

- Quantifying how fast the sample average, $\hat{\phi}_T$, converges to the true posterior average, $\bar{\phi} \triangleq \int \phi(\theta) \pi(\theta|X) d\theta$, for $\hat{\phi}_T \triangleq \frac{1}{T} \sum_{t=1}^{T} \phi(\theta_t)$, where $T$ is number of iterations.

**Theorem**

For the proposed SGMGT and SGMGT-D algorithms, if a fixed stepsize $h$ is used, we have:

- **Bias:** $\left| \mathbb{E} \hat{\phi}_T - \bar{\phi} \right| = O \left( 1/(Th) + h \right)$,

- **MSE:** $\mathbb{E} \left( \hat{\phi} - \bar{\phi} \right)^2 = O \left( 1/(Th) + h^2 \right)$.
Experiments overview

We evaluate our model on various tasks:

1. Toy task: multiple-well synthetic potential
2. Bayesian Logistic Regression
3. Latent Dirichlet Allocation
4. Discriminative RBM
5. Bayesian Recurrent Neural Network
Multiple-well Synthetic Potential

- Generate samples from a complex multimodal distribution.
- SGMGT-D: w/ 1st dynamics and resampling

Figure: Synthetic multimodal distribution. Left: empirical distributions for different methods. Right: traceplot for each method.
**Bayesian Logistic Regression**

Table: Average AUROC and median ESS. Dataset dimensionality is indicated in parenthesis after the name of each dataset.

<table>
<thead>
<tr>
<th>AUROC ($D$)</th>
<th>A (15)</th>
<th>G (25)</th>
<th>H (14)</th>
<th>P(8)</th>
<th>R (7)</th>
<th>C (87)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGNHT</td>
<td>0.89</td>
<td>0.75</td>
<td>0.90</td>
<td>0.86</td>
<td>0.95</td>
<td>0.65</td>
</tr>
<tr>
<td>SGMGT (a=1)</td>
<td>0.92</td>
<td>0.78</td>
<td>0.91</td>
<td>0.86</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>SGMGT-D (a=1)</td>
<td>0.95</td>
<td>0.86</td>
<td>0.95</td>
<td><strong>0.93</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.73</strong></td>
</tr>
<tr>
<td>SGMGT (a=2)</td>
<td>0.93</td>
<td>0.79</td>
<td>0.93</td>
<td>0.88</td>
<td>0.86</td>
<td>0.62</td>
</tr>
<tr>
<td>SGMGT-D (a=2)</td>
<td><strong>0.95</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.95</strong></td>
<td>0.90</td>
<td>0.97</td>
<td>0.69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SGNHT</td>
<td>869</td>
<td>941</td>
<td>1911</td>
<td>2077</td>
<td>1761</td>
<td>1873</td>
</tr>
<tr>
<td>SGMGT-D (a=1)</td>
<td><strong>3147</strong></td>
<td><strong>2131</strong></td>
<td>2448</td>
<td><strong>4244</strong></td>
<td>1494</td>
<td><strong>3605</strong></td>
</tr>
<tr>
<td>SGMGT-D (a=2)</td>
<td>2700</td>
<td>1989</td>
<td><strong>2768</strong></td>
<td>3430</td>
<td><strong>2265</strong></td>
<td>2969</td>
</tr>
</tbody>
</table>
Figure: Experimental results for DRBM. Upper-left: testing accuracies for SGLD, SGNHT, SGMGT and SGMGT-D. Upper-right through lower-right: traceplots for SGLD, SGNHT and SGMGT-D with $a = 2$, respectively.
Table: Test negative log-likelihood results on polyphonic music datasets and test perplexities on PTB using RNN.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Piano</th>
<th>Nott</th>
<th>Muse</th>
<th>JSB</th>
<th>PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGLD</td>
<td>11.37</td>
<td>6.07</td>
<td>10.83</td>
<td>11.25</td>
<td>127.47</td>
</tr>
<tr>
<td>SGNHT</td>
<td>9.00</td>
<td>4.24</td>
<td>7.85</td>
<td>9.27</td>
<td>131.3</td>
</tr>
<tr>
<td>SGMGT (a=1)</td>
<td>7.90</td>
<td>4.35</td>
<td>8.42</td>
<td>8.67</td>
<td>120.6</td>
</tr>
<tr>
<td>SGMGT (a=2)</td>
<td>10.17</td>
<td>4.64</td>
<td>8.51</td>
<td>8.84</td>
<td>250.5</td>
</tr>
<tr>
<td>SGMGT-D (a=1)</td>
<td><strong>7.51</strong></td>
<td><strong>3.33</strong></td>
<td>7.11</td>
<td>8.46</td>
<td>113.8</td>
</tr>
<tr>
<td>SGMGT-D (a=2)</td>
<td>7.53</td>
<td>3.35</td>
<td><strong>7.09</strong></td>
<td><strong>8.43</strong></td>
<td><strong>109.0</strong></td>
</tr>
<tr>
<td>SGD</td>
<td>11.13</td>
<td>5.26</td>
<td>10.08</td>
<td>10.81</td>
<td>120.44</td>
</tr>
<tr>
<td>RMSprop</td>
<td>7.70</td>
<td>3.48</td>
<td>7.22</td>
<td>8.52</td>
<td>120.45</td>
</tr>
<tr>
<td>ADAM</td>
<td>8.00</td>
<td>3.70</td>
<td>7.56</td>
<td>8.51</td>
<td>120.45</td>
</tr>
</tbody>
</table>

Figure: Learning curves of different SG-MCMC methods for RNN.
Conclusion and Future study

Conclusion:
- Scalable MCMC inference with improved stationary mixing efficiency.
- Remedies to alleviate practical issues with generalized HMC kinetics.
- Better theoretical guarantees.

Future research:
- Connection to optimization methods.
Q&A