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.
  - *Stochastic Gradient Monomial Gamma Sampler*, In ICML, Zhang et al., 2017.
  - *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.
- *Metropolis-Hastings* (MH) achieves great success.
- 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

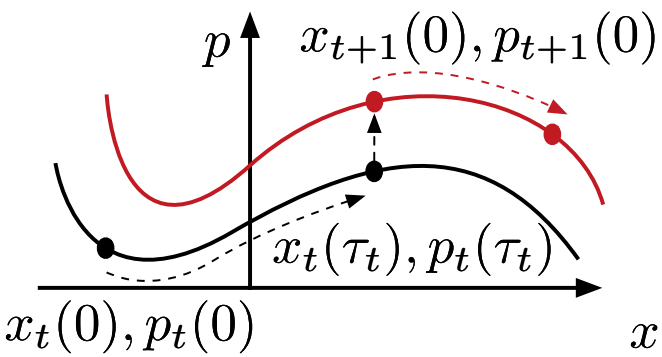
- Toward improving the mixing efficiency, two auxiliary variable MCMC methods were developed.
- *Hamiltonian Monte Carlo (HMC)* was proposed to allow long-range movement with a high acceptance ratio, which significantly improves mixing performance.
- *Slice sampler (SS)* use auxiliary slice variables 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 $p \in \mathbb{R}^d$. $K(p) = \frac{1}{2} p^T \boldsymbol{M}^{-1} p$ 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):

$$\frac{dx}{dt} = \nabla_p K(p) \ , \qquad \frac{dp}{dt} = -\nabla_x U(x) \, . \tag{1}$$

  2. Sample momentum p from its marginal distribution.



Hamiltonian Monte Carlo

Algorithm 1: Hamiltonian Monte Carlo

Input: Starting position $\theta^{(1)}$ and step size ϵ
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: $p(y_t|x_t) \propto 1 \quad , s.t. 0 < y_t < f(x_t)$

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

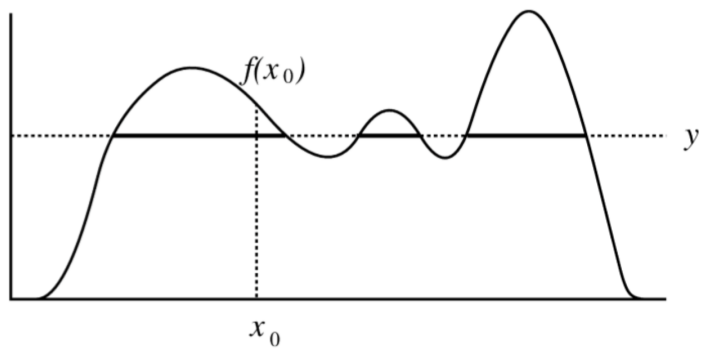


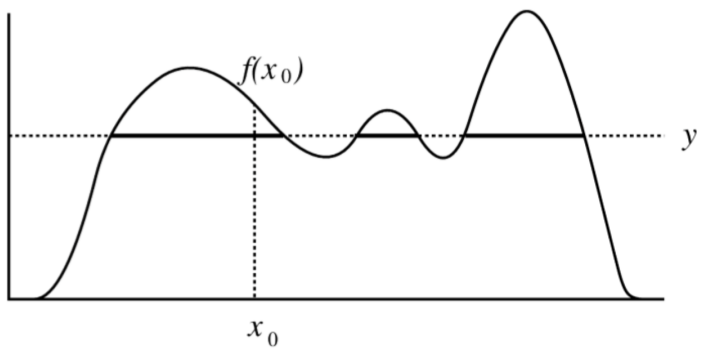
Figure: Slice sampling

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)].



Outline

Unifying HMC and slice sampling

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

- We consider **generalized HMC** with kinetic

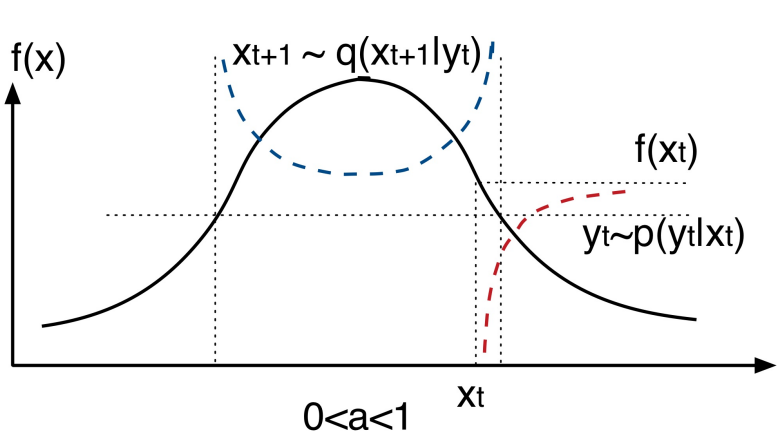
$$K(p) = |p|^{1/a}, a > 0 \tag{2}$$

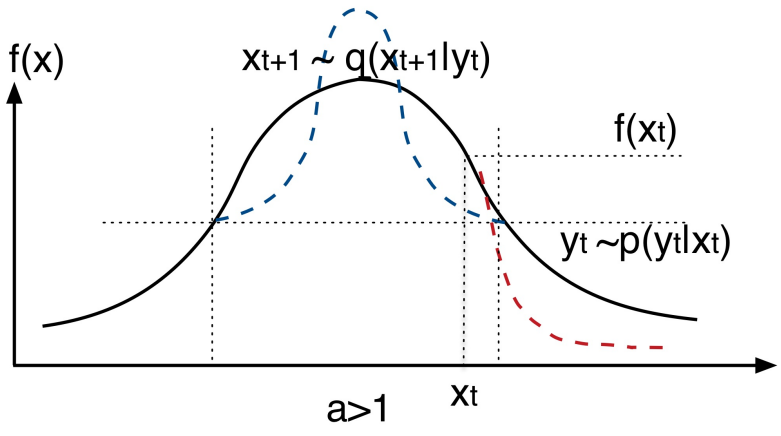
- We showed, this **generalized HMC** is indeed equivalent to a **generalized slice sampler** as following:

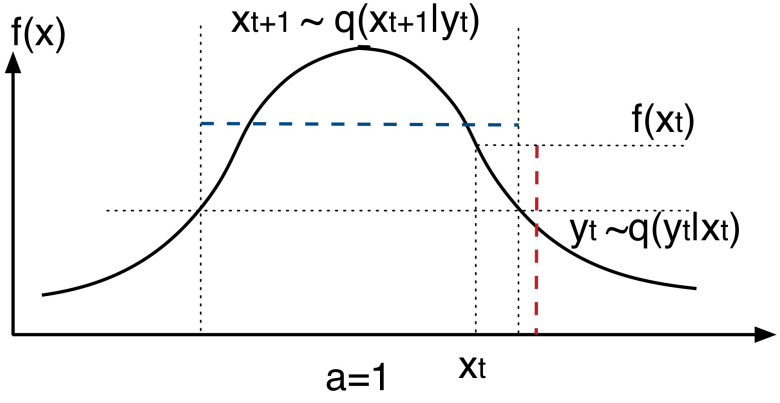
$$\text{Slicing:} p(y_t|x_t) = \frac{1}{\Gamma(a)f(x_t)}[\log f(x_t) - \log y_t]^{a-1}, \quad s.t.\, 0 < y_t < f(x_t) \tag{3}$$

$$\text{Sampling:} q(x_{t+1}|y_t) = \frac{1}{Z_2(y_t)}[\log f(x_{t+1}) - \log y_t]^{a-1} \quad s.t. f(x_t) > y_t \tag{4}$$

Unifying HMC and slice sampling (Cont'd)







a	HMC space	Canonical (SS) space
0.5	***Standard HMC***	MG-SS
1	MG-HMC	***Standard SS***
Otherwise (a>0)	MG-HMC	MG-SS

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.

Using Hamilton-Jacobi equation to solve the dynamic

- This connection between generalized HMC and generalized SS is revealed by *Hamilton-Jacobi equation* (HJE).
- In HJE, the original system (H, x, p, τ) is transformed to (H', x', p', τ), 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 τ.

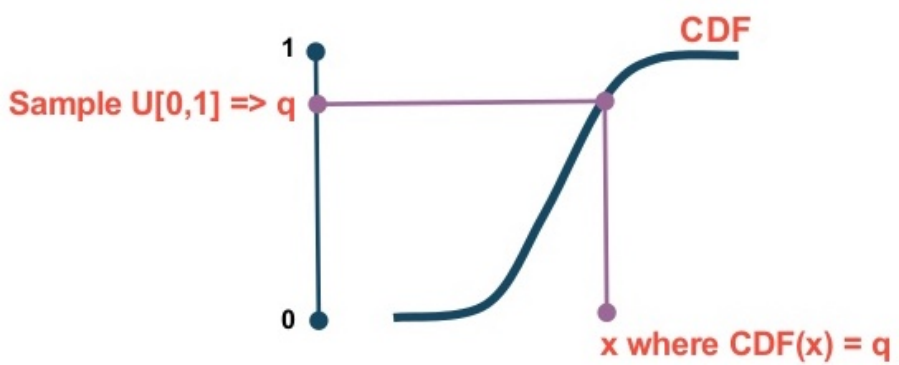
Using Hamilton-Jacobi equation to solve the dynamic

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

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

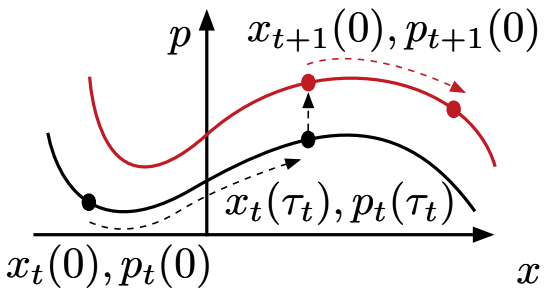
  where C is a constant, $x_{\min} = \text{argmin}_{U(x) \leq H} x$
- *(inverse transform sampling)* Uniformly sampling τ and solving x^* from (5), is equivalent to directly sampling x^* from the density:

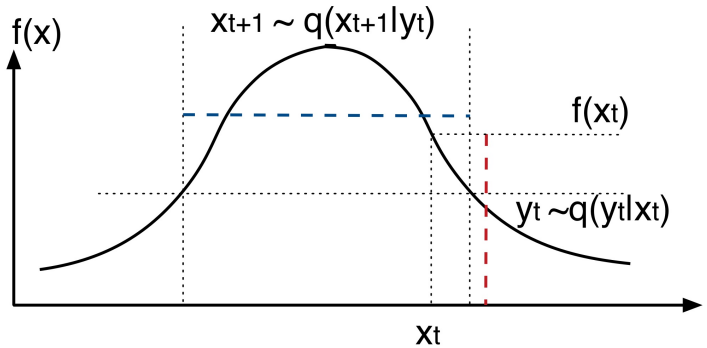
$$p(x^*|H) \propto \ [H - U(x^*)]^{a-1} \ , \quad \textit{s.t.,} \quad H - U(x^*) \geq 0 \, . \tag{6}$$



Intuitions about the connection

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





Interesting to know.. But so what?

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

Generalized HMC sampling in practice

- Generalized kinetic $K(p; m, a) = \frac{|p|^{1/a}}{m}, a, m > 0$
- *Monomial Gamma* (MG) distribution:

  $\pi(p; m, a) = \frac{m^{-a}}{2\Gamma(a+1)} e^{-\frac{|p|^{1/a}}{m}}$.
- $\text{MG}(a, m) = S \cdot G^a$ 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 $\frac{d\mathbf{x}}{dt} = \nabla_p K(\mathbf{p}), \frac{d\mathbf{p}}{dt} = -\nabla_x U(\mathbf{x})$ to get (x^*, p^*)
6: (Metropolis Hastings) accept x^* with probability $\min(1, \exp(-H(x^*, p^*) + H(x, p)))$
7: **end for**

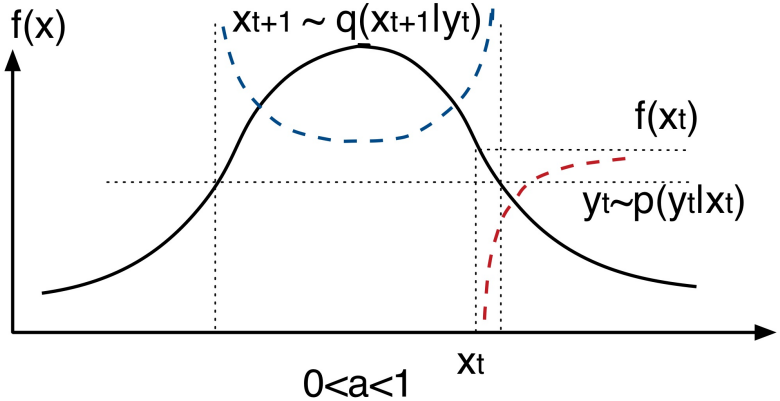
Analyzing mixing performance [autocorrelation]

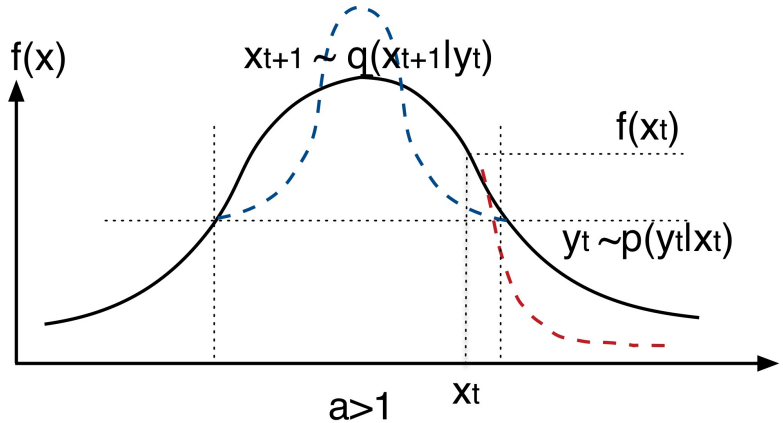
- Connection to generalized SS enables theoretical analysis for generalized HMC.
- The following theorem states that a *larger* a (heavier tail kinetics) would lead to *lower* autocorrelation during *stationary sampling period*.

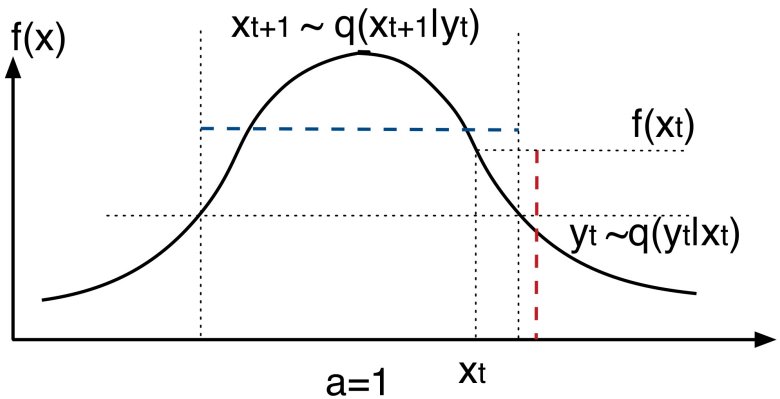
Theorem (Asymptotic autocorrelation)

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 stay close to $f(x_t)$] $\rightarrow$ [$f(x_{t+1})$ close to $f(x_t)$].







a	HMC space	Canonical (SS) space
0.5	***Standard HMC***	MG-SS
1	MG-HMC	***Standard SS***
Otherwise (a>0)	MG-HMC	MG-SS

Analyzing mixing performance [ESS and ergodicity]

- The following theorem states effective sample size (ESS) $\text{ESS} \triangleq N/(1+2\times\sum_{h=1}^{\infty}\rho(h))$ 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, $ESS\to N$

- 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.

Case study

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

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

- 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 $\mathcal{O}(a^{-1})$.

Additional advantage for sampling multimodal distribution

- 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 $a \to \infty$.

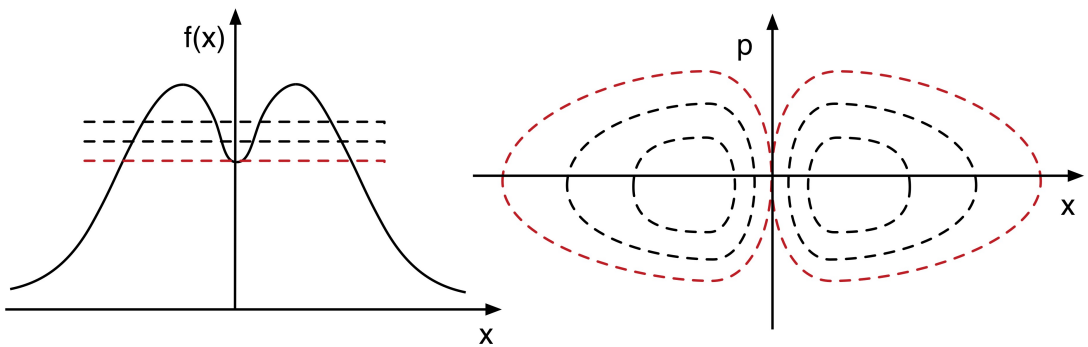


Figure: Disjoint components associated with same Hamiltonian H.

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

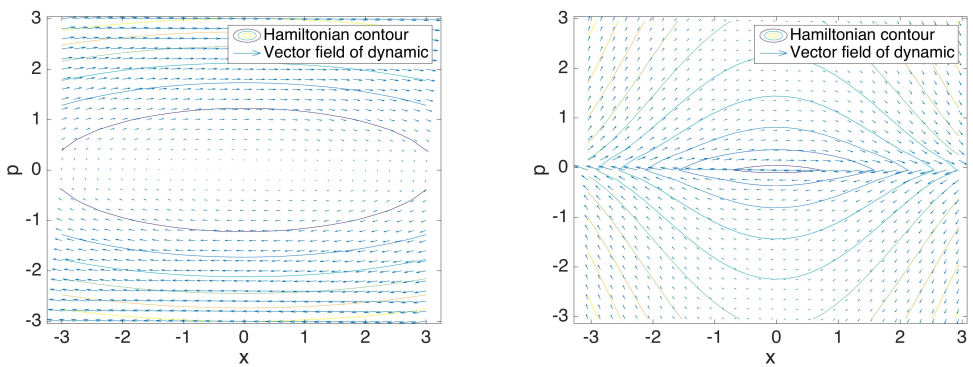


Figure: Hamiltonian contours when $a = 0.5$ and $a = 2$.

Simulation studies

- **1D unimodal problems** Univariate toy distributions $p(x) = \frac{1}{Z_1}\exp(-E(x))$, *s.t.* $x \geq 0$:
  1) **Exponential distribution** Exp(θ), where $E(x) = \theta x$.
  2) **Positive-truncated Gaussian** $\mathcal{N}_+(0, \theta)$, $E(x) = x^2$.
  3) **Gamma distributions** Gamma(r, θ), where $E(x) = -(r-1)\log x + \theta x$, where $r = 2$ and $r = 3$

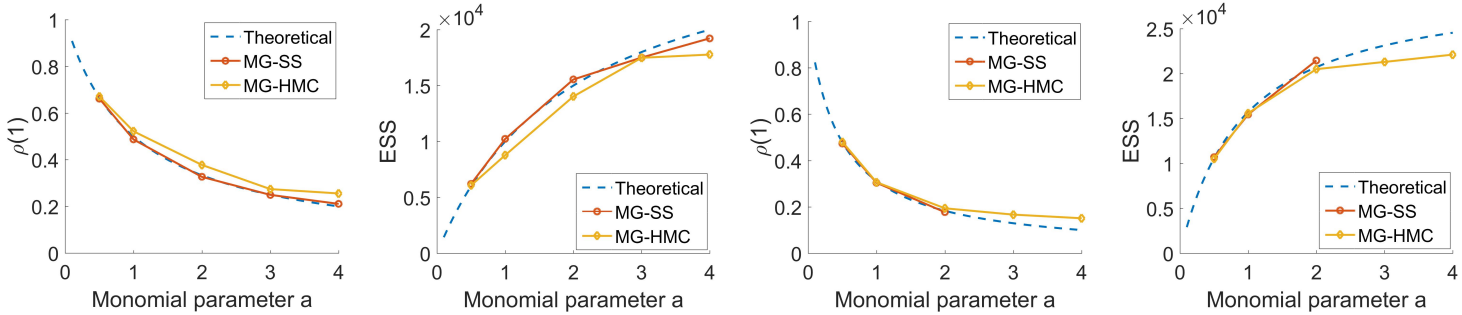


Figure: Theoretical and empirical $\rho(1)$ and ESS of exponential distribution (upper) and $\mathcal{N}_+$ (lower).

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.

	Aus(15)	Ger (25)	Hea(14)	Pim (8)	Rip (7)	Cav (87)
$a = 0.5$	3124	3447	3524	3434	3317	33
$a = 1$	**4308**	**4353**	**4591**	**4664**	**4226**	**36**
$a = 2$	1490	3646	4315	4424	1490	7

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

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

	min ESS	Time(s)	AR
$a = 0.5$	2677	525	0.98
$a = 1$	**3029**	517	0.97
$a = 2$	1534	512	0.77

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

Outline

Background: Stochastic Gradient MCMC

- 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_τ

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

where $\{\tau_1, \cdots, \tau_{N'}\}$ are random subsets.

Background: Stochastic Gradient MCMC

- Driven by a continuous-time *Markov stochastic process*.

$$d\Gamma = V(\Gamma)dt + D(\Gamma)dW\,, \tag{8}$$

- Γ denotes the parameters of the *augmented* system, *e.g.*, p and θ
- $V(\cdot)$ 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: 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

$$d\theta = \nabla_p K(p)dt \tag{9}$$

$$dp = -\nabla_\theta \tilde{U}(\theta)dt - B(\theta)\nabla_p K(p)dt + \mathcal{N}(0, 2B(\theta)dt). \tag{10}$$

- However, estimating $B(\theta)$ is difficult.

Background: Stochastic Gradient Nosé-Hoover thermostat

- **SGNHT** (stochastic gradient Nosé-Hoover thermostat) [Ding et al. (2014)] use thermostat for estimating the stochastic noise.

$$d\theta = \nabla_p K(p)dt \tag{11}$$

$$dp = -\nabla_\theta \tilde{U}(\theta)dt - \xi \nabla_p K(p)dt + \mathcal{N}(0, 2B(\theta)dt) \tag{12}$$

$$d\xi = (p^T p - 1)dt. \tag{13}$$

Outline

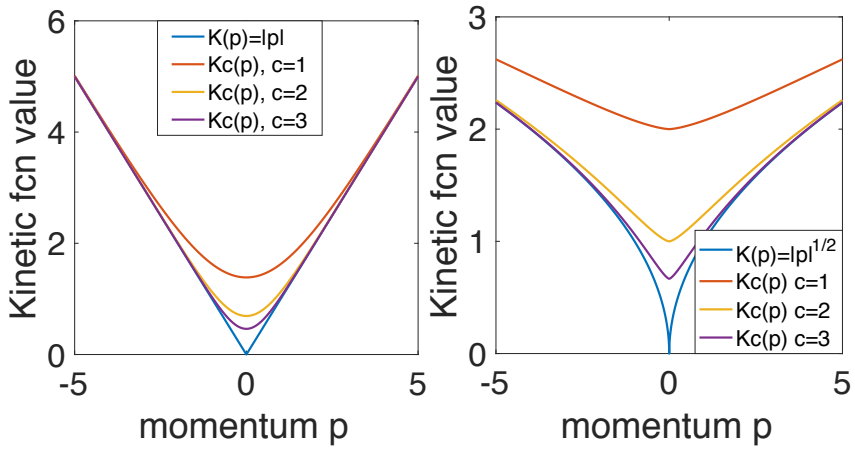
Improving over SGMCMC

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) 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.

$$H = K(p) + U(\theta) + F(\xi)\,, \tag{14}$$

- SDE under this generalized SGMCMC (denote as *SGMGT*)

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

- 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 $\rightarrow$ θ won't change.

Additional First Order Dynamics (Cont'd)

- Adding first-order dynamics to θ and ξ

$$
\begin{aligned}
d\theta =& \nabla K_c(p)dt - \sigma_\theta \nabla U(\theta)dt + \sqrt{2\sigma_\theta}dW \\
dp =& -\left(\sigma_p + \gamma \nabla F(\xi)\right) \odot \nabla K_c(p)dt \\
& - \nabla U(\theta)dt + \sqrt{2\sigma_p}dW, \\
d\xi =& \gamma \left[\nabla K_c(p) \odot \nabla K_c(p) - \nabla^2 K_c(p)\right] dt \\
& - \sigma_\xi \nabla F(\xi)dt + \sqrt{2\sigma_\xi}dW \,. \qquad (19)
\end{aligned}
$$

- 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

Stochastic resampling

- Resample p and ξ 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*

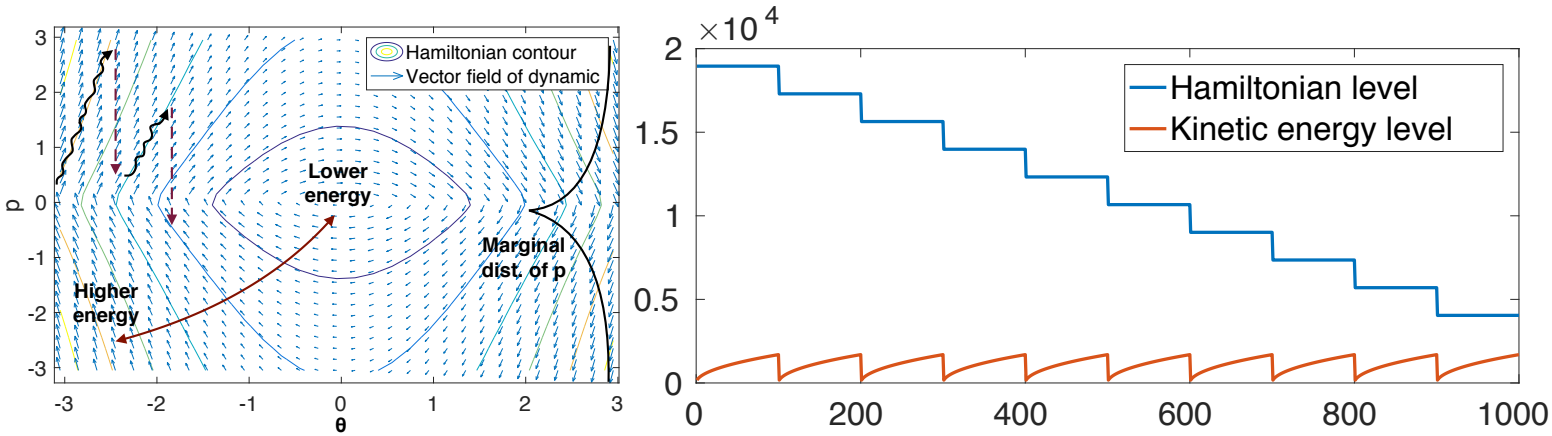


Figure: Stochastic resampling.

Theoretical properties

- Quantifying how fast the sample average, $\hat{\phi}_T$, converges to the true posterior average, $\bar{\phi} \triangleq \int \phi(\theta)\pi(\theta|X)\mathrm{d}\theta$, 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:

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

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

Outline

Multiple-well Synthetic Potential

- Generated samples has better stationary mixing.

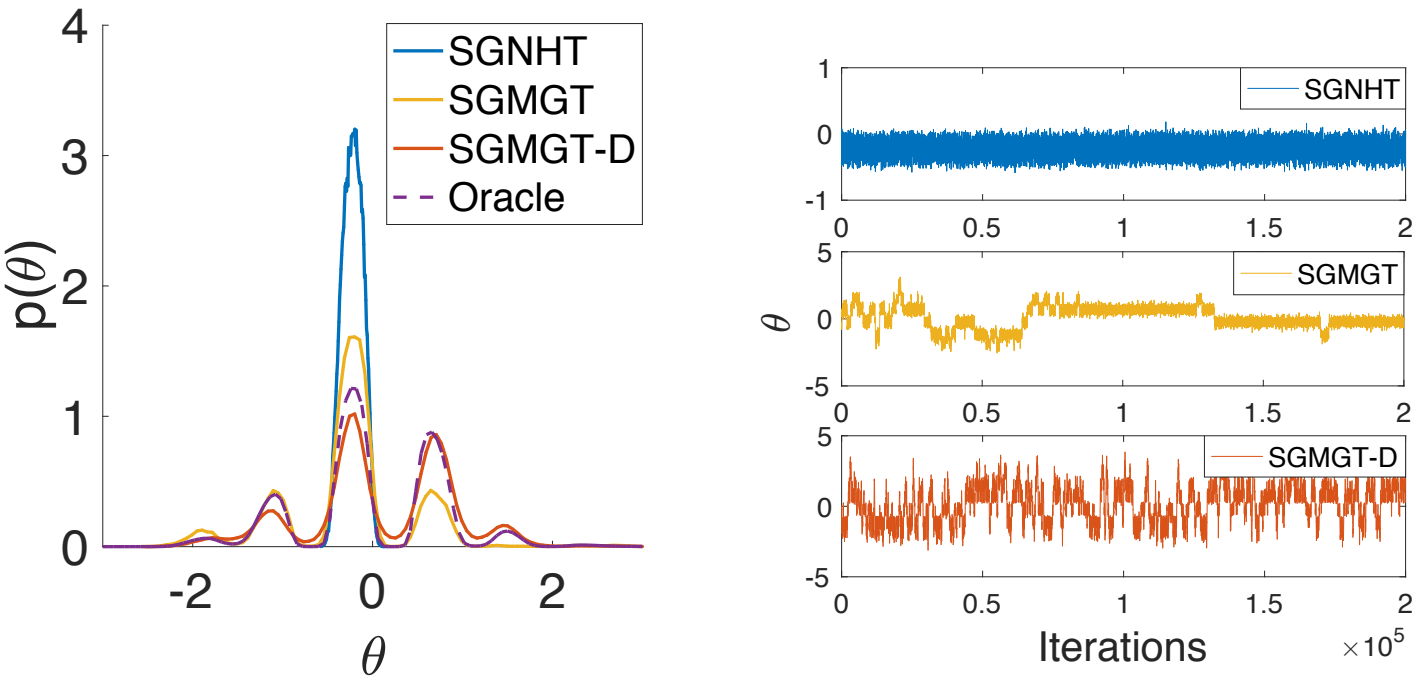


Figure: Synthetic multimodal distribution. Left: empirical distributions for different methods. Right: traceplot for each method.

Bayesian Logistic Regression

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

AUROC (D)	A (15)	G (25)	H (14)	P(8)	R (7)	C (87)
SGNHT	0.89	0.75	0.90	0.86	0.95	0.65
SGMGT(a=1)	0.92	0.78	0.91	0.86	0.87	0.70
SGMGT-D(a=1)	0.95	0.86	0.95	**0.93**	**0.98**	**0.73**
SGMGT(a=2)	0.93	0.79	0.93	0.88	0.86	0.62
SGMGT-D(a=2)	**0.95**	**0.90**	**0.95**	0.90	0.97	0.69
ESS (D)	A (15)	G (25)	H (14)	P(8)	R (7)	C (87)
SGNHT	869	941	1911	2077	1761	1873
SGMGT-D(a=1)	**3147**	**2131**	2448	**4244**	1494	**3605**
SGMGT-D(a=2)	2700	1989	**2768**	3430	**2265**	2969

Discriminative RBM

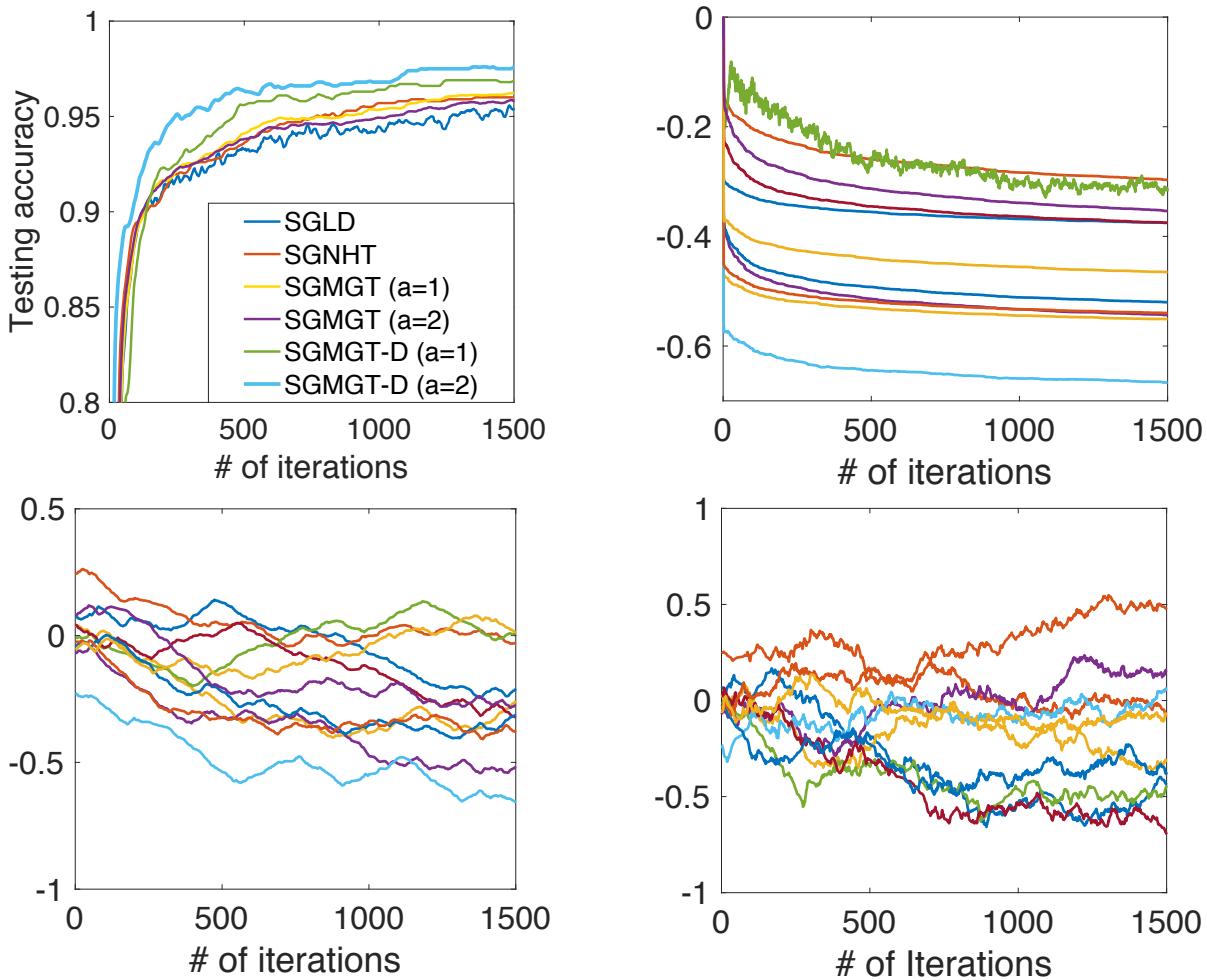


Figure: Left: testing accuracies for SGLD, SGNHT, SGMGT and SGMGT-D. Middle-left through right: traceplots for SGLD, SGNHT and SGMGT-D.

Recurrent Neural Network

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

Algorithms	Piano	Nott	Muse	JSB	PTB
SGLD	11.37	6.07	10.83	11.25	127.47
SGNHT	9.00	4.24	7.85	9.27	131.3
SGMGT (a=1)	7.90	4.35	8.42	8.67	120.6
SGMGT (a=2)	10.17	4.64	8.51	8.84	250.5
SGMGT-D (a=1)	**7.51**	**3.33**	7.11	8.46	113.8
SGMGT-D (a=2)	7.53	3.35	**7.09**	**8.43**	**109.0**
SGD	11.13	5.26	10.08	10.81	120.44
RMSprop	7.70	3.48	7.22	8.52	120.45

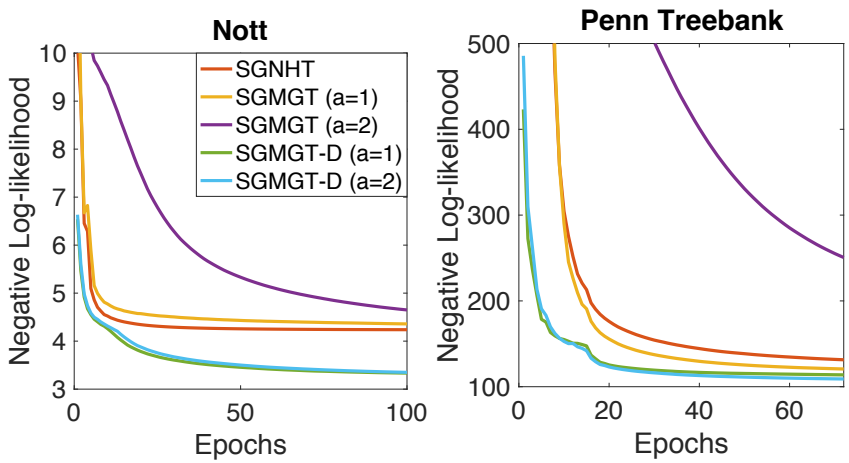


Figure: Learning curves of different SG-MCMC methods.

Take-aways

Conclusion:

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

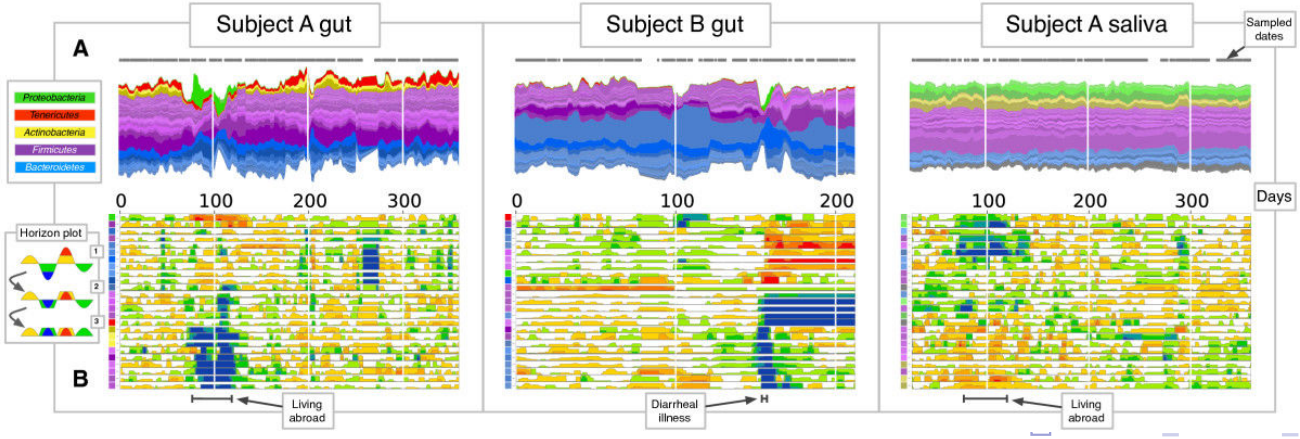
Future research:

- Adaptive selection of monomial parameters.
- Connection to optimization methods.

Outline

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 $K = 50$.



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

DPFA Model: emission

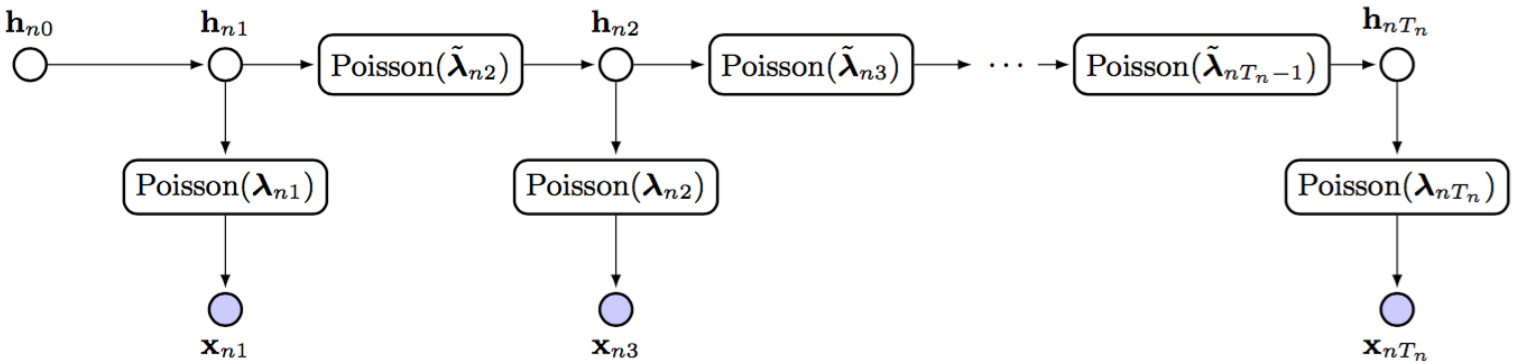
- **Emission model:** Poisson factor model [Henao et al. (2015)].

$$\mathbf{x}_{nt} \sim \text{Poisson}\left(\boldsymbol{\Psi}(\boldsymbol{\theta}_{nt} \circ \mathbf{h}_{nt})\right) , \tag{20}$$

- We specify prior distributions as,

$$\boldsymbol{\psi}_k \sim \text{Dirichlet}(\eta_\psi \mathbf{1}_M) , \theta_{knt} \sim \text{Gamma}(r_k, b_\theta) , \tag{21}$$

$$h_{knt} \sim \text{Bernoulli}(\pi_{knt}) , \tag{22}$$



DPFA Model: emission

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

$$\boldsymbol{h}_{nt} = 1\left(\boldsymbol{z}_{nt} > 0\right), \boldsymbol{z}_{nt} \sim \text{Poisson}\left(\tilde{\boldsymbol{\lambda}}_{nt}\right), \tag{23}$$

$$\tilde{\boldsymbol{\lambda}}_{nt} = \tau_{nt}^{-1}\boldsymbol{\Phi}(\mathbf{w}_{nt-1} \circ \mathbf{h}_{nt-1}) + \tilde{\boldsymbol{\lambda}}_0 \tag{24}$$

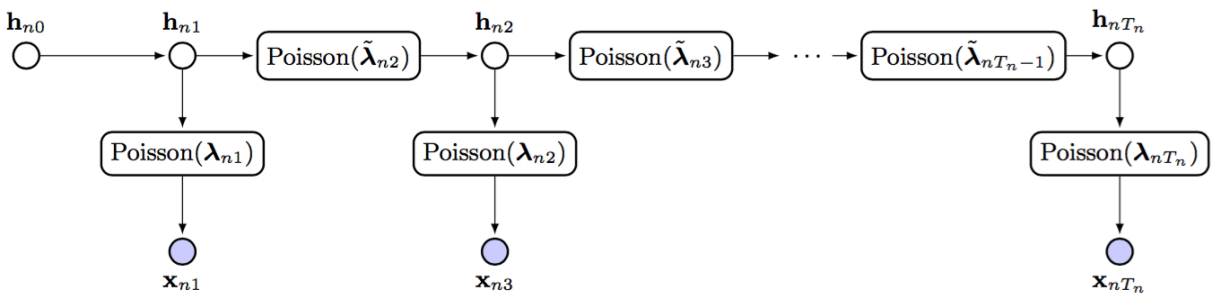
- Equivalently,

$$p(\boldsymbol{h}_{nt} = 1) = \text{Bernoulli}\left(1 - \exp(-\tilde{\boldsymbol{\lambda}}_{nt})\right),$$

- We specify prior distributions as,

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}\left(\eta_\phi \mathbf{1}_K\right), w_{knt-1} \sim \text{Gamma}(s_k, b_w), \tag{25}$$

- Sensitive to existence (vs abundance); inference conveniency;



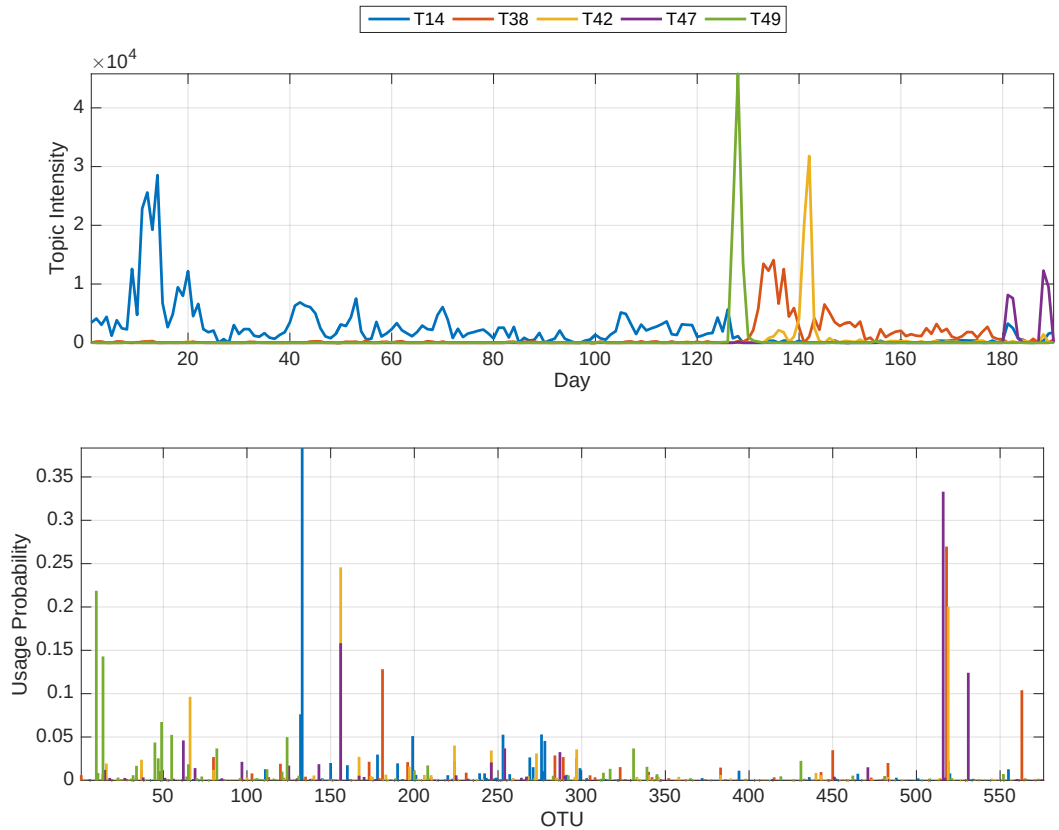
SGMGT-embedded Gibbs Inference

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

$$\pi(\boldsymbol{\Theta}|\mathbf{h}, \mathbf{z}) \propto p(\boldsymbol{\Theta})p(\mathbf{x}|\mathbf{h}, \boldsymbol{\Theta})p(\mathbf{h}, \mathbf{z}|\boldsymbol{\Theta}).$$

- 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

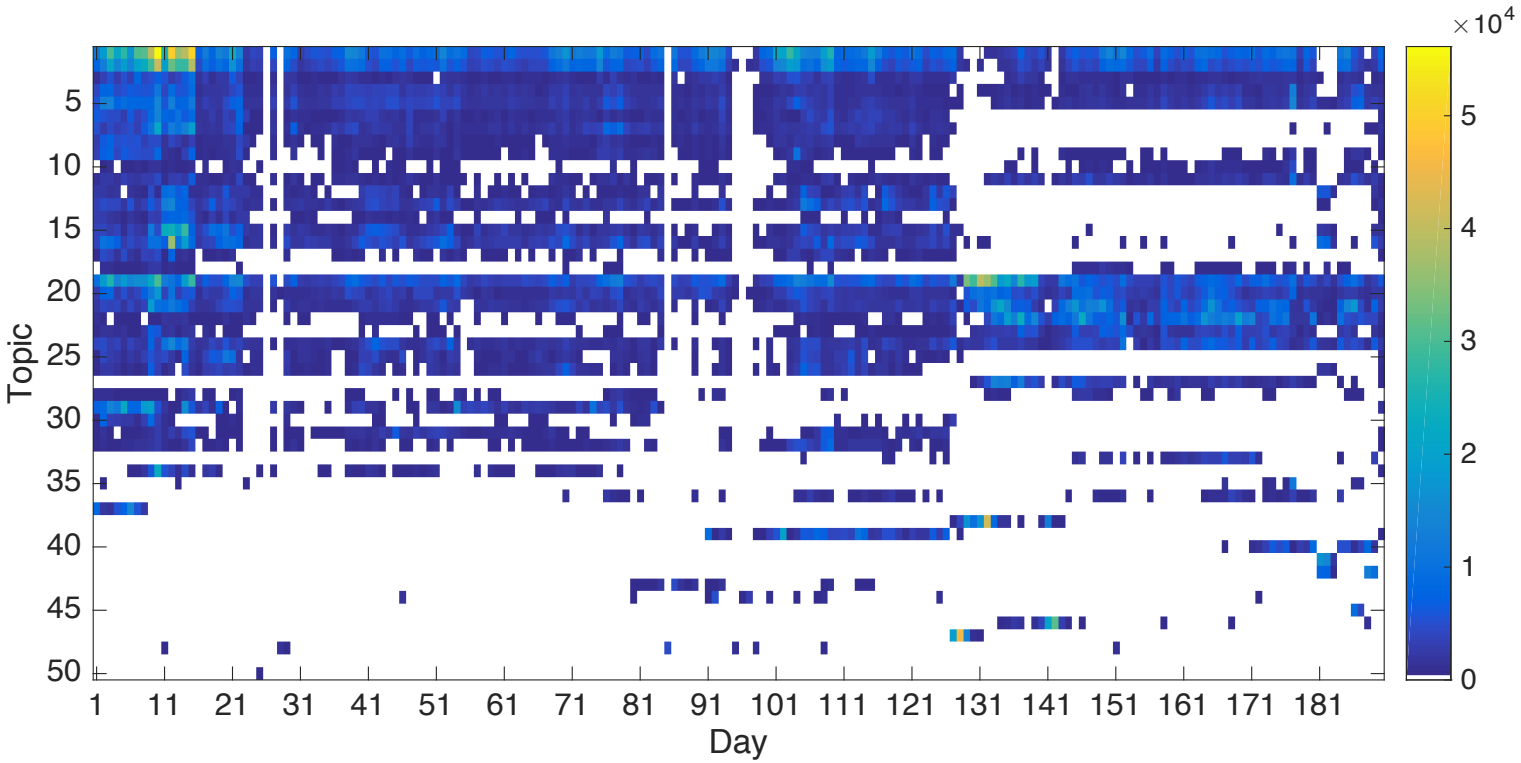


Figure: Intensity heatmap for microbiome data with 50 topics (y axis). The x axis represents time in days.

Quantitative analysis

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

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

- 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

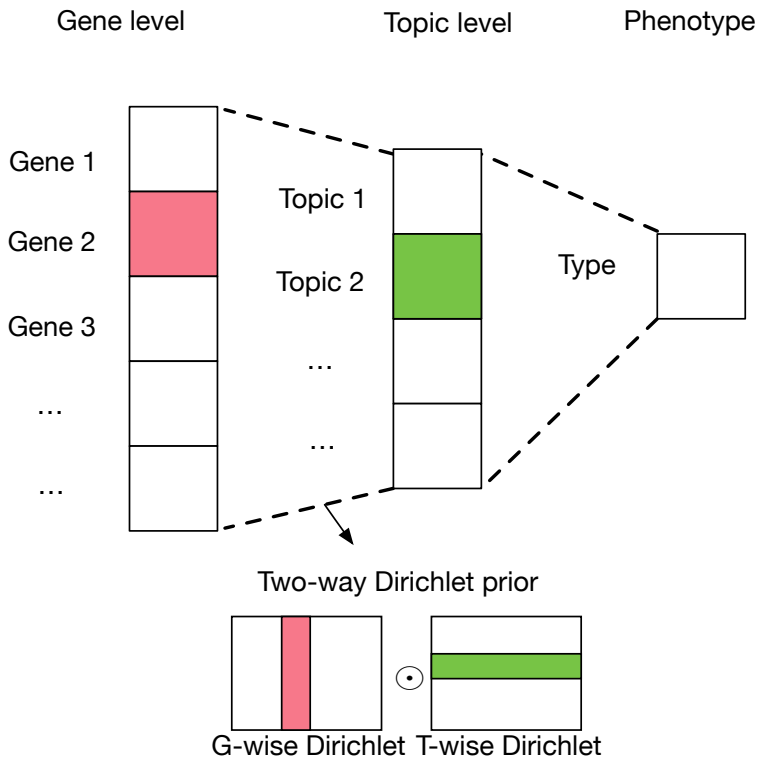
Motivation & Data

- **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 $d_i \in \mathbf{R}^V$ into a topic vector $h_i \in \mathbf{R}^K$.
- 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.

Two-way Dirichlet prior for neural topic loading



$$h_i = \sigma(W d_i),$$
$$y_i \sim \text{softmax}(U h_i),$$
$$W, U \sim p_W(\cdot), p_U(\cdot)$$

Two-way Dirichlet prior for neural topic loading

- 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.

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

- V is specified to have a *masked* two-way Dirichlet distribution.

$$p(V_{nl}) \propto C_{nl} D_{nl} M_{nl} \tag{32}$$
$$C_n \sim Dir(\gamma) \tag{33}$$
$$D_l \sim Dir(\eta) \tag{34}$$

- $M \in \{0, 1\}^{N \times L}$ indicates whether n-th isoform is from l-th gene.

Isoform composition inference

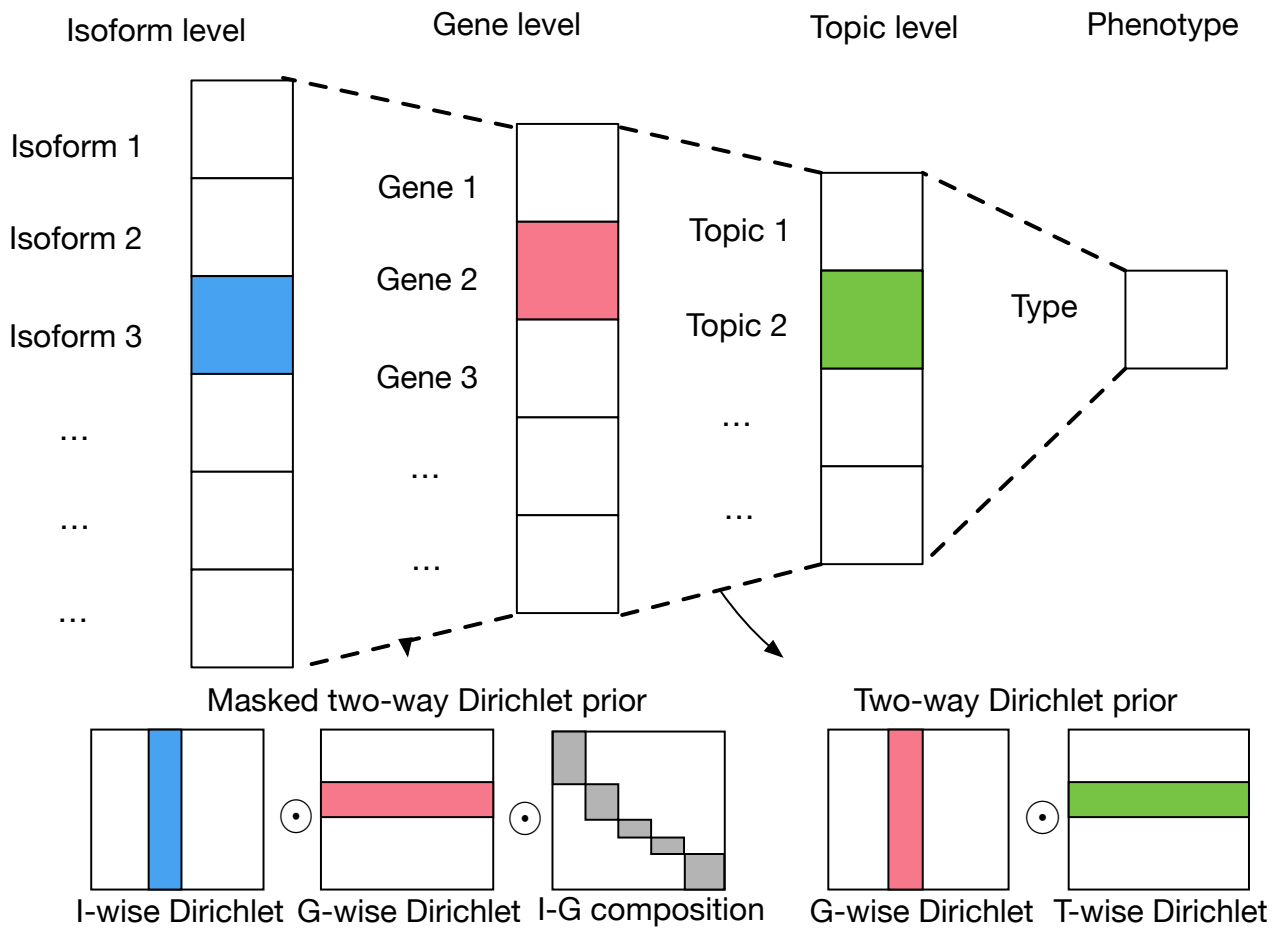


Figure: Isoform composition model

Inference details

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

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

  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.

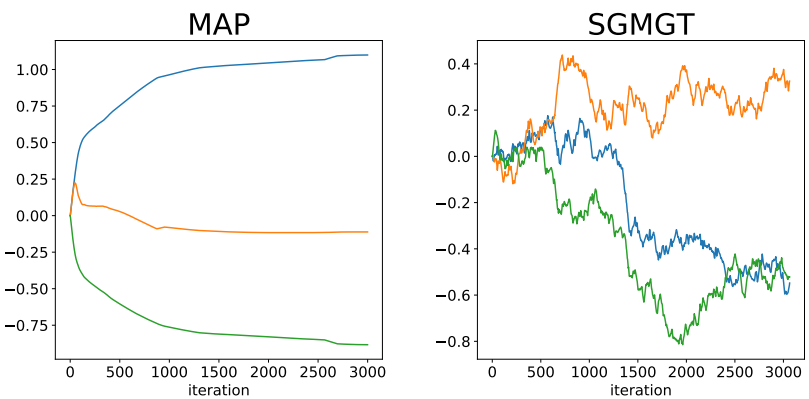
Results
Prediction

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

Model	Group Lasso	Ours (Two-way Dir)
Gene-level model	0.187±0.050	**0.162±0.083**
Isoform -level model	0.177 ± 0.066	**0.149±0.075**

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

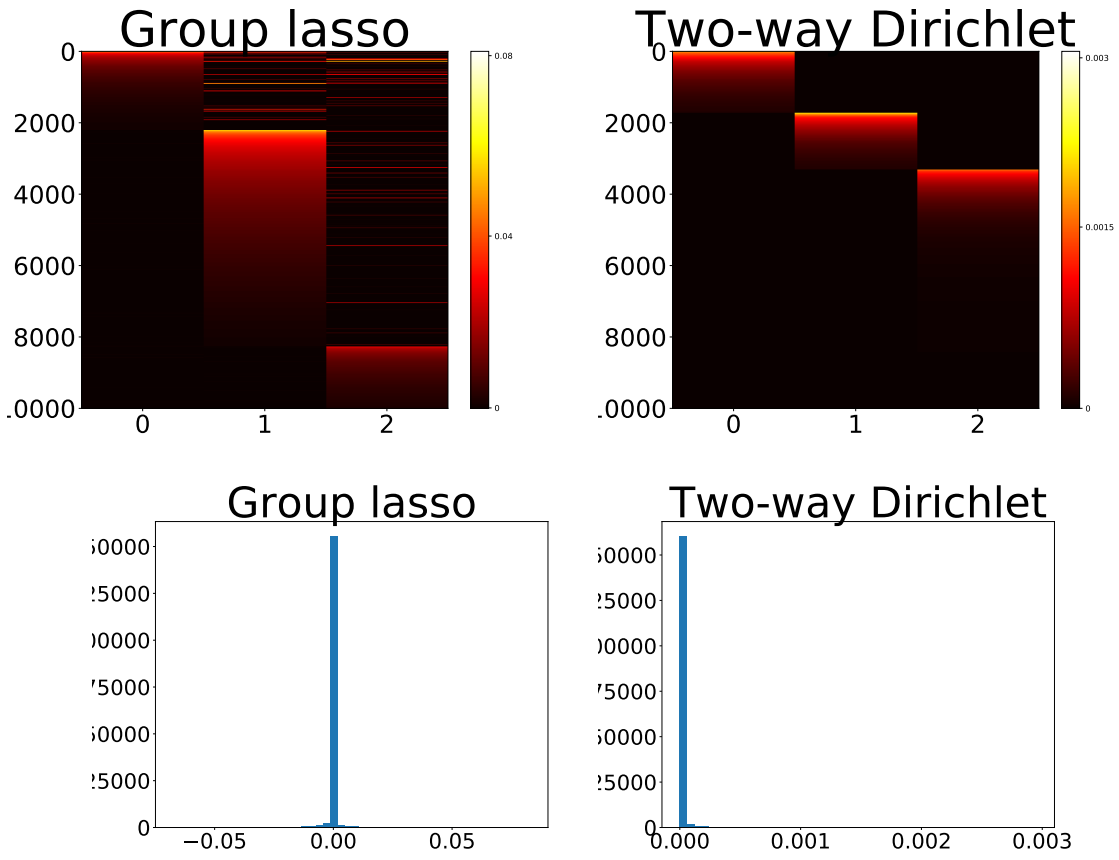
- Traceplot of weight parameters.



Results

Non-overlapping sparse topic loading

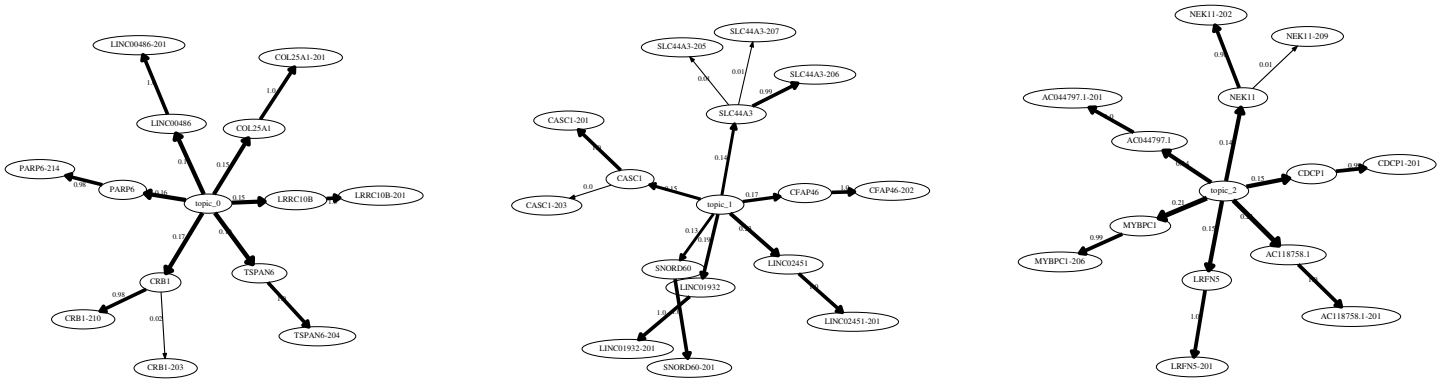
- Two-way Dirichlet prior shows non-negative, non-overlapping and sparse topic intensity.



Results

Interpretable topic inference

- Identified 3 topics correspond to 3 infection types.



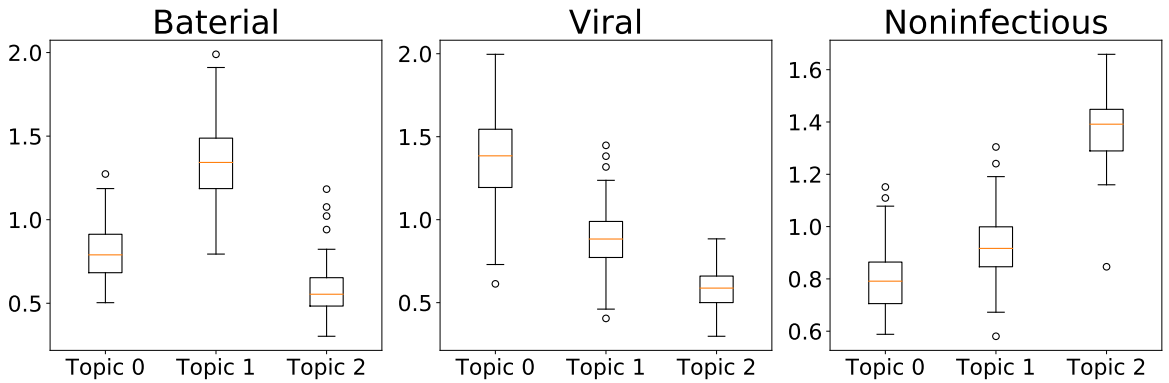


Figure: Topic intensity level for each group of infection type

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!

References I


Chen, Tianqi, Emily B Fox, and Carlos Guestrin (2014). "Stochastic Gradient Hamiltonian Monte Carlo". In: *ArXiv*.


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


Ding, Nan et al. (2014). "Bayesian sampling using stochastic gradient thermostats". In: *Neural Information Processing Systems*.


Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "A note on the group lasso and a sparse group lasso". In: *arXiv preprint arXiv:1001.0736*.


Girolami, Mark and Ben Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2.

References II

- Henao, Ricardo et al. (2015). "Deep Poisson Factor Modeling". In: *Neural Information Processing Systems*.
- Neal, Radford M (2003). "Slice sampling". In: *Annals of statistics*.
- Nishimura, Akihiko and David Dunson (2016a). "Geometrically Tempered Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1604.00872*.
- – (2016b). "Variable length trajectory compressible hybrid Monte Carlo". In: *arXiv preprint arXiv:1604.00889*.
- Nishimura, Akihiko, David Dunson, and Jianfeng Lu (2017). "Discontinuous Hamiltonian Monte Carlo for sampling discrete parameters". In: *arXiv preprint arXiv:1705.08510*.
- Roberts, Gareth O and Richard L Tweedie (1996). "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli*.

References III


Vigário, Ricardo et al. (1998). "Independent component analysis for identification of artifacts in magnetoencephalographic recordings". In: *NIPS*.


Zhou, Mingyuan (2015). "Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction". In: *Artificial Intelligence and Statistics Conference*.