Towards Improving the Efficiency and Scalability of MCMC inference

Yizhe Zhang

Ph.D. defense, Duke University

Committee chair:
Lawrence Carin

Committee members:
Katherine Heller, Alexander Hartemink,
Scott Schmidler, David Dunson

Jan, 26, 2018
Motivations

- **Why Bayesian inference?**
- Limited data, uncertainty estimation, model averaging...
- **What is MCMC?**
- MCMC simulates a Markov chain whose invariant states follow a given (target) probability.
- **Why MCMC?**
- Intractable integration.
- **What are some challenges in MCMC?**
- Efficiency and scalability.
Specific Aims

- The aim is to perform efficient and scalable Markov Chain Monte Carlo (MCMC) sampling from unnormalized density.
- Common in Bayesian inference, including many biomedical problems.
- Related publications:
  - *Laplacian Hamiltonian Monte Carlo*, Zhang et al., In ECML, 2016.
  - *Towards Unifying Hamiltonian Monte Carlo and Slice Sampling*, Zhang et al., In NIPS, 2016.
  - *Dynamic Poisson Factor Analysis*, In ICDM, Zhang et al., 2016.
Sampling from unnormalized density

- Suppose sampling from \( p(x) \propto \exp(-U(x)) \) is of interest, where \( U(x) \) represent the potential energy function.
- However, large proposal \( \rightarrow \) low acceptance ratio; small proposal \( \rightarrow \) slow move.
- Even with extensive tuning of proposals the *random walk* nature often delivers *inefficient mixing* of the Markov chain.
Auxiliary variable MCMC

- Toward improving the mixing efficiency, two auxiliary variable MCMC methods were developed.
- **Hamiltonian Monte Carlo (HMC)** was proposed to allow long-range movement with a high acceptance ratio, which significantly improves mixing performance.
- **Slice sampler (SS)** use auxiliary slice variables for efficient moves. These moves can be automatically adapted to match the relative scale of the local region being sampled.
Hamiltonian Monte Carlo

- Sampling from \( p(x) \propto \exp[-U(x)] \)
- HMC augment the density with auxiliary momentum \( p \in \mathbb{R}^d \). \( K(p) = \frac{1}{2}p^T M^{-1}p \) is the kinetic energy. \( H = U(x) + K(p) \) is the Hamiltonian.
- HMC iterates between two steps:
  1. Move along Hamiltonian contour to propose new samples for \( x \), driven by the following partial differential equations (PDEs):
     \[
     \frac{dx}{dt} = \nabla_p K(p) , \quad \frac{dp}{dt} = -\nabla_x U(x). \tag{1}
     \]
  2. Sample momentum \( p \) from its marginal distribution.
Hamiltonian Monte Carlo

Algorithm 1: Hamiltonian Monte Carlo

Input: Starting position $\theta^{(1)}$ and step size $\epsilon$

for $t = 1, 2 \cdots$ do

Resample momentum $r$

$r^{(t)} \sim \mathcal{N}(0, M)$

$(\theta_0, r_0) = (\theta^{(t)}, r^{(t)})$

Simulate discretization of Hamiltonian dynamics in Eq. (4):

$r_0 \leftarrow r_0 - \frac{\epsilon}{2} \nabla U(\theta_0)$

for $i = 1$ to $m$ do

$\theta_i \leftarrow \theta_{i-1} + \epsilon M^{-1} r_{i-1}$

$r_i \leftarrow r_{i-1} - \epsilon \nabla U(\theta_i)$

end

$r_m \leftarrow r_m - \frac{\epsilon}{2} \nabla U(\theta_m)$

$(\hat{\theta}, \hat{r}) = (\theta_m, r_m)$

Metropolis-Hastings correction:

$u \sim \text{Uniform}[0, 1]$

$\rho = e^{H(\hat{\theta}, \hat{r}) - H(\theta^{(t)}, r^{(t)})}$

if $u < \min(1, \rho)$, then $\theta^{(t+1)} = \hat{\theta}$

end
Slice sampling

- *Slice sampling* augments $x$ with a *slice variable* $y$.
- Iterates between two *uniform* sampling step:

  **Slicing:** $p(y_t| x_t) \propto 1$, s.t. $0 < y_t < f(x_t)$
  
  **Sampling:** $p(x_{t+1}| y_t) \propto 1$, s.t. $f(x_t) > y_t$

*Figure:* Slice sampling
Slice sampling (cont’d)

- Samples a joint distribution in a Gibbs sampling manner.

\[ p(x, y) = \begin{cases} 
\frac{1}{Z}, & 0 < y < f(x) \\
0, & \text{otherwise}
\end{cases} \]

where \( Z = \int f(x) \, dx \) is the normalizing constant.

- The evaluation of slice interval \( \mathbb{X} \triangleq \{x : f(x) > y\} \) is typically non-trivial.

- Iterative procedures are used to adaptively capture the boundaries [Neal (2003)].
Unifying HMC with slice sampling

Outline

1 Preliminaries

2 Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC

3 Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies

4 Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics

5 Conclusion

6 Acknowledgements
Unifying HMC and slice sampling

HMC and slice sampling share many similarity, are they connected?

• We consider **generalized HMC** with kinetic

\[ K(p) = |p|^{1/a}, \ a > 0 \]  \hspace{1cm} (2)

• We showed, this **generalized HMC** is indeed equivalent to a **generalized slice sampler** as following:

**Slicing**: \( p(y_t|x_t) = \frac{1}{\Gamma(a)f(x_t)}[\log f(x_t) - \log y_t]^{a-1}, \)

\[ s.t. \ 0 < y_t < f(x_t) \] \hspace{1cm} (3)

**Sampling**: \( q(x_{t+1}|y_t) = \frac{1}{Z_2(y_t)}[\log f(x_{t+1}) - \log y_t]^{a-1} \)

\[ s.t. f(x_t) > y_t \] \hspace{1cm} (4)
Unifying HMC and slice sampling (Cont’d)

**Figure:** Generalized HMC and equivalent generalized SS. Red and blue dashed lines denote the conditionals $p(y_t | x_t)$ and $q(x_{t+1} | y_t)$, respectively.

<table>
<thead>
<tr>
<th>$a$</th>
<th>HMC space</th>
<th>Canonical (SS) space</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td><strong>Standard HMC</strong></td>
<td>MG-SS</td>
</tr>
<tr>
<td>1</td>
<td>MG-HMC</td>
<td><strong>Standard SS</strong></td>
</tr>
<tr>
<td>Otherwise (a&gt;0)</td>
<td>MG-HMC</td>
<td>MG-SS</td>
</tr>
</tbody>
</table>
This connection between generalized HMC and generalized SS is revealed by *Hamilton-Jacobi equation* (HJE).

In HJE, the original system \((H, x, p, \tau)\) is transformed to \((H', x', p', \tau)\), while the Hamilton’s equation (1) is preserved.

The HJE is employed to find the particle position \(x^* \triangleq x(\tau)\) for dynamic evolution duration \(\tau\).
Using Hamilton-Jacobi equation to solve the dynamic

- From HJE, the $x^*$ can be achieved by solving (5), with an evolutionary time, $\tau$.

$$\tau = \int_{x_{\text{min}}}^{x^*} \max\{H - U(z), 0\}^{a-1} dz - C.$$  \hspace{1cm} (5)

where $C$ is a constant, $x_{\text{min}} = \arg\min_{U(x) \leq H^x} U(x)$

- (inverse transform sampling) Uniformly sampling $\tau$ and solving $x^*$ from (5), is equivalent to directly sampling $x^*$ from the density:

$$p(x^*|H) \propto [H - U(x^*)]^{a-1}, \; \text{s.t.,} \; H - U(x^*) \geq 0.$$  \hspace{1cm} (6)
Intuitions about the connection

- **dynamic updating step** in HMC \( \iff \) **conditional sampling step** (given slice variable) in SS.
- **resampling a momentum** \( p_t \) in HMC \( \iff \) **sampling a slice variable** in SS.
- **Hamiltonian** \( H \) \( \iff \) **slice variable** \( y \) \( (H_t = -\log y_t) \).

Interesting to know.. But so what?
Unifying HMC with slice sampling

What can we do based on this connection?

- **First**, this connection enables *theoretical characterization* of mixing rate of HMC, which has not been well-explored.
- **Second**, a more efficient HMC can be derived.
Outline

1 Preliminaries

2 Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC

3 Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies

4 Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics

5 Conclusion

6 Acknowledgements
Generalized HMC sampling in practice

- Generalized kinetic $K(p; m, a) = \frac{|p|^{1/a}}{m}$, $a, m > 0$
- **Monomial Gamma** (MG) distribution:
  $$\pi(p; m, a) = \frac{m^{-a}}{2^a \Gamma(a+1)} e^{-\frac{|p|^{1/a}}{m}}$$
- MG$(a, m) = S \cdot G^a$ where $G \sim \text{Gamma}(a, m)$

**Algorithm 1** Monomial Gamma HMC (MG-HMC)

1. **Input**: Total sample size $T$, MG parameter $a$.
2. **Output**: Sample results, $\{x_0, \cdots, x_T\}$.
3. **for** $t = 1$ to $T$ **do**
4. (Sample momentum) Sample $p_t \sim \text{MonomialGamma}(m, a)$.
5. (Hamiltonian dynamic flow) Numerically simulate $\frac{dx}{dt} = \nabla_p K(p), \frac{dp}{dt} = -\nabla_x U(x)$ to get $(x^*, p^*)$
6. (Metropolis Hastings) accept $x^*$ with probability
   $$\min(1, \exp(-H(x^*, p^*) + H(x, p)))$$
7. **end for**
Analyzing mixing performance [autocorrelation]

- Connection to generalized SS enables theoretical analysis for generalized HMC.
- The following theorem states that a larger \(a\) (heavier tail kinetics) would lead to lower autocorrelation during stationary sampling period.

**Theorem (Asymptotic autocorrelation)**

*For univariate target distribution, the one time lag autocorrelation \(\rho(x_t, x_{t+1})\) of the analytic generalized SS parameterized by \(a\) asymptotically approaches zero when \(a \to \infty\), under regularity condition of \(U(x)\) and stationary assumption.*
**Intuition:** [small \(a\)] \(\rightarrow\) \([y_t \text{ stay close to } f(x_t)] \rightarrow [f(x_{t+1}) \text{ close to } f(x_t)].\)
Analyzing mixing performance [ESS and ergodicity]

- The following theorem states effective sample size (ESS)
  \[ \text{ESS} \triangleq \frac{N}{1 + 2 \times \sum_{h=1}^{\infty} \rho(h)} \] goes to full when \( a \to \infty \), indicating approximating \( i.i.d. \) samples.

**Theorem (limiting ESS)**

*If 1) the variance of transition kernel \( \text{Var}_{\kappa_h}(\cdot, x)(x) \) is bounded, 2) uniform ergodicity can be established. When \( a \to \infty \), we have, \( \text{ESS} \to N \)*

- Establishing the geometric ergodicity requires
  \[
y \frac{d}{dy} \int_{f(x) > y} [\log f(x) - \log y]^{a-1}
  \]
  to be non-increasing with \( y \). For \( U(x) = x^\omega, \omega > 0 \), such condition holds.
Improving stationary efficiency of HMC

Case study

- **1D exponential distribution** $\text{Exp}(\theta)$, $U(x) = \theta x$, $x \geq 0$.
- After some algebra,

$$\rho(1) = \frac{1}{a+1}, \quad \rho(h) = \frac{1}{(a+1)^h}, \quad \text{ESS} = \frac{Na}{a+2}.$$ 

- For the exponential family class of model [Roberts and Tweedie (1996)], with potential energy $U(x) = x^\omega$, $x \geq 0$, $\omega > 0$, $\rho(1)$ decays at a rate of $O(a^{-1})$. 
MG-HMC with large $a$ is particularly advantageous for sampling *multimodal distributions*.

For multimodal distribution, there exist *disjoint components* with same Hamiltonian level, which HMC can not freely jump between.

We showed that the chance of being on a disjoint energy level *goes to zero*, when $a \to \infty$.

**Figure:** Disjoint components associated with same Hamiltonian $H$. 

---

**Improving stationary efficiency of HMC**

---

**Towards unifying HMC and SS**

---

**Scalable and efficient MCMC inference**

---

**Biomedical applications**

---

**Conclusion**

---

**Acknowledgements**

---

**References**

---
No free lunch

- Such a performance gain does not come in free.
- First, as $a$ gets larger, the numerical difficulty in Hamiltonian dynamic updating is increased.
- Second, with bad initialization, the sampler may have slow initial convergence to the true target distribution. ($a > 1$)

In addition, not scalable to larger datasets

**Figure:** Hamiltonian contours when $a = 0.5$ and $a = 2$. 
Simulation studies

- **1D unimodal problems** Univariate toy distributions

  \[ p(x) = \frac{1}{Z_1} \exp(-E(x)), \text{ s.t. } x \geq 0: \]

  1) **Exponential distribution** \( \text{Exp}(\theta) \), where \( E(x) = \theta x \).

  2) **Positive-truncated Gaussian** \( \mathcal{N}_+(0, \theta) \), \( E(x) = x^2 \).

  3) **Gamma distributions** \( \text{Gamma}(r, \theta) \), where

  \[ E(x) = -(r - 1) \log x + \theta x, \text{ where } r = 2 \text{ and } r = 3 \]

**Figure:** Theoretical and empirical \( \rho(1) \) and ESS of exponential distribution (upper) and \( \mathcal{N}_+ \) (lower).
Real-world problems

- **Bayesian logistic regression** with various dimensionality.
- \( a > 1 \): performance decrease quickly with increasing dimensionality.
- \( a = 1 \): exceptional and robust in most cases.

<table>
<thead>
<tr>
<th></th>
<th>Aus (15)</th>
<th>Ger (25)</th>
<th>Hea (14)</th>
<th>Pim (8)</th>
<th>Rip (7)</th>
<th>Cav (87)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a = 0.5 )</td>
<td>3124</td>
<td>3447</td>
<td>3524</td>
<td>3434</td>
<td>3317</td>
<td>33</td>
</tr>
<tr>
<td>( a = 1 )</td>
<td>4308</td>
<td>4353</td>
<td>4591</td>
<td>4664</td>
<td>4226</td>
<td>36</td>
</tr>
<tr>
<td>( a = 2 )</td>
<td>1490</td>
<td>3646</td>
<td>4315</td>
<td>4424</td>
<td>1490</td>
<td>7</td>
</tr>
</tbody>
</table>

**Table:** Minimum ESS for each method in BLR experiments (dimensionality in parenthesis)

- **Independent Component Analysis (ICA)** [Vigário et al. (1998)]

<table>
<thead>
<tr>
<th></th>
<th>min ESS</th>
<th>Time(s)</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a = 0.5 )</td>
<td>2677</td>
<td>525</td>
<td>0.98</td>
</tr>
<tr>
<td>( a = 1 )</td>
<td>3029</td>
<td>517</td>
<td>0.97</td>
</tr>
<tr>
<td>( a = 2 )</td>
<td>1534</td>
<td>512</td>
<td>0.77</td>
</tr>
</tbody>
</table>

**Table:** Results for ICA on MEG data. \( d = 25, N = 17730 \).
Take-aways

• A MCMC method that has a *faster stationary mixing* theoretically v.s. HMC, yielding lower variance for sample based estimator (with fixed sample size).

• Especially helpful when the target distribution is *multimodal*.

• Suffers from numeric difficulty, initial convergence and scalability issues.

• *Future directions*: higher-order numerical integrator, geometric adaptation [Girolami and Calderhead (2011) and Nishimura and Dunson (2016a)].
Outline

1. Preliminaries
2. Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC
3. Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies
4. Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics
5. Conclusion
6. Acknowledgements
Sampling from $f(\theta) \propto \exp(-U(\theta, X))$

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

$$\tilde{U}(\theta) = -\frac{N}{N'} \sum_{i=1}^{N'} \log p(x_{\tau_i} | \theta) - \log p(\theta), \quad (7)$$

where $\{\tau_1, \cdots, \tau_{N'}\}$ are random subsets.
Background on stochastic gradient MCMC

**Background: Stochastic Gradient MCMC**

- Driven by a continuous-time *Markov stochastic process*.
  \[ d\Gamma = V(\Gamma)dt + D(\Gamma)dW, \]  
  \[ (8) \]

- \( \Gamma \) 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 satisfied.
  \[ \nabla_\Gamma \cdot p(\Gamma)V(\Gamma) = \nabla_\Gamma \nabla_\Gamma^T : [p(\Gamma)D(\Gamma)] \]
**Background on stochastic gradient MCMC**

### Background: Stochastic Gradient Hamiltonian Monte Carlo

- **SGHMC** (stochastic gradient Hamiltonian Monte Carlo) [Chen, Fox, and Guestrin (2014)] use stochastic gradient
  \( \nabla_\theta \tilde{U}(\theta) \)

- A friction term \( B(\theta) \) is introduced to account for stochastic noise.

- The SDE is given as
  \[
  \begin{align*}
  d\theta &= \nabla_p K(p) dt \\
  dp &= -\nabla_\theta \tilde{U}(\theta) dt - B(\theta) \nabla_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt). \tag{10}
  \end{align*}
  \]

- However, estimating \( B(\theta) \) is difficult.
SGNHT (stochastic gradient Nosé-Hoover thermostat) [Ding et al. (2014)] use thermostat for estimating the stochastic noise.

\[
d\theta = \nabla_p K(p) dt \\
dp = -\nabla_\theta \tilde{U}(\theta) dt - \xi \nabla_p K(p) dt + \mathcal{N}(0, 2B(\theta) dt) \\
d\xi = (p^T p - 1) dt.
\]
Outline

1. Preliminaries

2. Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC

3. Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies

4. Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics

5. Conclusion

6. Acknowledgements
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) kinetics $K(p) = |p|^{1/a}$, where $a \geq 1$.
- Better 1) stationary mixing 2) exploring multimodal distribution.
- However, directly applying such $K(p)$ will not satisfy Fokker-Planck equation.
- We use a differentiable version of MG kinetics, which maintain same tail behavior with stiff kinetic.
Additional First Order Dynamics

- Augmented Hamiltonian system with kinetics and thermostat.

\[ H = K(p) + U(\theta) + F(\xi), \]  \hspace{1cm} (14)

- SDE under this generalized SGMCMC (denote as \textit{SGMGT})

\[ d\theta = \nabla K(p) dt \]  \hspace{1cm} (15)
\[ dp = - (\sigma_p + \gamma \nabla F(\xi)) \odot \nabla K(p) dt \]  \hspace{1cm} (16)
\[ - \nabla U(\theta) dt + \sqrt{2\sigma_p} dW, \]  \hspace{1cm} (17)
\[ d\xi = \gamma \left[ \nabla K_c(p) \odot \nabla K(p) - \nabla^2 K(p) \right] dt. \]  \hspace{1cm} (18)

- With \textbf{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.
Additional First Order Dynamics (Cont’d)

• Adding first-order dynamics to $\theta$ and $\xi$

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

(19)

• 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, $\nabla K(p)$ would be large.

• The proposed SDE also has better theoretic guarantee on the existence and convergence of bounded solutions
Efficient MCMC with batch data

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
- **immediately** move to lower energy levels.
- Denoting SGMGT with add. Langevin & resampling as $SGMGT-D$

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

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

1. Preliminaries
2. Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC
3. Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies
4. Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics
5. Conclusion
6. Acknowledgements
Empirical studies

Multiple-well Synthetic Potential

- Generated samples has better stationary mixing.

**Figure:** Synthetic multimodal distribution. Left: empirical distributions for different methods. Right: traceplot for each method.
Empirical studies

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>0.93</td>
<td>0.98</td>
<td>0.73</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>0.95</td>
<td>0.90</td>
<td>0.95</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>3147</td>
<td>2131</td>
<td>2448</td>
<td>4244</td>
<td>1494</td>
<td>3605</td>
</tr>
<tr>
<td>SGMGT-D(a=2)</td>
<td>2700</td>
<td>1989</td>
<td>2768</td>
<td>3430</td>
<td>2265</td>
<td>2969</td>
</tr>
</tbody>
</table>
Empirical studies

Discriminative RBM

Figure: Left: testing accuracies for SGLD, SGNHT, SGMGT and SGMGT-D. Middle-left through right: traceplots for SGLD, SGNHT and SGMGT-D.
Empirical studies

Recurrent Neural Network

Table: Test negative log-likelihood results on various datasets.

<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>7.09</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>
</tbody>
</table>

Figure: Learning curves of different SG-MCMC methods.
Conclusion:

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

Future research:

- Adaptive selection of monomial parameters.
- Connection to optimization methods.
Dynamic Poisson factor analysis for gut microbiome study

Outline

1. Preliminaries
2. Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC
3. Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies
4. Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics
5. Conclusion
6. Acknowledgements
Dynamic metagenomic topic modeling

- **Motivation:** identify “topics” in human gut microbiota.
- **Data:** longitudinal measurements of human gut microbiota over time, from 6 subjects spanning 3 different studies.
- DNA reads mapped into 33750 *Operational Taxonomic Units* (OTUs). OTU defines species, represented as counts.
- 129 time-steps (non-uniform over a year).
- **Challenges:** Nonuniform time span; relatively large scale; high sparsity level (85%); abundance vs existence.
- **Model:** Dynamic Poisson factor model with $K = 50$. 

![Graphs showing dynamic Poisson factor analysis for gut microbiome study](image)
Dynamic modelling for discrete time-series data

- A dynamic model for discrete time-series data.
- The model is specified by constructing a hierarchy of **Poisson factor analysis** blocks.
- In experimental results on microbiome data, we identified topics associated with disease infection and recovery, which can be verified from domain knowledge.
Dynamic Poisson factor analysis for gut microbiome study

**DPFA Model: emission**

- **Emission model:** Poisson factor model [Henao et al. (2015)].
  
  \[ x_{nt} \sim \text{Poisson}(\Psi(\theta_{nt} \circ h_{nt})) , \]  
  \( n = \{1, \ldots, N\} \), \( t = \{1, \ldots, T\} \), \( k = \{1, \ldots, K\} \)

- We specify prior distributions as,

  \[
  \psi_k \sim \text{Dirichlet}(\eta_{\psi} 1_M) ,
  \theta_{knt} \sim \text{Gamma}(r_k, b_\theta) ,
  \]

  \[
  h_{knt} \sim \text{Bernoulli}(\pi_{knt}) ,
  \]

  \( k = \{1, \ldots, K\} \), \( n = \{1, \ldots, N\} \), \( t = \{1, \ldots, T\} \)
Dynamic Poisson factor analysis for gut microbiome study

**DPFA Model: emission**

- **Transition model:** Bernoulli-Poisson link [Zhou (2015)]

\[
\begin{align*}
h_{nt} &= 1 (z_{nt} > 0) , z_{nt} \sim \text{Poisson} \left( \tilde{\lambda}_{nt} \right) , \\
\tilde{\lambda}_{nt} &= \tau_{nt}^{-1} \Phi(w_{nt-1} \circ h_{nt-1}) + \tilde{\lambda}_0
\end{align*}
\]  

- Equivalently,

\[
p(h_{nt} = 1) = \text{Bernoulli} \left( 1 - \exp(-\tilde{\lambda}_{nt}) \right) ,
\]

- We specify prior distributions as,

\[
\phi_k \sim \text{Dirichlet} \left( \eta_\phi 1_K \right) , w_{knt-1} \sim \text{Gamma}(s_k, b_w)
\]

- Sensitive to existence (vs abundance); inference conveniency;
SGMGT-embedded Gibbs Inference

- For local variables, the conditional posterior can be derived.
- Depend only on non-zero elements of $x_{nt}$ and $z_{nt}$; can be parallelized.
- For global variables $\Theta \triangleq \{\theta, \Psi, w, \Phi, \tilde{\lambda}_0\}$, use SGMGT for fast approximate inference.

$$
\pi(\Theta|h, z) \propto p(\Theta)p(x|h, \Theta)p(h, z|\Theta).
$$

- Compared with full Gibbs approach.
Dynamic metagenomic topic modeling

- **Topic 49** (Proteobacteria) is consistent with the onset of a Salmonella infection
- **Topic 38** (Firmicutes), **Topic 42** (Tenericutes) is related to its recovery period.
- **Topic 14** present up to the time of infection does not reappear after recovery.
- All are consistent with the findings of David et al. (2014)
**Topic intensities**

**Figure:** Intensity heatmap for microbiome data with 50 topics (y axis). The x axis represents time in days.
Quantitative analysis

**Table:** One-step ahead forecasting results on microbiome data.

<table>
<thead>
<tr>
<th>Sample</th>
<th>#OTU</th>
<th>T</th>
<th>DPFA (Gibbs)</th>
<th>DPFA (SGMGT)</th>
<th>Naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>5432</td>
<td>321</td>
<td>0.880±0.008</td>
<td>0.866±0.011</td>
<td>0.761</td>
</tr>
<tr>
<td>S2</td>
<td>5432</td>
<td>189</td>
<td>0.755±0.044</td>
<td>0.613±0.067</td>
<td>0.378</td>
</tr>
<tr>
<td>S3</td>
<td>9371</td>
<td>30</td>
<td>0.989±0.003</td>
<td>0.951±0.005</td>
<td>0.790</td>
</tr>
<tr>
<td>S4</td>
<td>9371</td>
<td>30</td>
<td>0.964±0.006</td>
<td>0.948±0.009</td>
<td>0.760</td>
</tr>
<tr>
<td>S5</td>
<td>33750</td>
<td>332</td>
<td>0.943±0.003</td>
<td>0.932±0.007</td>
<td>0.835</td>
</tr>
<tr>
<td>S6</td>
<td>33750</td>
<td>129</td>
<td>0.975±0.002</td>
<td>0.960±0.005</td>
<td>0.843</td>
</tr>
</tbody>
</table>

- Same amount of Gibbs burnin and collection iterations.
- Full Gibbs(S1) 16265s (single CPU+ Titan X GPU).
- SGMGT(S1) 6149s
- SGMGT is roughly 3 times faster comparing to full Gibbs.
Outline

1 Preliminaries

2 Towards unifying HMC and SS
   - Unifying HMC with slice sampling
   - Improving stationary efficiency of HMC

3 Scalable and efficient MCMC inference
   - Background on stochastic gradient MCMC
   - Efficient MCMC with batch data
   - Empirical studies

4 Biomedical applications
   - Dynamic Poisson factor analysis for gut microbiome study
   - Neural topic analysis for genetic infection diagnostics

5 Conclusion

6 Acknowledgements
Motivation & Data

- **Motivation**: Diagnostic approaches to accurately discriminate between *viral* and *bacterial* etiology versus *non-infectious* causes of febrile illness.
- **Task**: Predict 3 pathogen classes (bacterial, viral, non-infectious) from gene expression data.
- **Model**: Interpretable non-linear topics that characterize key genes/isoforms.
- **Data**: $p = 55,688$ isoforms; 21,203 genes; $N = 212$ subjects.
Method overviews

- Abstract counts $d_i \in \mathbb{R}^V$ into a topic vector $h_i \in \mathbb{R}^K$.
- Differs from traditional topic modeling strategy in:
  - Each topic represents a *non-linear* composition of vocabulary.
  - Topics are selected according to *supervised* signal.
  - A *two-way Dirichlet prior* is introduced for the topic loading weight matrix, to induce *sparsity, non-negativity* and *non-overlapping* property.
  - Instead of using full-batch MCMC or MAP, we consider *SGMGT* for inference.
Neural topic analysis for genetic infection diagnostics

Two-way Dirichlet prior for neural topic loading

\[ h_i = \sigma(W d_i), \]
\[ y_i \sim \text{softmax}(U h_i), \]
\[ W, U \sim p_W(\cdot), p_U(\cdot) \]
• Desired properties for $W$: *non-negative, sparse, interpretable* and exclusive.

• Consider a *two-way* Dirichlet prior for $W \in \mathbb{R}^{M \times N}$

\[
p(W_{mn}) \propto A_{mn}B_{mn} \tag{26}
\]
\[
A_m \sim Dir(\alpha), B_n \sim Dir(\beta) \tag{27}
\]

where $A_m$ is a row vector and $B_n$ is a column vector.

• Sample auxiliary variables $\tilde{A}_m$ and $\tilde{B}_n$, which has log-Gamma distribution as prior. Consequently,

\[
A_m = \text{softmax}(\tilde{A}_m), B_n = \text{softmax}(\tilde{B}_n) \tag{28}
\]
Isoform composition inference

- Further consider an additional layer to learn isoform composition for specific gene.

\[ g_i = \sigma(V d_i), \quad (29) \]
\[ h_i = \sigma(W g_i), \quad (30) \]
\[ y_i \sim \text{softmax}(U h_i) \quad (31) \]

- \( V \) is specified to have a \textit{masked} two-way Dirichlet distribution.

\[ p(V_{nl}) \propto C_{nl} D_{nl} M_{nl} \quad (32) \]
\[ C_n \sim \text{Dir}(\gamma) \quad (33) \]
\[ D_l \sim \text{Dir}(\eta) \quad (34) \]

- \( M \in \{0, 1\}^{N \times L} \) indicates whether \( n \)-th isoform is from \( l \)-th gene.
Neural topic analysis for genetic infection diagnostics

Isoform composition inference

**Figure:** Isoform composition model
Inference details

- Omitting constant the log-likelihood objective can be written as:

\[
\mathcal{L} = \log p(Y|X, \Theta) + (1 - \alpha) \sum_{ij} \log A_{ij} + (1 - \beta) \sum_{ij} \log B_{ij} \\
+ (1 - \gamma) \sum_{ki} \log C_{ki} + (1 - \eta) \sum_{ki} \log D_{ki} + \text{const.} \tag{35}
\]

where \( \Theta \triangleq \{A, B, C, D\} \)

- Inference via SGMGT

- For baseline we consider group lasso method [Friedman, Hastie, and Tibshirani (2010)]. Each column in the weight matrix is considered as one group.
Results
Prediction

- Testing error rate via 10-fold cross validation. 3 topics. For SGMGT we collect 500 posterior samples for testing.

<table>
<thead>
<tr>
<th>Model</th>
<th>Group Lasso</th>
<th>Ours (Two-way Dir)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene-level model</td>
<td>0.187 ± 0.050</td>
<td>0.162 ± 0.083</td>
</tr>
<tr>
<td>Isoform -level model</td>
<td>0.177 ± 0.066</td>
<td>0.149 ± 0.075</td>
</tr>
</tbody>
</table>

**Table:** Error rate on testset with 10 fold cross-validation

- Traceplot of weight parameters.
Results
Non-overlapping sparse topic loading

- Two-way Dirichlet prior shows non-negative, non-overlapping and sparse topic intensity.
Results
Interpretable topic inference

- Identified 3 topics correspond to 3 infection types.

Figure: Topic intensity level for each group of infection type
Conclusions and future works

**Conclusions:**
- Unifying HMC and SS in a theoretical perspective.
- Proposing MG-HMC with better stationary mixing.
- Proposing Scalable MCMC inference to remedy practical issues of previous method.
- Discussing scalable Bayesian inference to many biomedical problem.

**Future works:**
- Developing better numerical integrator; adaptive selection of hyper-parameters [Nishimura and Dunson (2016b)].
- Developing geometric adaptation on Riemannian manifold for higher dimensional cases.
- Potential for discrete HMC sampling [Nishimura, Dunson, and Lu (2017)].
Acknowledgment

- **My advisor:** Dr. Lawrence Carin
- **My committees:** Dr. Katherine Heller, Dr. Scott Schmidler, Dr. Alexander Hartemink, Dr. David Dunson
- **My collaborators:** Dr. Ricardo Henao, Dr. Changyou Chen, Zhe Gan, Dr. Xiangyu Wang, Kai Fan, Dinghan Shen, Guoyin Wang, Jianqiao Li, Siyang Yuan, Chunyuan Li, Yunchen Pu, Wenlin Wang, Dr. Jonathan Mattingly, Dr. Jianfeng Lu, Liqun Chen, Shuyang Dai...
Thank You!

David, Lawrence A et al. (2014). “Host lifestyle affects human microbiota on daily timescales”. In: Genome Biology 15.7.


References II


