A Contour Stochastic Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions

Wei Deng, Guang Lin, Faming Liang
Purdue University, West Lafayette, IN, USA

NEURAL INFORMATION PROCESSING SYSTEMS

Summary

We propose a contour stochastic gradient Langevin dynamics (CSGLD) for scalable Bayesian learning. The algorithm has the following properties:

- Sample from a *flat* density to accelerate the simulations and *adjust* the bias through importance weights.
- Adaptively estimate the latent vector through stochastic approximation and obtain a *sampling-optimization equilibrium* in the long run.
- The mean-field system satisfies a stability condition, which leads to the convergence of the latent vector to a ***unique fixed-point, regardless of the non-convexity*** of the original energy function.
- The convergence of the weighted averaging estimators is guaranteed.

Methodology development

Our interest is to simulate from a *multi-modal* distribution

$$\pi(\boldsymbol{x}) \propto \exp(-U(\boldsymbol{x})/\tau),$$

where $U(\boldsymbol{x})$ is the energy function and τ is the temperature.

To accelerate the simulations, we propose to simulate from a *flattened* density

$$\varpi_{\Psi_{\boldsymbol{\theta}}}(\boldsymbol{x}) \propto \frac{\pi(\boldsymbol{x})}{\Psi^{\zeta}_{\boldsymbol{\theta}}(U(\boldsymbol{x}))},$$

where $\zeta > 0$ is a hyperparameter and $\boldsymbol{\theta} = (\theta(1), \theta(2), \ldots, \theta(m))$ is an unknown latent vector which takes value in the space:

$$\boldsymbol{\Theta} = \left\{ (\theta(1), \theta(2), \cdots, \theta(m)) \,\middle|\, 0 < \theta(1), \theta(2), \cdots, \theta(m) < 1 \text{ and } \sum_{i=1}^{m} \theta(i) = 1 \right\}.$$

Now consider an energy *partition* $\{\mathcal{X}_i\}_{i=1}^m$. If we set ζ and $\Psi_{\boldsymbol{\theta}}$ as follows:

(i) $\zeta = 1$ and $\Psi_{\boldsymbol{\theta}}(U(\boldsymbol{x})) = \sum_{i=1}^{m} \theta(i) 1_{u_{i-1} < U(\boldsymbol{x}) \le u_i}$,

(ii) $\theta(i) = \theta_\star(i)$, where $\theta_\star(i) = \int_{\boldsymbol{\chi}_i} \pi(\boldsymbol{x}) d\boldsymbol{x}$ for $i \in \{1, 2, \cdots, m\}$,

the algorithm leads to a *random walk* in the *space of energy*.

A naïve extension to SGLD Note that this setup only works under the Metropolis setting. A naïve extension of (i) results in $\frac{\partial \log \Psi_\theta(u)}{\partial u} = \frac{1}{\Psi_\theta(u)} \frac{\partial \Psi_\theta(u)}{\partial u} = 0$ a.e..

Thus the algorithm behaves the same as SGLD and fails to simulate from a flat density.

To avoid the vanishing gradient, we set $\Psi_{\boldsymbol{\theta}}(u)$ as a piecewise continuous function:

$$\Psi_{\boldsymbol{\theta}}(u) = \sum_{i=1}^{m} \left(\theta(i-1) e^{(\log\theta(i) - \log\theta(i-1)) \frac{u - u_{i-1}}{\Delta u}} \right) 1_{u_{i-1} < u \le u_i}.$$

A direct calculation shows that

$$\nabla_{\boldsymbol{x}} \log \varpi_{\Psi_{\boldsymbol{\theta}}}(\boldsymbol{x}) = -\left[1 + \zeta\tau \frac{\log\theta(J(\boldsymbol{x})) - \log\theta((J(\boldsymbol{x})-1) \vee 1)}{\Delta u} \right] \frac{\nabla_{\boldsymbol{x}} U(\boldsymbol{x})}{\tau},$$

where $J(\boldsymbol{x}) \in \{1, 2, \cdots, m\}$ is a index that satisfies $u_{J(\boldsymbol{x})-1} < U(\boldsymbol{x}) \le u_{J(\boldsymbol{x})}$.

To obtain the the optimal $\boldsymbol{\theta}_\star$, we propose to estimate it via *stochastic approximation*.

Algorithm 1 Contour SGLD. The original density is recovered via importance weights.

[1.] (Data sampling) Simulate a batch data of size n from the full data of size N.

[2.] (Simulation step) Sample $\boldsymbol{x}_{k+1}$ using the SGLD algorithm based on $\boldsymbol{\theta}_k$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \epsilon_{k+1} \frac{N}{n} \left[1 + \zeta\tau \frac{\log\theta_k(\tilde{J}(\boldsymbol{x}_k)) - \log\theta_k((\tilde{J}(\boldsymbol{x}_k)-1) \vee 1)}{\Delta u} \right] \nabla_{\boldsymbol{x}} \widetilde{U}(\boldsymbol{x}_k) + \sqrt{2\tau\epsilon_{k+1}} \boldsymbol{w}_{k+1},$$

where ϵ is the learning rate, $\boldsymbol{w}$ is a Gaussian vector, $\nabla_{\boldsymbol{x}} \widetilde{U}(\cdot)$ is the stochastic gradient, and $\tilde{J}(\cdot)$ is the index obtained by the stochastic energy $\widetilde{U}(\cdot)$.

[3.] (Stochastic approximation) Update the estimate of $\boldsymbol{\theta}$

$$\theta_{k+1}(i) = \theta_k(i) + \omega_{k+1} \theta_k^{\zeta}(\tilde{J}(\boldsymbol{x}_{k+1})) \left(1_{i = \tilde{J}(\boldsymbol{x}_{k+1})} - \theta_k(i) \right),$$

where $1_{i = \tilde{J}(\boldsymbol{x}_{k+1})}$ is an indicator function which equals 1 if $i = \tilde{J}(\boldsymbol{x}_{k+1})$.

Convergence results

Convergence to a unique fixed point $\boldsymbol{\theta}_\star$ Rewrite the update of the latent vector $\boldsymbol{\theta}_k$ via stochastic approximation

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \omega_{k+1} \widetilde{H}(\boldsymbol{\theta}_k, \boldsymbol{x}_{k+1}),$$

where $\widetilde{H}(\boldsymbol{\theta}, \boldsymbol{x})$ is a random field under $\varpi_{\Psi_{\boldsymbol{\theta}}}(\boldsymbol{x}) \propto \frac{\pi(\boldsymbol{x})}{\Psi^{\zeta}_{\boldsymbol{\theta}}(U(\boldsymbol{x}))}$. Study the mean-field

$$h(\boldsymbol{\theta}) = \int_{\mathcal{X}} \widetilde{H}(\boldsymbol{\theta}, \boldsymbol{x}) \varpi_{\boldsymbol{\theta}}(\boldsymbol{x}) d\boldsymbol{x} \propto (\boldsymbol{\theta}_\star + \varepsilon\beta(\boldsymbol{\theta}) - \boldsymbol{\theta}) = 0.$$

Apply the perturbation theory and a Lyapunov function $\mathbb{V}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\|^2$ leads to:

Lemma 1 (Stability). *Given a small enough ε, there is a constant $\phi > 0$ s.t.*

$$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \langle h(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}_\star \rangle \le -\phi \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2 + \mathcal{O}\Big(\epsilon + \frac{1}{m} + \delta_n(\boldsymbol{\theta}) \Big).$$

Together with the tool of Poisson equation to control the fluctuation, we have

Theorem 1 (L^2 convergence). *Given proper regularity assumptions,* $\boldsymbol{\theta}_k$ converges to a unique $\boldsymbol{\theta}_\star$ even if $U(\cdot)$ is non-convex:

$$\mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_\star\|^2 \right] = \mathcal{O}\Big(\omega_k + \sup_{i \ge k_0} \epsilon_i + \frac{1}{m} + \sup_{i \ge k_0} \delta_n(\boldsymbol{\theta}_i) \Big).$$

Convergence of weighted averaging estimator We first show the convergence of $\frac{1}{k}\sum_{i=1}^{k} f(\boldsymbol{x}_i)$ by treating the adaptive gradient as a *biased* (but *decaying* fast) gradient.

Lemma 2 (Convergence of the Averaging Estimators). *Given proper regularity assumptions, for any bounded function f, we have*

$$\left| \mathbb{E}\left[\frac{\sum_{i=1}^{k} f(\boldsymbol{x}_i)}{k} \right] - \int_{\boldsymbol{\chi}} f(\boldsymbol{x}) \varpi_{\widetilde{\Psi}_{\boldsymbol{\theta}_\star}}(d\boldsymbol{x}) \right| = \mathcal{O}\Big(\frac{1}{k\epsilon} + \sqrt{\epsilon} + \sqrt{\frac{\sum_{i=1}^{k} \omega_k}{k}} + \frac{1}{\sqrt{m}} + \sup_{i \ge k_0} \sqrt{\delta_n(\boldsymbol{\theta}_i)} \Big).$$

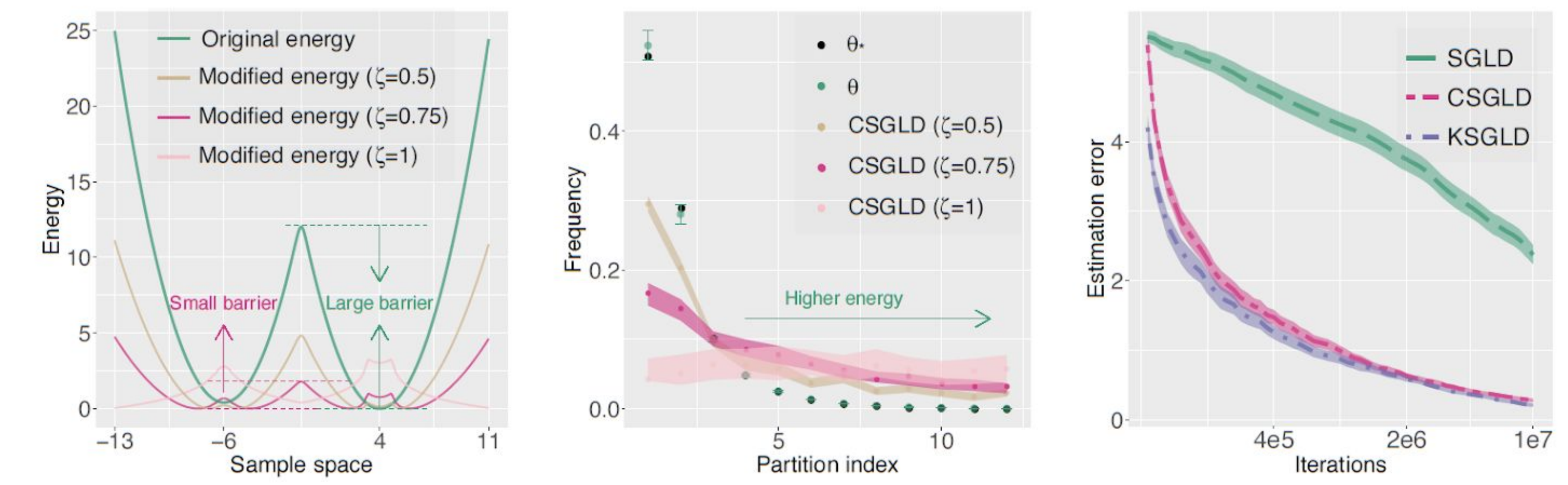
Then we study the convergence of $\frac{\sum_{i=1}^{k} \theta_i^{\zeta}(\tilde{J}(\boldsymbol{x}_i)) f(\boldsymbol{x}_i)}{\sum_{i=1}^{k} \theta_i^{\zeta}(\tilde{J}(\boldsymbol{x}_i))}$ by applying the previous results.

Theorem 2 (Convergence of weighted averaging estimators). *Given a test function f*

$$\left| \mathbb{E}\left[\frac{\sum_{i=1}^{k} \theta_i^{\zeta}(\tilde{J}(\boldsymbol{x}_i)) f(\boldsymbol{x}_i)}{\sum_{i=1}^{k} \theta_i^{\zeta}(\tilde{J}(\boldsymbol{x}_i))} \right] - \int_{\boldsymbol{\chi}} f(\boldsymbol{x}) \pi(d\boldsymbol{x}) \right| = \mathcal{O}\Big(\frac{1}{k\epsilon} + \sqrt{\epsilon} + \sqrt{\frac{\sum_{i=1}^{k} \omega_k}{k}} + \frac{1}{\sqrt{m}} + \sup_{i \ge k_0} \sqrt{\delta_n(\boldsymbol{\theta}_i)} \Big).$$

Experiments

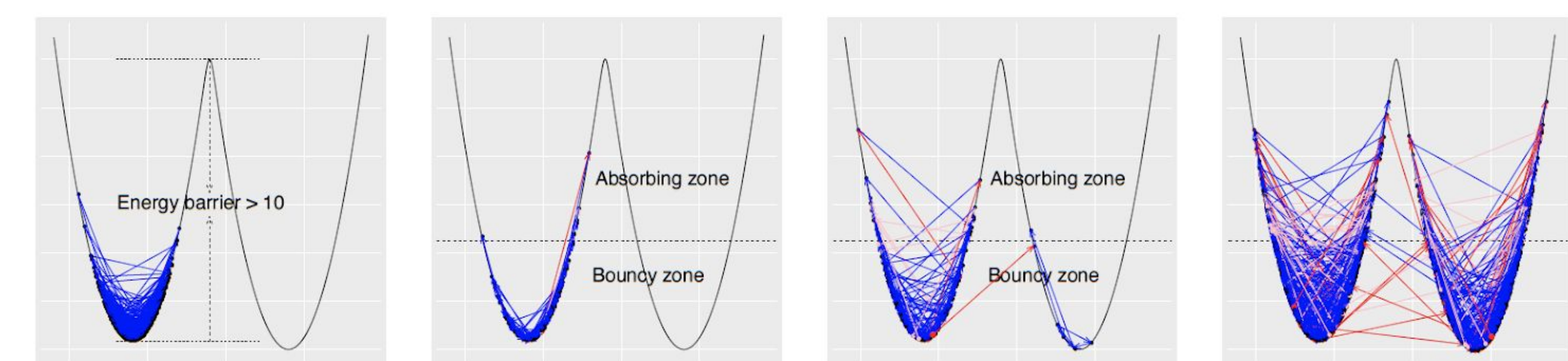
A Gaussian mixture distribution (a) CSGLD samples from a flattened energy with a reduced energy barrier; (b) $\boldsymbol{\theta}$ is well estimated for different ζ's; (c) CSGLD converges much faster than SGLD.



(a) Original v.s. trial (b) $\boldsymbol{\theta}$'s estimates (c) Estimation errors

Figure 1: Landscape and convergence of CSGLD.

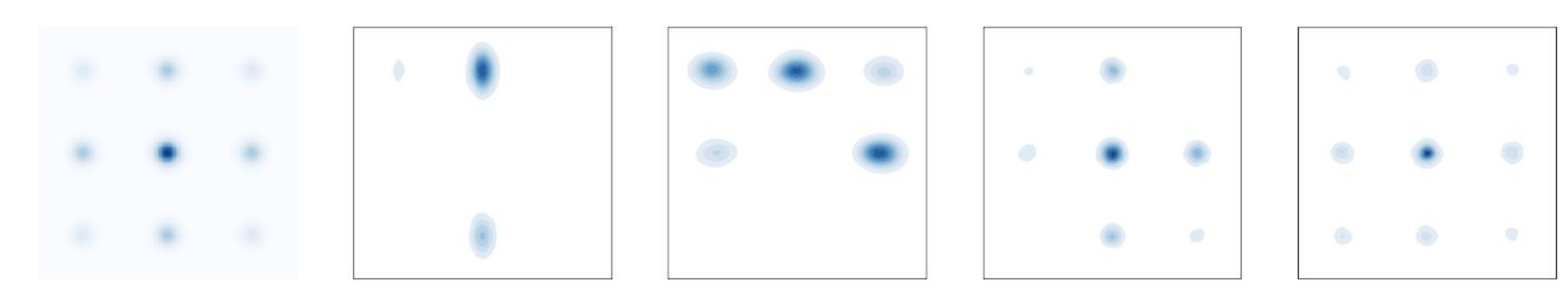
Sample trajectories Given a good estimate of $\boldsymbol{\theta}$, CSGLD yields *a smaller or even negative gradient multiplier* in low energy regions which *bounces the sampler back to high energy.* By contrast, SGLD gets stuck in a local region.



(a) SGLD paths (b) CSGLD (early) (c) CSGLD (mid) (d) CSGLD (late)

Figure 2: Sample trajectories of SGLD and CSGLD

A synthetic multi-modal distribution Compare CSGLD with SGLD, cycic SGLD (cycSGLD, ICLR'20), replica exchange SGLD (reSGLD, ICML'20).



(a) Ground truth (b) SGLD (c) cycSGLD (d) reSGLD (e) CSGLD

Figure 3: Simulations of a distribution. A resampling scheme is used for CSGLD.

Broader impact

CSGLD is a scalable dynamic importance sampler. It is an extension of the flat histogram algorithms from the Metropolis kernel to the Langevin kernel and paves the way for future research in adaptive biasing force techniques for big data problems.

Demo: github.com/WayneDW/Contour-Stochastic-Gradient-Langevin-Dynamics

PURDUE UNIVERSITY