Constrained Exploration via Reflected Replica Exchange Stochastic Gradient Langevin Dynamics

Haoyang Zheng[1], Hengrong Du[2], Qi Feng[3], Wei Deng[4], Guang Lin[1]

[1] Purdue University [2] Vanderbilt University [3] Florida State University [4] Morgan Stanley

Problem Formulation

How **expensive** is it to generate samples which will converge to a target probability density π:

$$d\pi(x_1,x_2)=\frac{1}{Z}\underbrace{e^{-\frac{U(x_1)}{\tau_1}-\frac{U(x_2)}{\tau_2}}}_{p(x_1,x_2)}dx_1dx_2,$$

where Z is a normalizing constant:

$$Z=\int_{\Omega\times\Omega}p(x_1,x_2)dx_1dx_2.$$

Motivations

- reSGLD may **over-explore** if high-temp chain delves too deeply into the distribution tails.
- It deteriorates the model's stability and lead to **poor predictions**.
- We proposed reflected reSGLD, which utilizes **reflection** steps within a bounded domain for **constrained non-convex** exploration.

Contributions

Theory:

- We prove the proposed method outperforms the naïve reSGLD.
- Reducing the domain diameter enhances mixing rates with a **quadratic** behavior.

Experiments:

- We introduces the novel use of the method in dynamic system identification.
- Extensive testing of r2SGLD against multi-modal distribution simulation and large-scale deep learning tasks.

Methodology

Reflected Replica Exchange Langevin Diffusion:

The system dynamics are described by the following SEDs:

$$d\boldsymbol{\beta}_t^{(1)}=-\nabla U(\boldsymbol{\beta}_t^{(1)})dt+\sqrt{2\tau_1}\,dW_t^{(1)}+\nu(\boldsymbol{\beta}_t^{(1)})L^{(1)}(dt),$$
$$d\boldsymbol{\beta}_t^{(2)}=-\nabla U(\boldsymbol{\beta}_t^{(2)})dt+\sqrt{2\tau_2}\,dW_t^{(2)}+\nu(\boldsymbol{\beta}_t^{(2)})L^{(2)}(dt),$$

Notes: W_t is Wiener process; $\nu(\boldsymbol{\beta}_t)$ is inner unit vector; L is independent local time.

The Swap Function:

$$S(\boldsymbol{\beta}_t^{(1)},\boldsymbol{\beta}_t^{(2)}):=e^{\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(U(\boldsymbol{\beta}_t^{(1)})-U(\boldsymbol{\beta}_t^{(2)})\right)}$$

The r2SGLD Algorithm:

Input Initial parameters $\widetilde{\boldsymbol{\beta}}_1^{(1)},\widetilde{\boldsymbol{\beta}}_1^{(2)}$; temperatures τ_1,τ_2; learn rate η.

for $k=1,2,\cdots,K$ **do**

Sampling Step

$$\widetilde{\boldsymbol{\beta}}_{k+1}^{(i)}=\mathcal{R}\left(\widetilde{\boldsymbol{\beta}}_k^{(i)}-\eta\nabla\widetilde{U}(\widetilde{\boldsymbol{\beta}}_k^{(i)})+\sqrt{2\eta\tau_1}\boldsymbol{\xi}_k^{(i)}\right),i=1,2.$$

Swapping Step

Generate a uniform random number $u\in[0,1]$.

Compute $\widetilde{S}(\widetilde{\boldsymbol{\beta}}_k^{(1)},\widetilde{\boldsymbol{\beta}}_k^{(2)})=e^{\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\left(\widetilde{U}(\widetilde{\boldsymbol{\beta}}_k^{(1)})-\widetilde{U}(\widetilde{\boldsymbol{\beta}}_k^{(2)})-\left(\frac{1}{\tau_1}-\frac{1}{\tau_2}\right)\frac{\widetilde{\sigma}^2}{C}\right)}$.

if $u<\widetilde{S}$ **then**

Swap $\widetilde{\boldsymbol{\beta}}_{k+1}^{(1)}$ and $\widetilde{\boldsymbol{\beta}}_{k+1}^{(2)}$.

Output Parameters $\{\widetilde{\boldsymbol{\beta}}_k^{(1)}\}_{k=1}^{K+1}$.

Theoretical Analysis

(a) Assumptions

Assumption A1. Ω is a compact domain with a boundary $\partial\Omega$ whose *second fundamental form* is bounded below by some constant $\kappa\leq 0$.

Assumption A2. The function $U\in C^2(\Omega)$. Since Ω is compact, there exists an $L>0$ such that for all $x,y\in\Omega$,

$$\|\nabla U(x)-\nabla U(y)\|\leq L\|x-y\|.$$

(b) Continuous-time Analysis

Theorem 3.4. *Given any initial measure μ_0 for which $d\mu_0/d\pi$ satisfies (5), the χ^2-divergence to the invariant distribution π decays exponentially according to:*

$$\chi^2(\mu_t\|\pi)\leq\chi^2(\mu_0\|\pi)\exp\left(-2t(1+\eta_S)C_P^{-1}\right),\quad(11)$$

where $\eta_S:=\inf_{t>0}\frac{\mathcal{E}_S\left(\frac{d\mu_t}{d\pi}\right)}{\mathcal{E}\left(\frac{d\mu_t}{d\pi}\right)}-1$ *is the* acceleration effect *in χ^2-divergence.*

Theorem 3.8. *Given any initial measure μ_0 for which $d\mu_0/d\pi\geq 0$ and satisfying (5), the 2-Wasserstein distance between μ_t and π satisfies the following accelerated exponential decay estimate:*

$$\mathcal{W}_2(\mu_t,\pi)\leq\sqrt{2C_{LS}D(\mu_0\|\pi)}\exp\left(-t(1+\delta_S)C_{LS}^{-1}\right),\quad(15)$$

where $\delta_S:=\inf_{t>0}\frac{\mathcal{E}_S\left(\sqrt{\frac{d\mu_t}{d\pi}}\right)}{\mathcal{E}\left(\sqrt{\frac{d\mu_t}{d\pi}}\right)}-1$ *is the* acceleration effect *in $\mathcal{W}_2$-distance.*

(c) Discretization Analysis

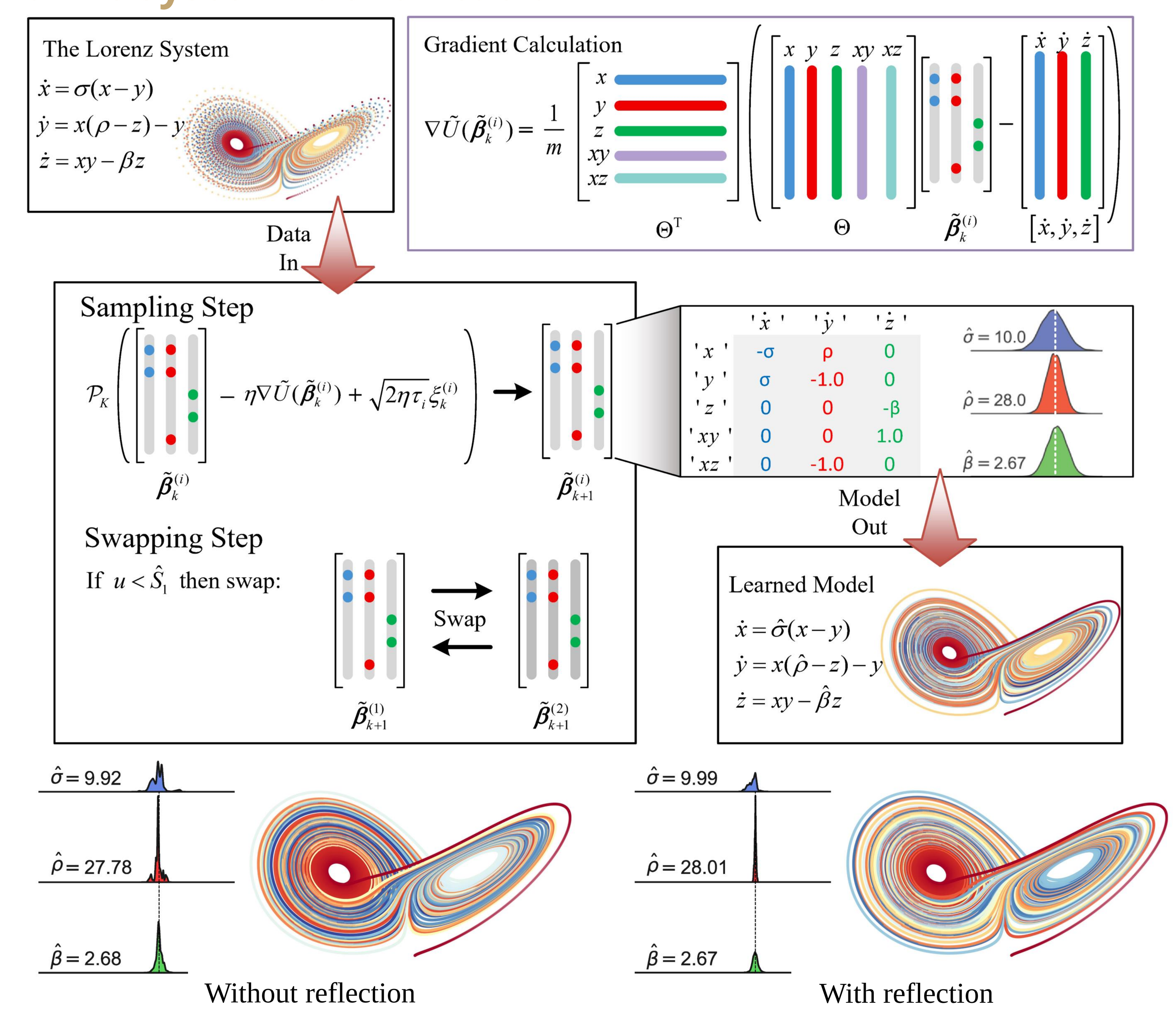
Theorem 3.11 (Discretization error). *Assume that the domain Ω is convex, and Assumptions A1, A2 hold true, then*

$$\mathcal{W}_1(\mu_T,\widetilde{\mu}_T)\leq\tilde{\mathcal{O}}\Bigg(\eta^{1/4}+\sqrt{\max_k\mathbb{E}[\|\phi_k\|^2]}+\sqrt{\eta^{-1/2}\max_k\sqrt{\mathbb{E}\left[|\psi_k|^2\right]}}\Bigg).$$

where $\widetilde{\mu}_T$ denotes the distribution of $\widetilde{\boldsymbol{\beta}}_T^{\eta}$, which is the continuous-time interpolation for r2SGLD, $\phi_k:=\nabla\widetilde{U}-\nabla U$ is the noise in the stochastic gradient, and $\psi_k:=\widetilde{S}-S$ is the noise in the stochastic swapping rate.

Experiments

Dynamic System Identification



Constrained Multi-modal Simulations

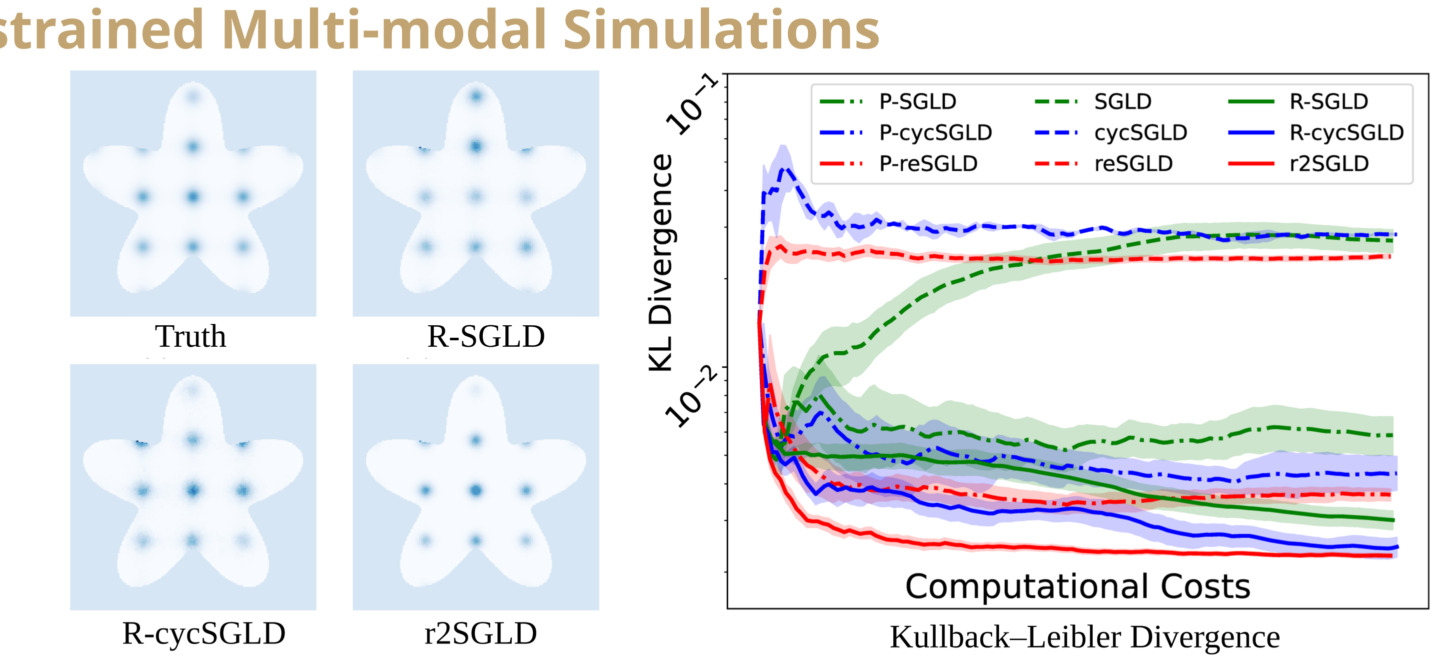
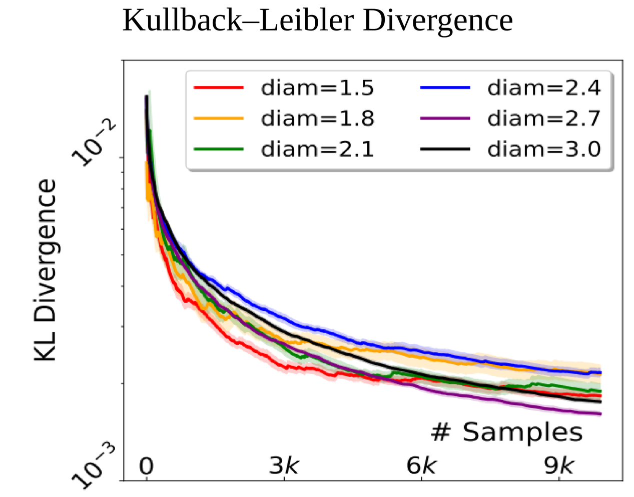


Image Classification

Methods	Metrics (ResNet20)		
	ACC (%) ↑	NLL ↓	Brier (‰) ↓
SGDM	72.13 ± 0.60	9667 ± 108	2.78 ± 0.05
SGHMC	72.47 ± 0.45	9543 ± 157	2.75 ± 0.05
cycSGHMC	73.49 ± 0.17	8913 ± 76	2.65 ± 0.02
reSGHMC	75.01 ± 0.14	8552 ± 69	2.50 ± 0.01
R-SGDM	72.43 ± 0.35	9626 ± 94	2.75 ± 0.03
R-SGHMC	72.85 ± 0.51	9501 ± 167	2.73 ± 0.05
R-cycSGHMC	73.77 ± 0.22	8953 ± 52	2.62 ± 0.02
r2SGHMC	$\mathbf{75.38\pm0.17}$	$\mathbf{8489\pm66}$	$\mathbf{2.46\pm0.02}$

Methods	Metrics (ResNet56)		
	ACC (%) ↑	NLL ↓	Brier (‰) ↓
SGDM	74.40 ± 0.71	9724 ± 169	3.59 ± 0.23
SGHMC	74.22 ± 0.66	9723 ± 214	3.23 ± 0.21
cycSGHMC	77.98 ± 0.61	8303 ± 161	3.19 ± 0.20
reSGHMC	78.87 ± 0.44	7406 ± 130	2.94 ± 0.06
R-SGDM	74.70 ± 0.68	9507 ± 106	3.53 ± 0.18
R-SGHMC	75.10 ± 0.55	9232 ± 158	3.36 ± 0.23
R-cycSGHMC	78.41 ± 0.67	7711 ± 144	3.12 ± 0.11
r2SGHMC	$\mathbf{79.39\pm0.30}$	$\mathbf{7155\pm91}$	$\mathbf{2.89\pm0.02}$



Mixing Rates vs. Domain Diameters

Takeaway

- The proposed r2SGLD algorithm performs the best.
- A smaller domain diameter in multi-modal simulation can improve the mixing rate.
- Large initial learning rate in CIFAR100 facilitates exploration.