1. Big Data MCMC & SGLD

**Problem:** MCMC requires calculations over full dataset each iteration. High computational cost at large datasets.

**Popular Solution:** SGLD [3], which uses a subset of data at each iteration. First, construct unbiased estimate of log posterior gradient using subsample of the data:

\[ \nabla \log \pi(\theta) = \nabla \log \pi_0(\theta) + \frac{N}{|S|} \sum_{x \in S} \nabla \log f(x_i | \theta), \quad S \subset \{1, \ldots, N\} \]

**SGLD Algorithm**
- Set starting value \( \theta_0 \) and stepsize \( h \).
- Iterate and store the following:
  \[ \theta_{m+1} = \theta_m + h \nabla \log \pi(\theta) + \sqrt{2h} \zeta_m, \]
  where \( \zeta_m \sim N(0,1) \).

When \( h \to 0 \) & \( m \to \infty \), SGLD samples from the true posterior. In practice use small \( h \); approximation quality depends on dataset size \( N \). As \( N \) increases, approximation of SGLD gets better.

![Graph showing number of observations](image1)

2. Problem

**Problem:** SGLD designed for unbounded spaces \( \mathbb{R}^d \). Based on constrained spaces like \( [0, \infty) \); especially at boundary. SGLRD [2] aimed to fix problem for simplex space but still biased due to discretisation by \( h \).

![Graph showing number of observations](image2)

3. CIR Process

**Aim:** Construct a scalable, approximate MCMC algorithm for the simplex (i.e., Dirichlet distribution) with no discretisation error. This removes biases at the boundary.

**Useful Transformation:** If \( \theta \sim \text{Gamma}(a_j, 1) \), for \( j = 1, \ldots, d \); then \( \omega = (\theta_1, \ldots, \theta_d) / \sum \theta_j \sim \text{Dirichlet}(a) \).

**Useful Process:** Cox-Ingersoll-Ross (CIR) process [1]. Fix \( h > 0 \). Aim to target \( \text{Gamma}(a, 1) \) distribution. Update is:

\[ \theta_{m+1} = \frac{1 - e^{-h}}{2} W, \quad W \theta_m \sim \chi^2 \left( 2a, 2b + \frac{h}{1 - e^{-h}} \right), \]

\( \chi^2(\nu, \mu) \) is a noncentral Chi sq distribution. As \( m \to \infty \) CIR samples Gamma(\(a, 1)\) for any \( h > 0 \). So \( h \) no longer discretising.

![Graph showing number of observations](image3)

4. Making Scalable: SCIR Process

**Idea:** Use CIR process to target \( \text{Gamma}(a_j, 1), \quad j = 1, \ldots, d \); then use transformation to get simplex samples \( \omega \sim \text{Dirichlet}(a) \).

**Problem:** Typically Dirichlet posterior of form

\[ \text{Dirichlet} \left( \sum_{i=1}^d \alpha_i, \ldots, \sum_{i=1}^d \alpha_i \right) \]

i.e., CIR process calculates \( O(N) \) sum each iteration. Expensive!

**Solution – SCIR Process:** Similar to SGLD, run CIR process with unbiased estimate of \( \alpha \) using subsample of data:

\[ \hat{\alpha} = \sum_{i=1}^d \hat{z}_i \]

SCIR process still discretisation free! \( h \) only determines how often \( \hat{\alpha} \) resampled. Similar to SGLD: as \( m \to \infty \), \( h \to 0 \) SCIR samples from target \( \text{Gamma}(a, 1) \). As discretisation free, provably unbiased.

![Graph showing number of observations](image4)

5. Extensions and Theory

**Extensions:** Method not limited to simplex. SCIR process can sample from Gamma distributed \([0, \infty)\) spaces. Common for inference of variances. Also, processes similar to CIR can be exploited; e.g., geometric Brownian process can be used for scalable inference on lognormal distributed \([0, \infty)\) spaces.

**Theory:** The CIR process is tractable. Allows us to derive theoretical results for SCIR. Can prove asymptotic unbiasedness and derive the non-asymptotic moment generating function.

![Graph showing number of observations](image5)

6. Experiments

**LDA:** Comparison between SCIR, SGLRD and (non-scalable) Gibbs on a Latent Dirichlet Allocation model on Wikipedia corpus.

**Dirichlet Process Mixture:** Comparison between SCIR, SGLRD on a Dirichlet Process Mixture model on the Microsoft user dataset. Stochastic DP sampler not been considered before.

![Graph showing number of observations](image6)