Cyclic SGMCMC and some implementations

-Summary-

Motivations - Reviews

- SGLD: the target is to maximize the posterior distribution by avoiding the local mode by injecting some noise.

$$\theta_{t+1} - \theta_t = -\epsilon_t \cdot \nabla\widetilde{U}(\theta_t) + \eta_t = \epsilon_t \cdot \left(\nabla \log p(\theta_{ti}) + \frac{N}{n} \sum_{i=1}^{n} \log p(x_{ti}|\theta_t) \right) + \eta_t$$

where ϵ_t : learning rate and $\eta_t \sim N(0, 2\epsilon_t)$

- SGHMC : introduce an auxiliary momentum variable r to improve mixing over SGLD.

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

where C is the momentum coefficient, and $\hat{B}$ is estimate of the noise from $\nabla\widetilde{U}$.

Motivations - Reviews

- To guarantee the asymptotic consistency (i.e : $\mathbb{P}(\|\theta_t - \theta\|_2 > \epsilon) \to 0$ as $t \to \infty$),

  we require decreasing step sizes condition (by Robbin-Monro algorithm):

$$\sum_{n=1}^{\infty} \epsilon_n = \infty \text{ and } \sum_{n=1}^{\infty} \epsilon_n^2 < \infty$$

(Typical decaying step size example : : $\epsilon_t = a(b + t)^{-\gamma}$ with $\gamma \in (0.5,1]$)

- Although many empirical successes have been reported by current SG-MCMC, the sampling becomes ineffective when the loss surface becomes highly multimodal (especially when the dimension of θ ($\sim$ # of parameters of NN) becomes larger)

  $\to$ **Require more efficient sampling techniques for highly multimodal parameter space.**

Ideas

- Q: What is the practical role of step size in SG-MCMC?
  - High ϵ : high possibility to escape mode (Exploration)
  - Low ϵ : low possibility to escape mode (Sampling)
- Combine both schemes to efficiently explore and sample the parameter space.

$$\epsilon_k = \frac{\epsilon_0}{2}\left[\cos\left(\frac{\pi \cdot mod\left(k-1, \left\lceil \frac{K}{M} \right\rceil\right)}{\lceil K/M \rceil}\right) + 1\right]$$

where M = the number of cycles and K = the number of total iterations.

Ideas

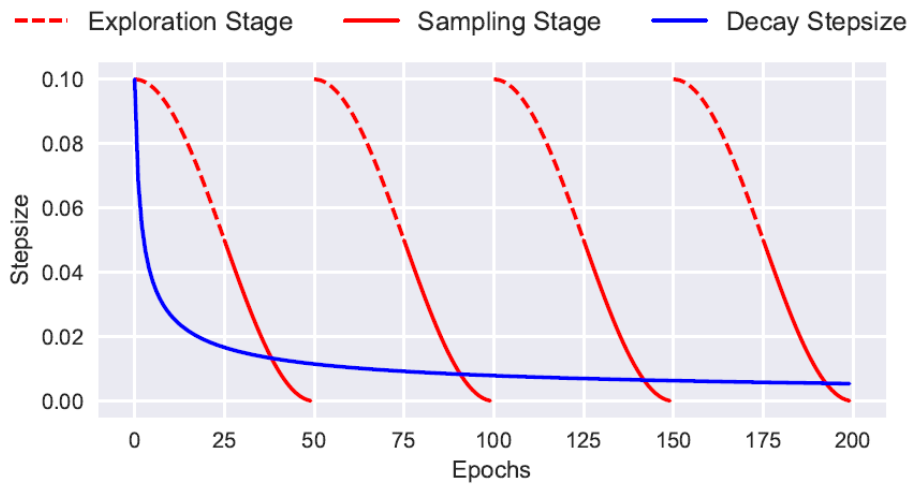


Figure 1: Illustration of the proposed cyclical stepsize schedule (red) and the traditional decreasing stepsize schedule (blue) for SG-MCMC algorithms.

Algorithm 1 Cyclical SG-MCMC.

Input: The initial stepsize α_0, number of cycles M, number of training iterations K and the proportion of exploration stage β.

for k = 1:K **do**

$\alpha \leftarrow \alpha_k$ according to Eq equation 1.

if $\frac{\mathrm{mod}\ (k-1,\lceil K/M \rceil)}{\lceil K/M \rceil} < \beta$ **then**

% Exploration stage

$\theta \leftarrow \theta - \alpha \nabla \tilde{U}_k(\theta)$

else

% Sampling stage

Collect samples using SG-MCMC methods

Output: Samples $\{\theta_k\}$

No collection of samples on exploration stage

- To balance the proportion of the exploration and sampling stages in cSG-MCMC:

  Consider the threshold $\beta \in (0,1)$ and perform exploration for k satisfying

$$r(k) = \frac{mod(k-1, \lceil K/M \rceil)}{\lceil K/M \rceil} < \beta$$

Theoretical analysis

Traditional SG-MCMC can reduce BIAS, MSE → 0 as $k \to \infty$ | **Weak convergence**

- But, Does this strategy can guarantee the Bias and MSE → 0 as $k \to \infty$? : Bias (Δ) / MSE (Δ)

Theorem 1. *Under Assumptions 2 in the appendix, for a smooth test function ϕ, the bias and MSE of cSGLD are bounded as:*

$$\textit{BIAS: } \left|\mathbb{E}\tilde{\phi} - \bar{\phi}\right| = O\left(\frac{1}{\alpha_0 K} + \alpha_0\right), \quad \textit{MSE: } \mathbb{E}\left(\tilde{\phi} - \bar{\phi}\right)^2 = O\left(\frac{1}{\alpha_0 K} + \alpha_0^2\right). \tag{2}$$

- How about the asymptotic distribution difference measured by Wasserstein distance?

Theorem 2. *Under Assumption 3 in the appendix, there exist constants (C_0, C_1, C_2, C_3) independent of the stepsizes such that the convergence rate of our proposed cSGLD with cyclical stepsize sequence equation 1 is bounded for all K satisfying (K mod M =0), as $W_2(\mu_K, \nu_\infty) \leq$*

$$C_3 \exp(-\frac{K\alpha_0}{2C_4}) + \left(6 + \frac{C_2 K\alpha_0}{2}\right)^{\frac{1}{2}} [(C_1 \frac{3\alpha_0^2 K}{8} + \sigma C_0 \frac{K\alpha_0}{2})^{\frac{1}{2}} + (C_1 \frac{3\alpha_0^2 K}{16} + \sigma C_0 \frac{K\alpha_0}{4})^{\frac{1}{4}}] .$$

Particularly, if we further assume $\alpha_0 = O(K^{-\beta})$ for $\forall \beta > 1$, $W_2(\mu_K, \nu_\infty) \leq C_3 +$ $(6 + \frac{C_2}{K^{\beta-1}})^{\frac{1}{2}} [(\frac{2C_1}{K^{2\beta-1}} + \frac{2C_0}{K^{\beta-1}})^{\frac{1}{2}} + (\frac{C_1}{K^{2\beta-1}} + \frac{C_0}{K^{\beta-1}})^{\frac{1}{4}}]$.

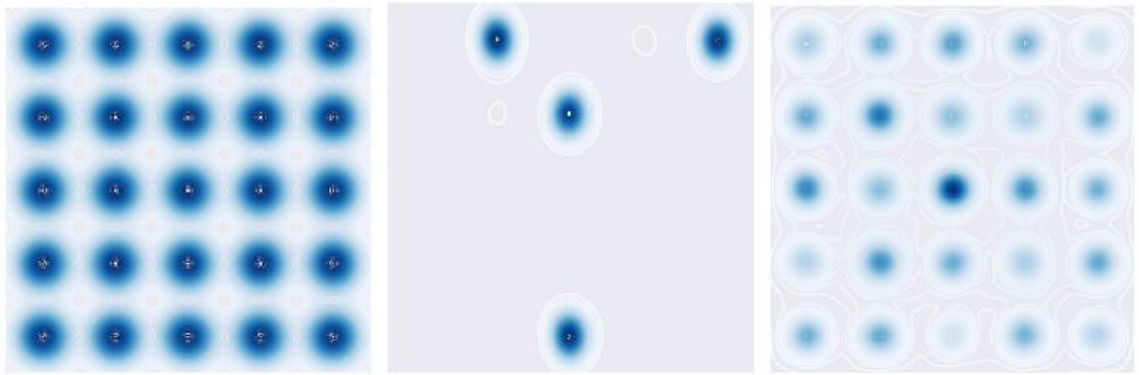
Notations:

μ_k : sample distributions of the sample at step k.

ν_t : distribution from SDE at time t.

If we choose small α_0, then, $W_2(\mu_k, \nu_\infty) \to 0$ as $K \to \infty$.

Experiment on synthetic multimodal data



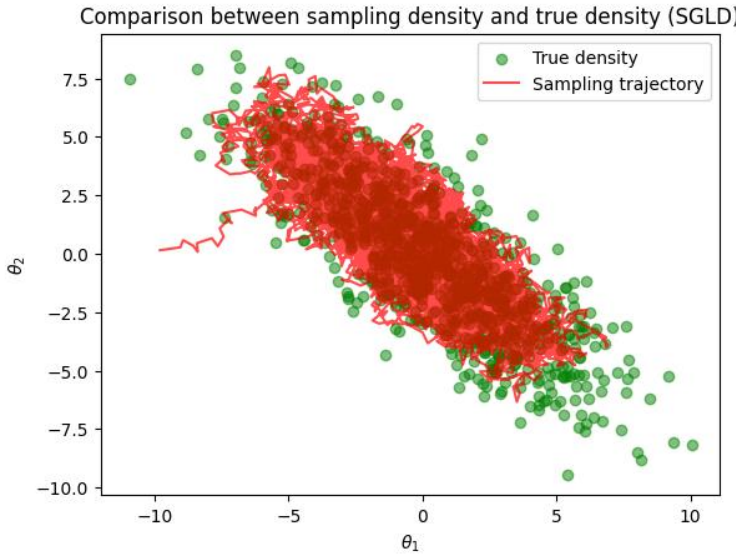
(a) Target (b) SGLD (c) cSGLD

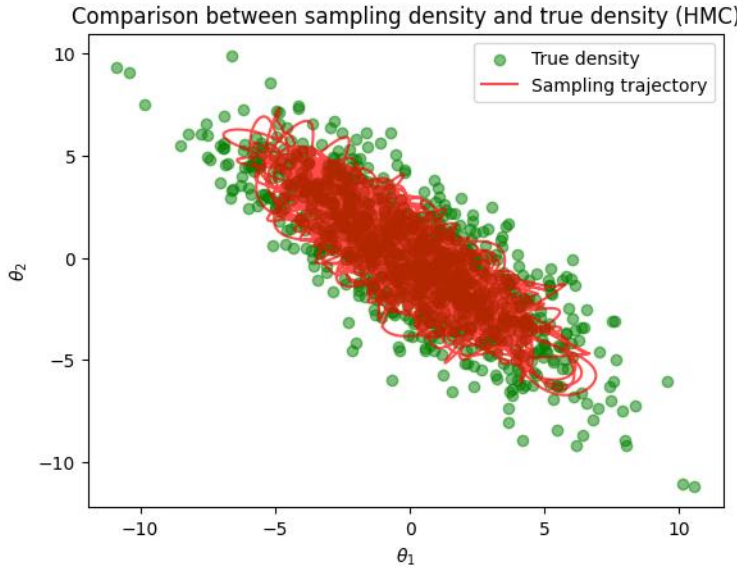
- To demonstrate the multi-modal sampling ability of cSG-MCMC, they checked the sample distributions from SGLD and cSGLD on mixture of 25 Gaussians.

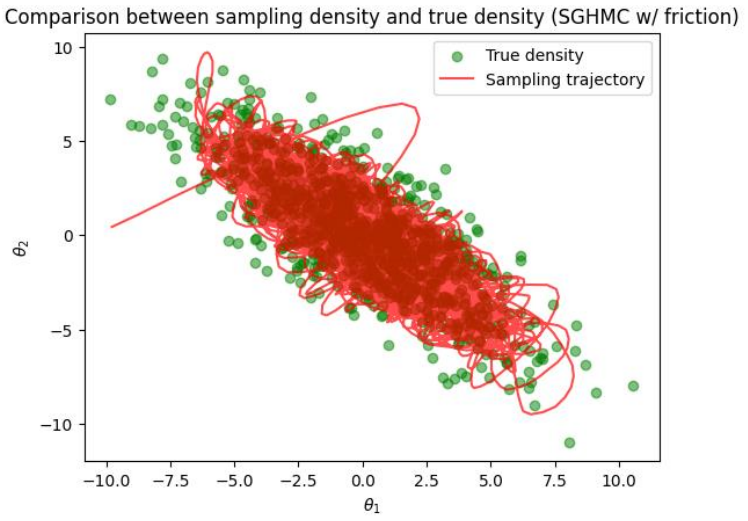
- It turns out that cSGLD leverages the large step sizes to discover a new mode, and small step sizes to explore local modes → significantly favorable strategy in non-asymptotic setting.

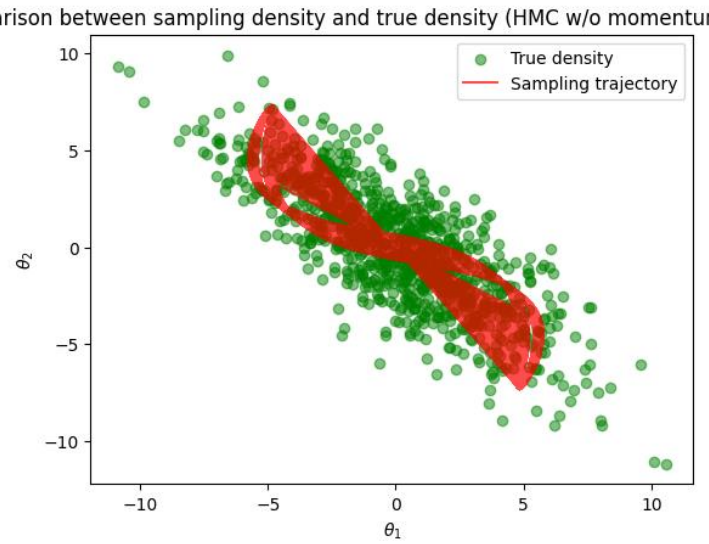
Implementations of SGLD / SGHMC

- Two-dimensional correlated gaussian sampling (iteration : 15000):







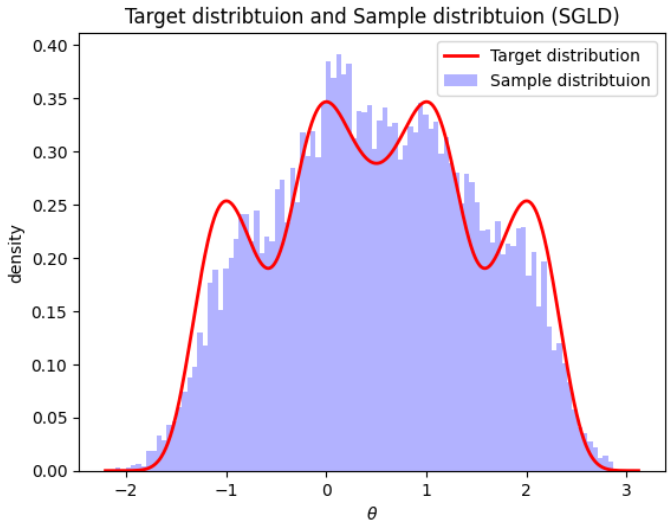


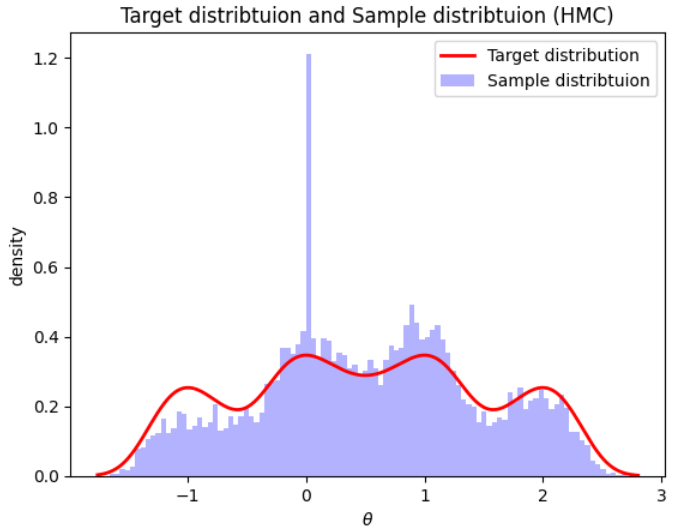
Note :

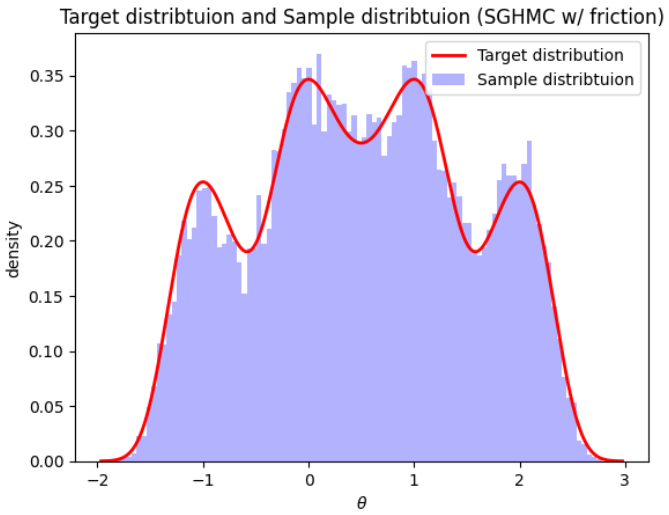
1. **For single-modal density, both SGLD, SGHMC works well.**
2. **Momentum resampling for naïve HMC is crucial for better exploration.**

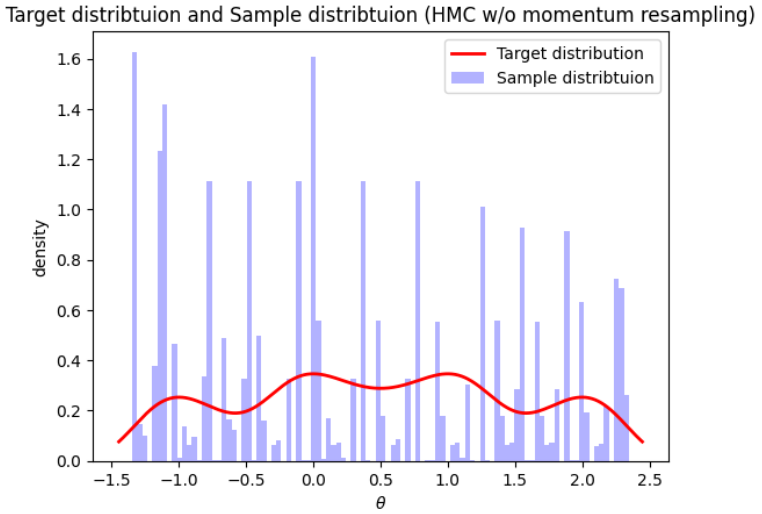
Implementations of SGLD / SGHMC

- One-dimensional multi-modal distribution (iteration : 15000):

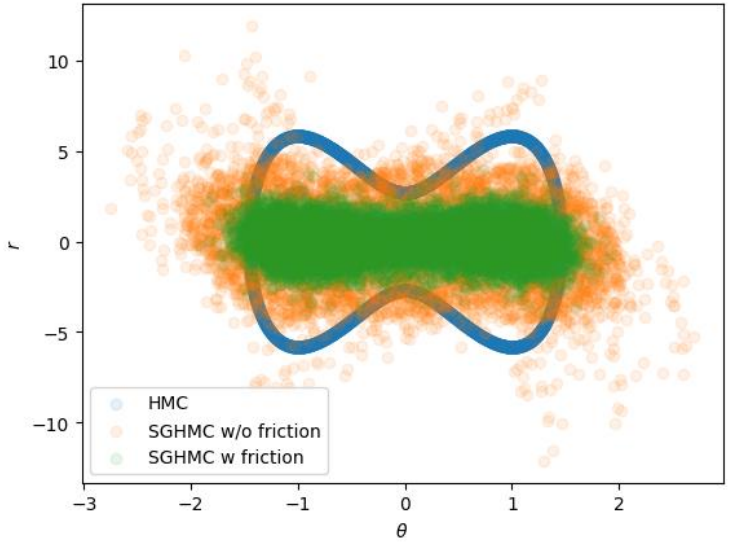








(θ, r) trajectories for HMCs for two-modes distribution



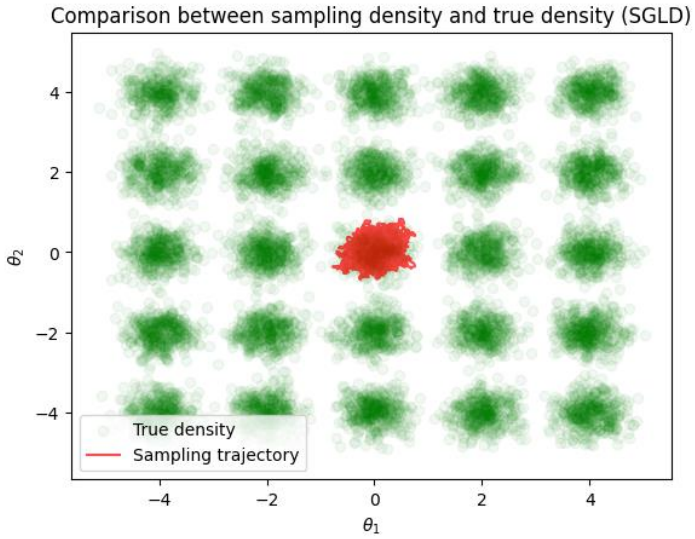
Note :

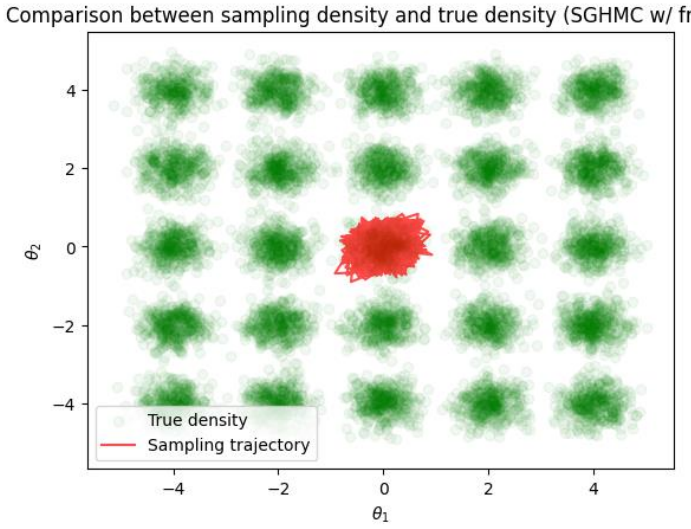
1. **For multi-modal case, the benefits of HMC appears.**

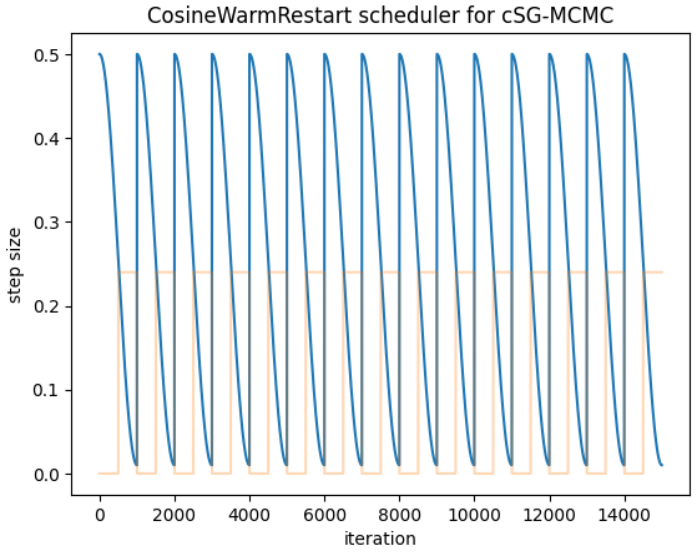
2. **As we studied, SGHMC w/ friction does not require momentum resampling, while it is crucial for naïve HMC .**

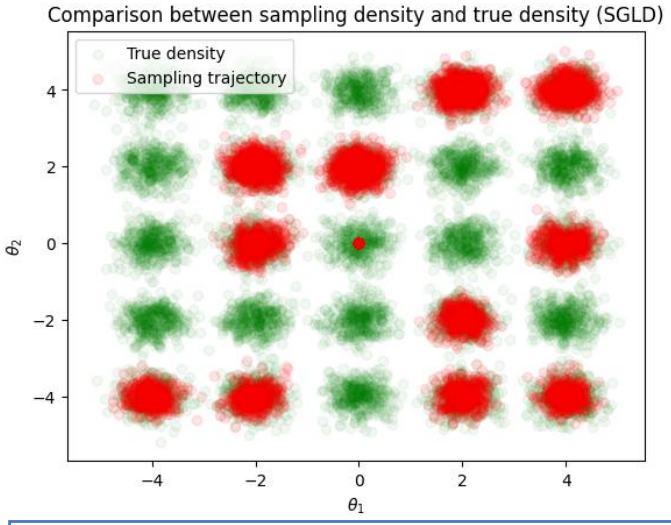
Implementations of cyclic SG-MCMC

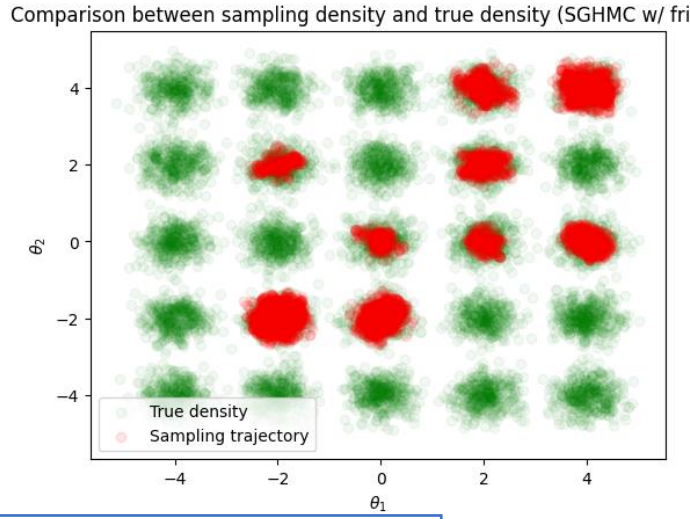
- Multi-modal two dimensional mixtures of gaussians (iteration : 15000):











w/o cyclic scheduler (above), w/ cyclic scheduler (below)

Note :

1. **The tuning for cyclic scheduler was easier for SGLD.**
2. **The maximum step size for scheduler must be carefully chosen to prevent wrong sampling (ex : NaN values)**