A Complete Recipe for Stochastic Gradient MCMC

-Summary-

Suggested SDE

- Consider the following SDE: ($z \in \mathbb{R}^d$)

$$dz = f(z)dt + \sqrt{2D(z)}dW(t)$$

where $f(z)$: deterministic drift, $W(t)$: d-dimensional Brownian motion, $D(z)$: P.S.D diffusion matrix

- **[Idea]** Set $f(z) = -[D(z) + Q(z)]\nabla H(z) + \Gamma(z)$, where $\Gamma(z)_i \coloneqq \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right)$

where $Q(z)$ is skew-symmetric ($\Leftrightarrow Q^T = -Q$)

Theorems related to given SDE

- **[Theorem]** $p^s(z) \propto \exp(-H(z))$ is a stationary distribution of the given dynamics if $f(z)$ is restricted to $f(z) = -[D(z) + Q(z)]\nabla H(z) + \Gamma(z)$ with $D(z)$: P.S.D, $Q(z)$: skew-symmetric. (Furthermore, Ii $D(z)$ is P.D or ergodicity can be shown, then $p^s(z)$ is unique)

- Proof sketch:
  - By Fokker-Planck description of the dynamics, it follows that:

$$\partial_t p(\mathbf{z}, t) = -\sum_i \frac{\partial}{\partial \mathbf{z}_i}\big(\mathbf{f}_i(\mathbf{z})p(\mathbf{z}, t)\big) + \sum_{i,j} \frac{\partial^2}{\partial \mathbf{z}_i \partial \mathbf{z}_j}\big(\mathbf{D}_{ij}(\mathbf{z})p(\mathbf{z}, t)\big).$$

Theorems related to given SDE

- Proof sketch:
  - By Fokker-Planck description of the dynamics, it follows that:

$$\partial_t p(\mathbf{z}, t) = -\sum_i \frac{\partial}{\partial \mathbf{z}_i}\big(\mathbf{f}_i(\mathbf{z}) p(\mathbf{z}, t)\big) + \sum_{i,j} \frac{\partial^2}{\partial \mathbf{z}_i \partial \mathbf{z}_j}\big(\mathbf{D}_{ij}(\mathbf{z}) p(\mathbf{z}, t)\big).$$

  - When Q is skew-symmetric, the following holds:

$$\partial_t p(\mathbf{z}, t) = \nabla^T \cdot \Big([\mathbf{D}(\mathbf{z}) + \mathbf{Q}(\mathbf{z})] \left[p(\mathbf{z}, t) \nabla H(\mathbf{z}) + \nabla p(\mathbf{z}, t)\right] \Big).$$

  - Note that $p(\boldsymbol{z}, t)\nabla H(\boldsymbol{z}) + p(\boldsymbol{z}, t) = 0$ when $p(z) \propto \exp\big(-H(z)\big)$, which proves the stationary of target distribution.

Completeness of the framework

- Question: what portion of samplers defined by continuous Markov processes with the target invariant distribution can we define by given SDE with certain $D(z)$ and $Q(z)$?

- By chapman-Kolmogorov equation, any continuous Markov process with stationary distribution $p^s(z)$ can be described by SDE: (which determines $D(z)$.)

$$dz = f(z)dt + \sqrt{2D(z)}dW(t)$$

Completeness of the framework

- **[Theorem]** Suppose $p^s(z)$ uniquely exists, and that $f_i(z)p^s(z) - \sum_{j=1}^{d} \frac{\partial}{\partial \theta_j}\left(D_{ij}(z)p^s(z)\right)$ is integrable with respect to Lebesgue measure, the there exists a skew-symmetric $Q(z)$ such that $f(z) = -[D(z) + Q(z)]\nabla H(z) + \Gamma(z)$, where $\Gamma(z)_i \coloneqq \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right)$ holds.

- This theorem implies that there exists a bijection between the set of all continuous Markov processes with $p^s(z) \propto \exp\left(-H(z)\right)$ and the SDE representation of $dz = f(z)dt + \sqrt{2D(z)}dW(t)$, where $f(z) = -[D(z) + Q(z)]\nabla H(z) + \Gamma(z)$.

Algorithm for generic SGMCMC

- To realize the continuous SDE, use ϵ-discretization (Full-data update version):

$$z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla H(z_t) + \Gamma(z_t)\right] + N\left(0, 2\epsilon_t D(z_t)\right)$$

- As we did in SGLD, SGHMC, use approximation (unbiased estimate) of $U(\theta)$:

$$\widetilde{U}(\theta) = -\frac{|\mathcal{S}|}{|\hat{\mathcal{S}}|}\sum_{x\in\hat{\mathcal{S}}} \log p(x|\theta) - \log p(\theta)$$

- Now, we should consider noise from stochastic gradient. From the central limit theorem, assume $\nabla\widetilde{U}(\theta) = \nabla U(\theta) + N\left(0, V(\theta)\right)$, which results $\nabla\widetilde{H}(z) = \nabla H(z) + \left[N\left(0, V(\theta)\right), 0\right]^T$ (Assuming $z = [\theta, r]$)

Algorithm for generic SGMCMC

- Then, the stochastic gradient variant of the above sampler becomes as follows:

$$z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla\widetilde{H}(z_t) + \Gamma(z_t)\right] + N\left(0, \epsilon_t\left(2D(z_t) - \epsilon_t\hat{B}_t\right)\right)$$

  where $\hat{B}_t$ is the estimate of the variance of $\left(D(z_t) + Q(z_t)\right)N\left(0, V(\theta)\right)$ with a condition $2D(z_t) - \epsilon_t\hat{B}_t \succcurlyeq 0$.

- Note that as $\epsilon_t^2 \to 0$ faster than ϵ_t, the discrepancy induced by estimate $\hat{B}_t$ approaches zero as $\epsilon_t \to 0$.

Applying the theory to construct samplers

<HMC>

- The discrete Hamiltonian dynamics used on HMC:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t M^{-1} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla U(\theta_t) \end{cases}$$

  where θ = position, r = momentum, M = mass / environment = frictionless surface

- To match HMC with suggested framework, set $z = (\theta, r), H(\theta, r) = U(\theta) + \frac{1}{2} r^T M^{-1} r$, and

  $Q(\theta, r) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$ and $D(\theta, r\,) = \mathbf{0}$.

Applying the theory to construct samplers

[Double check]:

Theory : $z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla H(z_t) + \Gamma(z_t)\right] + N\left(0{,}2\epsilon_t D(z_t)\right)$

Note: $Q(\theta, r) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$ and $D(\theta, r\ ) = \mathbf{0}$.

1. $\nabla H(z) = [\nabla U(\theta)^T, M^{-1}r]^T$ and $\Gamma(z)_i\ =\ \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right) = 0$

2. $\left(D(z) + Q(z)\right)\nabla H(z) + \Gamma(z) = \begin{pmatrix} -M^{-1}r \\ \nabla U(\theta) \end{pmatrix} + 0$

3. $N\left(0{,}2\epsilon D(z)\right) = 0$

Applying the theory to construct samplers

<SGHMC>

- The discrete dynamics used on Naïve-SGHMC:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t M^{-1} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla U(\theta_t) + N(0, \epsilon_t^2 V(\theta_t) \end{cases}$$

- Note that the above equation cannot be converted to the suggested theory!, which means the target distribution is not stationary.

- This is the reason we are required to impose friction term C to achieve stationary target distribution.

Applying the theory to construct samplers

<SGHMC>

- The discrete 2nd order Langevin dynamics used on SGHMC (w/ friction term C):

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t M^{-1} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla \widetilde{U}(\theta_t) - \epsilon_t C M^{-1} r_t + N\left(0, \epsilon_t\left(2C - \epsilon_t \hat{B}_t\right)\right) \end{cases}$$

where $\hat{B}_t$ is an estimate of $V(\theta_t)$.

- To match HMC with suggested framework, set $z = (\theta, r), H(\theta, r) = U(\theta) + \frac{1}{2} r^T M^{-1} r$, and

$Q(\theta, r) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$ and $D(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$.

Applying the theory to construct samplers

[Double check]:

Theory : $z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla\widetilde{H}(z_t) + \Gamma(z_t)\right] + N\left(0, \epsilon_t\left(2D(z_t) - \epsilon_t\hat{B}_{ext,t}\right)\right)$

Note: $Q(\theta, r) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$ and $D(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$.

1. $\nabla\widetilde{H}(z) = \left[\nabla\widetilde{U}(\theta)^T,\ M^{-1}r\right]^T$ and $\Gamma(z)_i = \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right) = 0$

2. $\left(D(z) + Q(z)\right)\nabla\widetilde{H}(z) + \Gamma(z) = \begin{pmatrix} -M^{-1}r \\ \nabla\widetilde{U}(\theta) + CM^{-1}r \end{pmatrix} + 0$

3. $N\left(0, \epsilon\left(2D(z) - \epsilon\hat{B}_{ext}\right)\right) = N\left(0, \epsilon\left(2C - \epsilon\hat{B}\right)\right)$ *[dimension reduction]*

Applying the theory to construct samplers

<SGLD>

- The discrete 1st order Langevin dynamics used on SGLD:

$$\theta_{t+1} \leftarrow \theta_t - \epsilon_t D \nabla \widetilde{U}(\theta_t) + N(0, 2\epsilon_t D)$$

- To match HMC with suggested framework, set $z = \theta, H(\theta) = U(\theta)$, and $Q(\theta) = 0$ and $D(\theta) = D$ and $\hat{B}_t = 0$.

Applying the theory to construct samplers

[Double check]:

Theory : $z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla\widetilde{H}(z_t) + \Gamma(z_t)\right] + N\left(0, \epsilon_t\left(2D(z_t) - \epsilon_t\hat{B}_{ext,t}\right)\right)$

Note: $Q(\theta) = 0$ and $D(\theta) = D$.

1. $\nabla\widetilde{H}(z) = \nabla\widetilde{U}(\theta)$ and $\Gamma(z)_i = \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right) = 0$

2. $\left(D(z) + Q(z)\right)\nabla\widetilde{H}(z) + \Gamma(z) = D\nabla\widetilde{U}(\theta) + 0$

3. $N\left(0, \epsilon\left(2D(z) - \epsilon\hat{B}_{ext}\right)\right) = N(0{,}2\epsilon D)$

Applying the theory to construct samplers

<SGRLD (Stochastic Gradient Riemannian Langevin Dynamics)>

- It is a generalized version of SGLD by adopting adaptive diffusion matrix $D(\theta) = G^{-1}(\theta)$,

  where $G(\theta)_{ij} = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta_i}\log p(x|\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log p(x|\theta)\right)\right]$ is the fisher information matrix.

- The discrete dynamics used on SGRLD:

$$\theta_{t+1} \leftarrow \theta_t - \epsilon_t\left[G(\theta_t)^{-1}\nabla\widetilde{U}(\theta_t) + \Gamma(\theta_t)\right] + N(0,2\epsilon_t G(\theta_t)^{-1})$$

  where $\Gamma(\theta)_i = \sum_{j=1}^{d}\frac{\partial D_{ij}(\theta)}{\partial\theta_j}$

- To match HMC with suggested framework, set $z = \theta, H(\theta) = U(\theta)$, and $Q(\theta) = 0$ and $D(\theta) = G^{-1}(\theta)$ and $\hat{B}_t = 0$.

Applying the theory to construct samplers

[Double check]:

Theory : $z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla\widetilde{H}(z_t) + \Gamma(z_t)\right] + N\left(0, \epsilon_t\left(2D(z_t) - \epsilon_t\hat{B}_{ext,t}\right)\right)$

Note: $Q(\theta) = 0$ and $D(\theta) = G^{-1}(\theta)$.

1. $\nabla\widetilde{H}(z) = \nabla\widetilde{U}(\theta)$ and $\Gamma(z)_i = \sum_{j=1}^{d} \frac{\partial}{\partial z_j}\left(D_{ij}(z) + Q_{ij}(z)\right) = \sum_{j=1}^{d} \frac{\partial G(\theta)^{-1}}{\partial \theta_j}$

2. $\left(D(z) + Q(z)\right)\nabla\widetilde{H}(z) + \Gamma z = G^{-1}(\theta)\nabla\widetilde{U}(\theta) + \Gamma(\theta)$

3. $N\left(0, \epsilon\left(2D(z) - \epsilon\hat{B}_{ext}\right)\right) = N(0,2\epsilon G(\theta)^{-1})$

Applying the theory to construct samplers

<SGNHT (Stochastic Gradient Nose-Hoover Thermostat)>

- It is augmented version of SGHMC with additional scalar variable ζ.
- The discrete dynamics used on SGNHT:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \nabla \widetilde{U}(\theta_t) - \epsilon_t \zeta_t r_t + N\left(0, \epsilon_t\left(2A \cdot I - \epsilon_t\, \hat{B}_t\right)\right) \\ \zeta_{t+1} \leftarrow \zeta_t + \epsilon_t \left(\frac{1}{d} r_t^T r_t - 1\right) \end{cases}$$

- To match HMC with suggested framework, set $z = (\theta, r, \zeta)$, $H(\theta, r, \zeta) = U(\theta) - \frac{1}{2} r^T r + \frac{d}{2}(\zeta + A)^2$, and

$$Q(\theta, r, \zeta) = \begin{pmatrix} 0 & +I & 0 \\ I & 0 & r/d \\ 0 & +r^T/d & 0 \end{pmatrix} \text{ and } D(\theta, r, \zeta) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & A \cdot I & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ where } \theta, r \in \mathbb{R}^d, \zeta \in \mathbb{R}$$

Applying the theory to construct samplers

[Double check]:

Theory : $z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla\widetilde{H}(z_t) + \Gamma(z_t)\right] + N\left(0, \epsilon_t\left(2D(z_t) - \epsilon_t\hat{B}_{ext,t}\right)\right)$

Note: $Q(\theta, r, \zeta) = \begin{pmatrix} 0 & +I & 0 \\ I & 0 & r/d \\ 0 & +r^T/d & 0 \end{pmatrix}$ and $D(\theta, r, \zeta) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & A \cdot I & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

1. $\nabla\widetilde{H}(z) = \left[\nabla\widetilde{U}(\theta)^T, \ -r^T, d(\zeta + A)\right]^T$ and $\Gamma(z) = [0,0,1]^T$

2. $\left(D(z) + Q(z)\right)\nabla\widetilde{H}(z) + \Gamma(z) = \left[-r^T, \ \nabla\,\widetilde{U}(\theta)^T - A \cdot r^T + r^T(\zeta + A), \ \ -\frac{r^T r}{d}\right]^T + [\mathbf{0}, \mathbf{0}, 1]^T$

3. $N\left(0, \epsilon\left(2D(z) - \epsilon\hat{B}_{ext}\right)\right) = N\left(0, \epsilon\left(2A \cdot I \ - \epsilon\hat{B}\right)\right)$

Devising new samplers

<SGRHMC (Stochastic Gradient Riemann Hamiltonian Monte Carlo)>

- Intuition : Let's take into account underlying target distribution geometry on SGHMC
  - How to? : (SGHMC → SGRHMC)

$$z = (\theta, r), \qquad H(\theta, r) = U(\theta) + \frac{1}{2} r^T M^{-1} r, \qquad Q(\theta, r) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}, \qquad D(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$$

$$\Downarrow$$

$$z = (\theta, r), \qquad H(\theta, r) = U(\theta) + \frac{1}{2} r^T r, \qquad Q(\theta, r) = \begin{pmatrix} 0 & -G(\theta)^{-1/2} \\ G(\theta)^{-1/2} & 0 \end{pmatrix}, \qquad D(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & G(\theta)^{-1} \end{pmatrix}$$

where $G(\theta)_{ij} = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta_i} \log p(x|\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log p(x|\theta) \right) \right]$ is the fisher information matrix.

- When $G(\theta)$ is any positive definite matrix, then it is called gSGRMHC (generalized SGRHMC).

Devising new samplers

<(g)SGRHMC (Stochastic Gradient Riemann Hamiltonian Monte Carlo)>

- Then, we have following discrete dynamics:

$$\begin{cases} \theta_{t+1} \leftarrow \theta_t + \epsilon_t G(\theta_t)^{-1/2} r_t \\ r_{t+1} \leftarrow r_t - \epsilon_t \left[G(\theta_t)^{-\frac{1}{2}} \nabla \widetilde{U}(\theta_t) + \nabla \left(G(\theta_t)^{-\frac{1}{2}} \right) - G(\theta_t)^{-1} r_t \right] + N\left(0, \epsilon_t \left(2G(\theta_t)^{-1} - \epsilon_t \hat{B}_t\right)\right) \end{cases}$$

Note: $\nabla \left(G(\theta)^{-\frac{1}{2}} \right)_i = \sum_{j=1}^{d} \frac{\partial}{\partial \theta_j} \left(G(\theta)^{-\frac{1}{2}} \right)_{ij}$

Algorithm 1: Generalized Stochastic Gradient Riemann Hamiltonian Monte Carlo

initialize (θ_0, r_0)
for $t = 0, 1, 2 \cdots$ **do**
  optionally, periodically resample momentum r as $r^{(t)} \sim \mathcal{N}(0, \mathbf{I})$
  $\theta_{t+1} \leftarrow \theta_t + \epsilon_t \mathbf{G}(\theta_t)^{-1/2} r_t, \quad \Sigma_t \leftarrow \epsilon_t (2\mathbf{G}(\theta_t)^{-1} - \epsilon_t \hat{\mathbf{B}}_t)$
  $r_{t+1} \leftarrow r_t - \epsilon_t \mathbf{G}(\theta_t)^{-1/2} \nabla_\theta \widetilde{U}(\theta_t) - \epsilon_t \nabla_\theta (\mathbf{G}(\theta_t)^{-1/2}) + \epsilon_t \mathbf{G}(\theta_t)^{-1} r_t + \mathcal{N}\Big(0, \Sigma_t\Big)$
end

Applying the theory to construct samplers

[Double check]:

Theory : $z_{t+1} \leftarrow z_t - \epsilon_t\left[\left(D(z_t) + Q(z_t)\right)\nabla\widetilde{H}(z_t) + \Gamma(z_t)\right] + N\left(0, \epsilon_t\left(2D(z_t) - \epsilon_t\hat{B}_{ext,t}\right)\right)$

Note: $Q(\theta, r) = \begin{pmatrix} 0 & -G(\theta)^{-1/2} \\ G(\theta)^{-1/2} & 0 \end{pmatrix}$ and $D(\theta, r) = \begin{pmatrix} 0 & 0 \\ 0 & G(\theta)^{-1} \end{pmatrix}$.

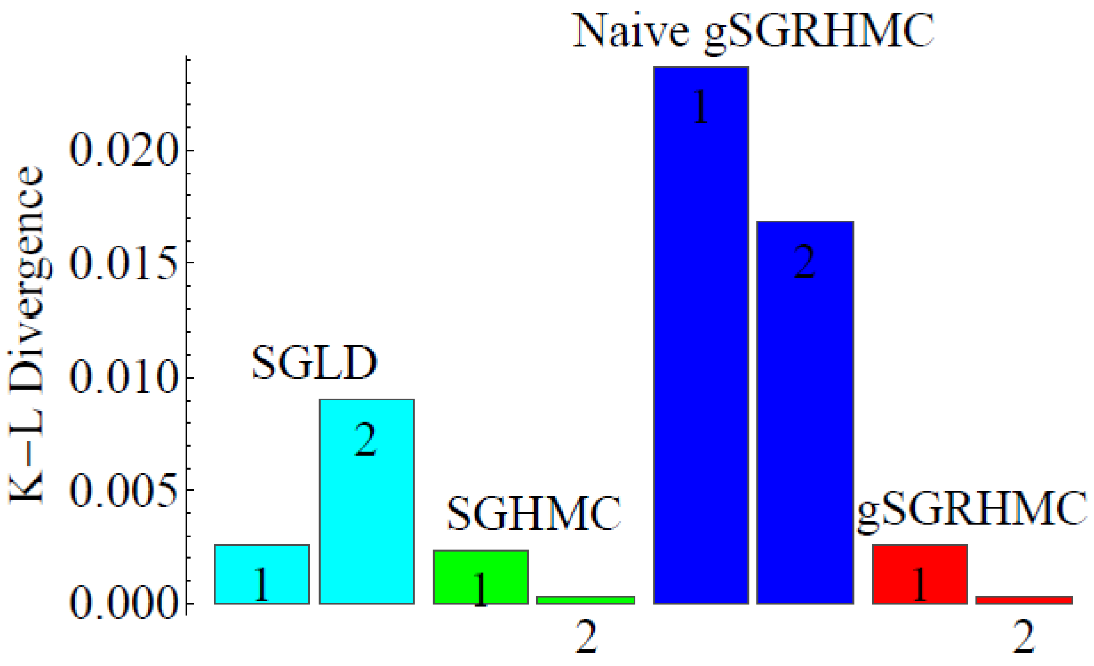
1. $\nabla\widetilde{H}(z) = \left[\nabla\widetilde{U}(\theta)^T,\ r^T\right]^T$ and $\Gamma(z) = \left[0, \nabla\left(G(\theta)^{-\frac{1}{2}}\right)^T\right]^T$

2. $\left(D(z) + Q(z)\right)\nabla\widetilde{H}(z) + \Gamma(z) = \begin{pmatrix} -G(\theta)^{-1/2}r \\ G(\theta)^{-\frac{1}{2}}\nabla\widetilde{U}(\theta) + G(\theta)^{-1}r \end{pmatrix} + \begin{pmatrix} 0 \\ \nabla\left(G(\theta)^{-\frac{1}{2}}\right) \end{pmatrix}$

3. $N\left(0, \epsilon\left(2D(z) - \epsilon\hat{B}_{ext}\right)\right) = N\left(0, \epsilon\left(2G(\theta)^{-1} - \epsilon\hat{B}\right)\right)$ *[dimension reduction]*

Experiments

- KL divergence of two simulated 1D distribution for several SGMCMC algorithms
  - Two 1D distributions : $U(\theta) = \theta^2/2$ (one peak), $U(\theta) = \theta^4 - 2\theta^2$ (two peaks)

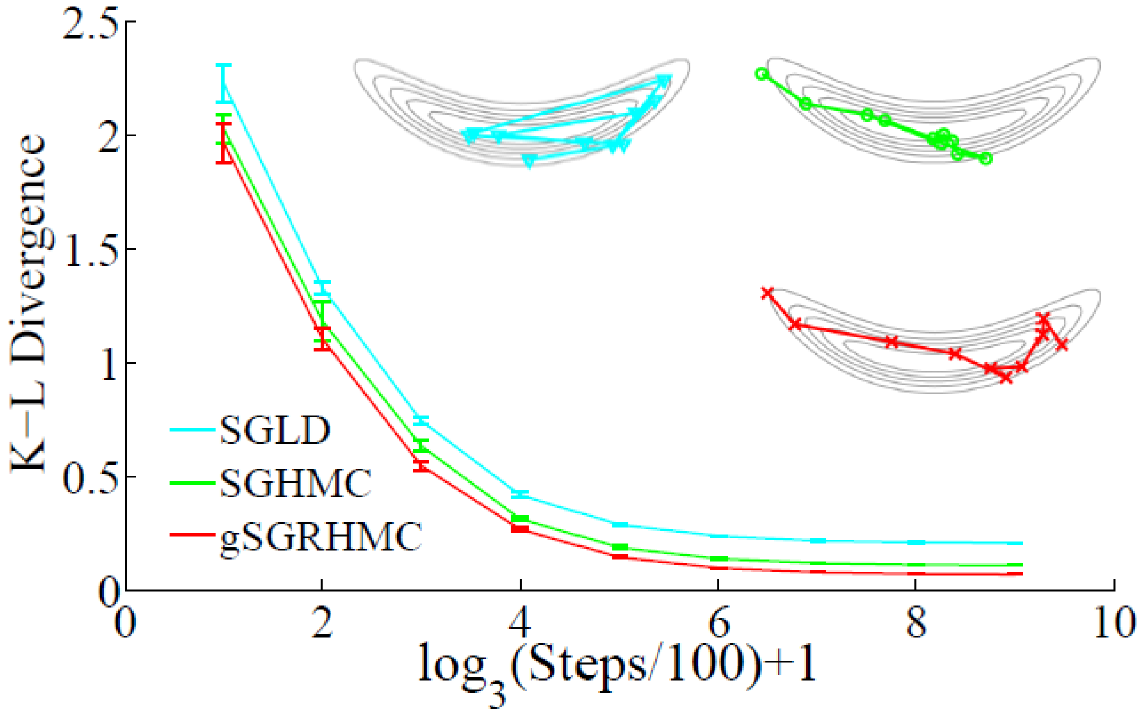


Note:

- **SGHMC is still strong sampler compared to SGRMHC.**
- **SGMHC ≅ gSGRHMC ≥ SGLD on this experiment**
- $\boldsymbol{G(\theta)^{-1} = 1.5\sqrt{|\tilde{U}(\theta) + 0.5|}}$ **on this experiment.**

Experiments

- KL divergence of a simulated 2D distribution for several SGMCMC algorithms
  - Target Distribution $U(\theta_1, \theta_2) = \frac{\theta_1^4}{10} + \frac{\left(4\cdot(\theta_2+1.2)-\theta_1^2\right)^2}{2}$ (having strong correlation)



Note:

- **SGHMC and gSGRHMC can efficiently explores the distribution.**
- **gSGRHMC shows better KL divergence compared to SGHMC on this experiment.**