Replica Exchange Stochastic Gradient Langevin Dynamics

Xuda Ye

May 19, 2024

1 Non-convex Learning via Replica Exchange Stochastic Gradient MCMC

The Stochastic Gradient Langevin Dyanmics (SGLD) is a popular technique in large-scale optimization and sampling, while Replica Exchange is a classical melocular dynamics tool to explore the nonconvex and multimodal potential. However, the combination of these two approaches are not trivial, because it is of difficult to keep the reversibility (guaranteed by detailed balance) in the stochastic gradient setting. In the paper [1], the authors propose a viable strategy to blend these two techniques by adaptively estimating the variance of the random potential at each iteration.

Replica exchange dynamics

The Langevin dynamics to sample the Boltzmann distribution is defined by

$$\mathrm{d}\beta_t^{(1)} = -\nabla U(\beta_t^{(1)}) + \sqrt{2\tau_1}\mathrm{d}W_t^{(1)},$$

where $U(\cdot)$ is the potential function, $\tau_1 > 0$ is the temperature, and $W_t^{(1)}$ is the Brownian motion. When the target distribution is multimodal, the Langevin dynamics can trap at the local minima. To resolve this issue, consider a higher temperature and define the coupled Langevin dynamics

$$\begin{aligned}\mathrm{d}\beta_t^{(1)} &= -\nabla U(\beta_t^{(1)}) + \sqrt{2\tau_1}\mathrm{d}W_t^{(1)},\\ \mathrm{d}\beta_t^{(2)} &= -\nabla U(\beta_t^{(2)}) + \sqrt{2\tau_2}\mathrm{d}W_t^{(2)},\end{aligned}$$

where $W_t^{(2)}$ is an independent Brownia motion. The invariant distribution is clearly

$$\pi(\beta^{(1)}, \beta^{(2)}) \propto e^{-\frac{U(\beta^{(1)})}{\tau_1} - \frac{U(\beta^{(2)})}{\tau_2}},$$

whose marginal in $\beta^{(1)}$ is exactly the target distribution.

The principle of Replica Exchange is switching the states $\beta^{(1)}, \beta^{(2)}$ to accelarate the convergence in $\beta^{(1)}$. To maintain the invariant distribution, the swapping rate can be chosen as

$$S_0(\beta^{(1)}, \beta^{(2)}) = e^{(\frac{1}{\tau_1} - \frac{1}{\tau_2})(U(\beta^{(1)}) - U(\beta^{(2)}))}.$$

In this way, Replica Exchange dynamics is a Markov jump process.

Naive stochastic gradient setting

For the potential $U(\beta)$, suppose $\tilde{U}(\beta)$ is the mini-batch approxmation of $U(\beta)$. On the one hand, the time-average effect ensures the Langevin dynamics captures the full gradient $\nabla U(\beta)$. On the other hand, one has to accurately estimate the swapping rate $S_0(\beta^{(1)}, \beta^{(2)})$. Since we only have the mini-batch approximation $\tilde{U}(\beta^{(1)})$ and $\tilde{U}(\beta^{(2)})$, a natural choice of the swapping rate is

$$\mathbb{S}_0(\beta^{(1)}, \beta^{(2)}) = e^{(\frac{1}{\tau_1} - \frac{1}{\tau_2})(\tilde{U}(\beta^{(1)}) - \tilde{U}(\beta^{(2)}))}.$$

However this choice causes large bias since $\tilde{U}(\beta)$ has its own variance.

Variance reduced stochastic gradient setting

We make the following assumption on the mini-batch approximation potential:

$$\tilde{U}(\beta) \sim \mathcal{N}(U(\beta), \sigma^2),$$

where σ^2 is the variance of the approximation. Next, for independent mini-batch samples at the states $\beta^{(1)}$ and $\beta^{(2)}$, we have

$$\tilde{U}(\beta^{(1)}) - \tilde{U}(\beta^{(2)}) = U(\beta^{(1)}) - U(\beta^{(2)}) + \sqrt{2}\sigma\xi,$$

where $\xi \sim \mathcal{N}(0, 1)$ is the standard normal random variable. Then we construct the swapping rate

$$\begin{aligned}\tilde{S}_1(\beta^{(1)}, \beta^{(2)}) &= e^{(\frac{1}{\tau_1} - \frac{1}{\tau_2})\left((\tilde{U}(\beta^{(1)}) - \tilde{U}(\beta^{(2)})) - (\frac{1}{\tau_1} - \frac{1}{\tau_2})\sigma^2\right)} \\ &= e^{(\frac{1}{\tau_1} - \frac{1}{\tau_2})\left((U(\beta^{(1)}) - U(\beta^{(2)})) - (\frac{1}{\tau_1} - \frac{1}{\tau_2})\sigma^2 + \sqrt{2}\sigma\xi\right)}.\end{aligned}$$

By taking the expectation with respect to ξ, we conclude that $\tilde{S}_1(\beta^{(1)}, \beta^{(2)})$ is an unbiased approximation to $S_0(\beta^{(1)}, \beta^{(2)})$.

Algorithm design

In practical computation, one can estimate the variance by the variance of the minibatch approximation, and adaptively update the variance in the simulation, because the variance σ^2 can depends on the state β. Then we can use

$$\hat{S}_1(\beta^{(1)}, \beta^{(2)}) = e^{(\frac{1}{\tau_1} - \frac{1}{\tau_2})\left((U(\beta^{(1)}) - U(\beta^{(2)})) - \frac{(\frac{1}{\tau_1} - \frac{1}{\tau_2})\sigma^2}{F}\right)},$$

where $F > 1$ is a constant to reduce the variance.

2 Constrained exploration via Reflected Replica Exchange Stochastic Gradient Langevin Dynamics

The paper [2] considers the large-scale constrained sampling in a bounded region $\Omega \subset \mathbb{R}^d$. The original Replica Exchange Stochastic Gradient Langevin Dynamics is then added a reflection part to address the sampling problem. The exponential convergence and the error analysis of this approach are obtained.

Motivation

The need for constrained sampling here is not from molecular dynamics, but from the over-exploration effect in machine learning. To force the sampling explores the reasonable region, the constrained sampling can be emlpoyed. The paper says the following:

> Specifically, over-exploration can result in either exploding or oscillating losses in deep learning training. This phenomenon can deteriorate the model's stability and optimization performance, and lead to poor predictions. ... To address this, constrained sampling techniques in MCMC play a pivotal role for different purposes in various forms, such as sampling on explicitly defined manifolds, implicitly defined manifolds, and sampling with moment constraints.

Reflected Replica Exchange Langevin dynamics

In the following, let $\Omega \subset \mathbb{R}^d$ be a compact region and $\partial\Omega$ be its boundary. Let the temperature $\tau_1 < \tau_2$, and the target distribution is

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

where Z is the normalization constant given by

$$Z = \int_{\Omega\times\Omega} e^{-\frac{U(x_1)}{\tau_1} - \frac{U(x_2)}{\tau_2}} \mathrm{d}x_1 \mathrm{d}x_2.$$

To sample $\pi(x_1, x_2)$, the Replica Exchange dynamics is defined as

$$\begin{aligned}
\dot{\beta}_t^{(1)} &= -\nabla U(\beta_t^{(1)})\mathrm{d}t + \sqrt{2\tau_1}\mathrm{d}W_t^{(1)} + \nu(\beta_t^{(1)})L^{(1)}(\mathrm{d}t), \\
\dot{\beta}_t^{(2)} &= -\nabla U(\beta_t^{(2)})\mathrm{d}t + \sqrt{2\tau_2}\mathrm{d}W_t^{(2)} + \nu(\beta_t^{(2)})L^{(2)}(\mathrm{d}t),
\end{aligned}$$

where $\beta^{(1)}, \beta^{(2)}$ represent the states corresponding the temperatures τ_1, τ_2, $W_t^{(1)}, W_t^{(2)}$ are independent Brownian motions. The function $\nu(\beta)$ is defined in the boundary $\partial\Omega$, and

represents the inner unit normal vector on $\partial\Omega$. $L^{(1)}$ and $L^{(2)}$ denote local times with reference to $\partial\Omega$.

The infinitesimal generator of the dynamics is given by

$$\begin{aligned}\mathcal{L}f = &-\langle\nabla_{x_1}f(x_1,x_2),\nabla U(x_1)\rangle + \tau_1\Delta_{x_1}f(x_1,x_2)\\ &-\langle\nabla_{x_2}f(x_1,x_2),\nabla U(x_2)\rangle + \tau_2\Delta_{x_2}f(x_1,x_2) + rS(x_1,x_2)\cdot(f(x_2,x_1)-f(x_1,x_2)),\end{aligned}$$

and the time evolution of $f(x_1,x_2)$ obeys the Neumann boundary condition

$$\nabla_{x_1}f(x_1,x_2)\cdot\nu(x_1)=0,\qquad \nabla_{x_2}f(x_1,x_2)\cdot\nu(x_2)=0.$$

The Dirichlet forms of the dynamics is defined as

$$\begin{aligned}\mathcal{E}(f) &= \int\Big(\tau_1\|\nabla_{x_1}f\|^2+\tau_2\|\nabla_{x_2}f\|^2\Big)\mathrm{d}\pi(x_1,x_2),\\ \mathcal{E}_S(f) &= \mathcal{E}(f)+\frac{r}{2}\int S(x_1,x_2)\cdot\big(f(x_2,x_1)-f(x_1,x_2)\big)^2\mathrm{d}\pi(x_1,x_2),\end{aligned}$$

and it can be seen that

$$-\int_{\Omega\times\Omega}f(x_1,x_2)(\mathcal{L}f)(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2=\mathcal{E}_S(f)\Longrightarrow\mathcal{L}f=\frac{1}{2}\frac{\delta\mathcal{E}_S}{\delta f}(f).$$

Algorithm design in the stochastic gradient setting

As indicated in the previous paper, there is a variance correction term in the swapping rate. Let $\tilde{U}(\beta)$ be the mini-batch approximation to $U(\beta)$, then the swapping rate is computed as

$$\tilde{S}(\beta^{(1)},\beta^{(2)})=e^{(\frac{1}{\tau_1}-\frac{1}{\tau_2})\left(\tilde{U}(\beta^{(1)})-\tilde{U}(\beta^{(2)})-(\frac{1}{\tau_1}-\frac{1}{\tau_2})\frac{\sigma^2}{C}\right)},$$

where σ^2 is the estimated variance of the approximation $\tilde{U}(\beta)$. Finally, r2SGLD is defined as in Algorithm 1.

Convergence analysis

We focus on the exponential convergence of the Reflected Replica Exchange Langevin Dynamics without stochastic gradient. Let μ_t be distribution law of the dynamics, then

$$\mathcal{W}_2(\mu_t,\pi)\leqslant\sqrt{2C_{\mathrm{LS}}D(\mu_0\|\pi)}\exp\big(-t(1+\delta_S)C_{\mathrm{LS}}^{-1}\big),$$

where $D(\cdot\|\cdot)$ is the KL divergence, C_{LS} is the log-Sobolev constant corresponding to the Dirichlet form $\mathcal{E}(f)$, and

$$\delta_S:=\inf_{t>0}\frac{\mathcal{E}_S(\sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}})}{\mathcal{E}(\sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}})}-1$$

is the accelaration effect. The explicit expression of δ_S seems unknown.

Proof. The proof is a direct application of the functional inequalities. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}D(\mu_t\|\pi) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega\times\Omega} e^{t\mathcal{L}}\left(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\right)\ln\left[e^{t\mathcal{L}}\left(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\right)\right]\mathrm{d}\pi \\ &= \int_{\Omega\times\Omega}\ln\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\right)\mathcal{L}\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\right)\mathrm{d}\pi \\ &= \frac{1}{2}\int_{\Omega\times\Omega}\ln\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\right)\frac{\delta\mathcal{E}_S}{\delta f}\left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\right)\mathrm{d}\pi \\ &= -\frac{1}{2}\int_{\Omega\times\Omega}\left(\frac{d\mu_t}{\mathrm{d}f}\right)^{-1}\mathcal{E}_S\left(\frac{d\mu_t}{\mathrm{d}f}\right)\mathrm{d}\pi = -2\mathcal{E}_S\left(\sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}}\right).\end{aligned}$$

On the other hand, the log-Sobolev inequality implies

$$D(\mu_t\|\pi) \leqslant C_{\mathrm{LS}}\mathcal{E}\left(\sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}}\right) \leqslant C_{\mathrm{LS}}(1+\delta_S)^{-1}\mathcal{E}_S\left(\sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}}\right),$$

hence we obtain the exponential convergence decay of the KL divergence:

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\mu_t\|\pi) \leqslant -2C_{\mathrm{LS}}^{-1}(1+\delta_S)D(\mu_t\|\pi),$$

and

$$D(\mu_t\|\pi) \leqslant D(\mu_0\|\pi)\exp\{-2C_{\mathrm{LS}^{-1}}(1+\delta_S)t\}, \quad \forall t \geqslant 0.$$

This error analysis framework is similar to Birth-Death Dynamics [3] by Yulong Lu and Jianfeng Lu. It is interesting that Birth-Death Dynamics also aims to sample to non-convex potentials efficiently.

References

[1] Wei Deng, Qi Feng, Liyao Gao, Faming Liang, and Guang Lin. Non-convex learning via replica exchange stochastic gradient mcmc. In *International Conference on Machine Learning*, pages 2474–2483. PMLR, 2020.

[2] Haoyang Zheng, Hengrong Du, Qi Feng, Wei Deng, and Guang Lin. Constrained exploration via reflected replica exchange stochastic gradient langevin dynamics. *arXiv preprint arXiv:2405.07839*, 2024.

[3] Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.