# Supplementary Material for "*Non-convex Learning via Replica Exchange Stochastic Gradient MCMC*"

**Wei Deng** [1] **Qi Feng** [*2] **Liyao Gao** [*1] **Faming Liang** [1] **Guang Lin** [1]

In this supplementary material, we prove the convergence in §1 and show the experimental setup in §2.

## 1. Convergence Analysis

### 1.1. Background

The continuous-time replica exchange Langevin diffusion (reLD) $\{\boldsymbol{\beta}_t\}_{t\geq 0} := \left\{\begin{pmatrix}\boldsymbol{\beta}_t^{(1)}\\ \boldsymbol{\beta}_t^{(2)}\end{pmatrix}\right\}_{t\geq 0}$ is a Markov process compounded with a Poisson jump process. In particular, the Markov process follows the stochastic differential equations

$$\begin{aligned} d\boldsymbol{\beta}_t^{(1)} &= -\nabla U(\boldsymbol{\beta}_t^{(1)})dt + \sqrt{2\tau_1}d\boldsymbol{W}_t^{(1)} \\ d\boldsymbol{\beta}_t^{(2)} &= -\nabla U(\boldsymbol{\beta}_t^{(2)})dt + \sqrt{2\tau_2}d\boldsymbol{W}_t^{(2)}, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)}$ are the particles (parameters) at time $t$ in $\mathbb{R}^d$, $\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)} \in \mathbb{R}^d$ are two independent Brownian motions, $U : \mathbb{R}^d \to \mathbb{R}$ is the energy function, $\tau_1 < \tau_2$ are the temperatures. The jumps originate from the swaps of particles $\boldsymbol{\beta}_t^{(1)}$ and $\boldsymbol{\beta}_t^{(2)}$ and follow a Poisson process where the jump rate is specified as the Metropolis form $rS(\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)})dt$. Here $r \geq 0$ is a constant, and $S$ follows

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

Under such a swapping rate, the probability $\nu_t$ associated with reLD at time $t$ is known to converge to the invariant measure (Gibbs distribution) with density

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

In practice, obtaining the exact energy and gradient for reLD (1) in a large dataset is quite expensive. We consider the replica exchange stochastic gradient Langevin dynamics (reSGLD), which generates iterates $\{\widetilde{\boldsymbol{\beta}}^{\eta}(k)\}_{k\geq 1}$ as follows

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}^{\eta(1)}(k+1) &= \widetilde{\boldsymbol{\beta}}^{\eta(1)}(k) - \eta\nabla\widetilde{U}(\widetilde{\boldsymbol{\beta}}^{\eta(1)}(k)) + \sqrt{2\eta\tau_1}\boldsymbol{\xi}_k^{(1)} \\ \widetilde{\boldsymbol{\beta}}^{\eta(2)}(k+1) &= \widetilde{\boldsymbol{\beta}}^{\eta(2)}(k) - \eta\nabla\widetilde{U}(\widetilde{\boldsymbol{\beta}}^{\eta(2)}(k)) + \sqrt{2\eta\tau_2}\boldsymbol{\xi}_k^{(2)}, \end{aligned} \tag{2}$$

where $\eta$ is considered to be a fixed learning rate for ease of analysis, and $\boldsymbol{\xi}_k^{(1)}$ and $\boldsymbol{\xi}_k^{(2)}$ are independent Gaussian random vectors in $\mathbb{R}^d$. Moreover, the positions of the particles swap based on the stochastic swapping rate. In particular,

[*] Equal contribution [1] Purdue University, West Lafayette, IN, USA [2] University of Southern California, Los Angeles, CA, USA. Correspondence to: Wei Deng <deng106@purdue.edu>, Guang Lin <guanglin@purdue.edu>, Faming Liang <fmliang@purdue.edu>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

$\widetilde{S}(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}) := S(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}) + \psi$, and the stochastic gradient $\nabla\widetilde{U}(\cdot)$ can be written as $\nabla U(\cdot) + \boldsymbol{\phi}$, where both $\psi \in \mathbb{R}^1$ and $\boldsymbol{\phi} \in \mathbb{R}^d$ are random variables with mean not necessarily zero. We also denote $\mu_k$ as the probability measure associated with $\{\widetilde{\boldsymbol{\beta}}^{\eta}(k)\}_{k\geq 1}$ in reSGLD (2) at step $k$, which is close to $\nu_{k\eta}$ in a suitable sense.

### 1.2. Overview of the analysis

We aim to study the convergence analysis of the probability measure $\mu_k$ to the invariant measure $\pi$ in terms of 2-Wasserstein distance,

$$\mathcal{W}_2(\mu, \nu) := \inf_{\Gamma\in\text{Couplings}(\mu,\nu)} \sqrt{\int \|\boldsymbol{\beta}_\mu - \boldsymbol{\beta}_\nu\|^2 d\Gamma(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\nu)}, \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm, and the infimum is taken over all joint distributions $\Gamma(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\nu)$ with $\mu$ and $\nu$ being the marginals distributions.

By the triangle inequality, we easily obtain that for any $k \in \mathbb{N}$ and $t = k\eta$, we have

$$\mathcal{W}_2(\mu_k, \pi) \leq \underbrace{\mathcal{W}_2(\mu_k, \nu_t)}_{\text{Discretization error}} + \underbrace{\mathcal{W}_2(\nu_t, \pi)}_{\text{Exponential decay}} .$$

We start with the discretization error first by analyzing how reSGLD (2) tracks the reLD (1) in 2-Wasserstein distance. The critical part is to study the discretization of the Poisson jump process in mini-batch settings. To handle this issue, we follow Dupuis et al. (2012) and view the swaps of positions as swaps of temperatures. Then we apply standard techniques in stochastic calculus (Chen et al., 2019; Yin and Zhu, 2010; Sato and Nakagawa, 2014; Raginsky et al., 2017) to discretize the Langevin diffusion and derive the corresponding discretization error.

Next, we quantify the evolution of the 2-Wasserstein distance between $\nu_t$ and $\pi$. The key tool is the exponential decay of entropy (Kullback-Leibler divergence) when $\pi$ satisfies the log-Sobolev inequality (LSI) (Bakry et al., 2014). To justify LSI, we first verify LSI for reSGLD without swaps, which is a direct result given a proper Lyapunov function criterion (Cattiaux et al., 2010) and the Poincaré inequality (Chen et al., 2019). Then we follow Chen et al. (2019) and verify LSI for reLD with swaps by analyzing the Dirichlet form. Finally, the exponential decay of the 2-Wasserstein distance follows from the Otto-Villani theorem by connecting the 2-Wasserstein distance with the entropy (Bakry et al., 2014).

Before we move forward, we first lay out the following assumptions:

**Assumption 1** (Smoothness)**.** *The energy function $U(\cdot)$ is $C$-smoothness, which implies that there exists a Lipschitz constant $C > 0$, such that for every $x, y \in \mathbb{R}^d$, we have $\|\nabla U(x) - \nabla U(y)\| \leq C\|x - y\|$.* [1]

**Assumption 2** (Dissipativity)**.** *The energy function $U(\cdot)$ is $(a, b)$-dissipative, i.e. there exist constants $a > 0$ and $b \geq 0$ such that $\forall x \in \mathbb{R}^d$, $\langle x, \nabla U(x)\rangle \geq a\|x\|^2 - b$.*

Here the smoothness assumption is quite standard in studying the convergence of SGLD, and the dissipativity condition is widely used in proving the geometric ergodicity of dynamic systems (Raginsky et al., 2017; Xu et al., 2018). Moreover, the convexity assumption is not required in our theory.

[1] $\|\cdot\|$ denotes the Euclidean $L^2$ norm.

### 1.3. Analysis of discretization error

The key to deriving the discretization error is to view the swaps of positions as swaps of the temperatures, which has been proven equivalent in distribution (Dupuis et al., 2012). Therefore, we model reLD using the following SDE,

$$d\boldsymbol{\beta}_t = -\nabla G(\boldsymbol{\beta}_t)dt + \Sigma_t d\boldsymbol{W}_t, \tag{4}$$

where $G(\boldsymbol{\beta}_t) = \begin{pmatrix} U(\boldsymbol{\beta}_t^{(1)}) \\ U(\boldsymbol{\beta}_t^{(1)}) \end{pmatrix}$, $\boldsymbol{W} \in \mathbb{R}^{2d}$ is a Brownian motion, $\Sigma_t$ is a random matrix in continuous-time that swaps between the diagonal matrices $\mathbb{M}_1 = \begin{pmatrix} \sqrt{2\tau_1}\mathbf{I}_d & 0 \\ 0 & \sqrt{2\tau_2}\mathbf{I}_d \end{pmatrix}$ and $\mathbb{M}_2 = \begin{pmatrix} \sqrt{2\tau_2}\mathbf{I}_d & 0 \\ 0 & \sqrt{2\tau_1}\mathbf{I}_d \end{pmatrix}$ with probability $rS(\boldsymbol{\beta}_t^{(1)}, \boldsymbol{\beta}_t^{(2)})dt$, and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is denoted as the identity matrix.

Moreover, the corresponding discretization of replica exchange SGLD (reSGLD) follows:

$$\widetilde{\boldsymbol{\beta}}^{\eta}(k+1) = \widetilde{\boldsymbol{\beta}}^{\eta}(k) - \eta\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}(k)) + \sqrt{\eta}\widetilde{\Sigma}^{\eta}(k)\boldsymbol{\xi}_k, \tag{5}$$

where $\boldsymbol{\xi}_k$ is a standard Gaussian distribution in $\mathbb{R}^{2d}$, and $\widetilde{\Sigma}^{\eta}(k)$ is a random matrix in discrete-time that swaps between $\mathbb{M}_1$ and $\mathbb{M}_2$ with probability $r\widetilde{S}(\widetilde{\boldsymbol{\beta}}^{\eta(1)}(k), \widetilde{\boldsymbol{\beta}}^{\eta(2)}(k))\eta$. We denote $\{\widetilde{\boldsymbol{\beta}}_t^{\eta}\}_{t\geq 0}$ as the continuous-time interpolation of $\{\widetilde{\boldsymbol{\beta}}^{\eta}(k)\}_{k\geq 1}$, which satisfies the following SDE,

$$\widetilde{\boldsymbol{\beta}}_t^{\eta} = \widetilde{\boldsymbol{\beta}}_0 - \int_0^t \nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})ds + \int_0^t \widetilde{\Sigma}^{\eta}_{\lfloor s/\eta\rfloor\eta} d\boldsymbol{W}_s. \tag{6}$$

Here the random matrix $\widetilde{\Sigma}^{\eta}_{\lfloor s/\eta\rfloor\eta}$ follows a similar trajectory as $\widetilde{\Sigma}^{\eta}(\lfloor s/\eta\rfloor)$. For $k \in \mathbb{N}^+$ with $t = k\eta$, the relation $\widetilde{\boldsymbol{\beta}}_t^{\eta} = \widetilde{\boldsymbol{\beta}}_{k\eta}^{\eta} = \widetilde{\boldsymbol{\beta}}^{\eta}(k)$ follows.

**Lemma 1** (Discretization error)**.** *Given the smoothness and dissipativity assumptions* (1) *and* (2)*, and the learning rate* $\eta$ *satisfying* $0 < \eta < 1 \wedge a/C^2$*, there exists constants* $D_1, D_2$ *and* $D_3$ *such that*

$$\mathbb{E}[\sup_{0\leq t\leq T} \|\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}_t^{\eta}\|^2] \leq D_1\eta + D_2 \max_k \mathbb{E}[\|\boldsymbol{\phi}_k\|^2] + D_3 \max_k \sqrt{\mathbb{E}\left[|\psi_k|^2\right]}, \tag{7}$$

*where* $D_1$ *depends on* $\tau_1, \tau_2, d, T, C, a, b$*;* $D_2$ *depends on* $T$ *and* $C$*;* $D_3$ *depends on* $r, d, T$ *and* $C$.

**Proof** Based on the replica exchange Langevin diffusion $\{\boldsymbol{\beta}_t\}_{t\geq 0}$ and the continuous-time interpolation of the stochastic gradient Langevin diffusion $\{\widetilde{\boldsymbol{\beta}}_t^{\eta}\}_{t\geq 0}$, we have the following SDE for the difference $\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}_t^{\eta}$. For any $t \in [0, T]$, we have

$$\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}_t^{\eta} = -\int_0^t (\nabla G(\boldsymbol{\beta}_s) - \nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})ds + \int_0^t (\Sigma_s - \widetilde{\Sigma}^{\eta}_{\lfloor s/\eta\rfloor\eta})d\boldsymbol{W}_s$$

Indeed, note that

$$\sup_{0\leq t\leq T} \|\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}_t^{\eta}\| \leq \int_0^T \|\nabla G(\boldsymbol{\beta}_s) - \nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\|)ds + \sup_{0\leq t\leq T}\left\|\int_0^t (\Sigma_s - \widetilde{\Sigma}^{\eta}_{\lfloor s/\eta\rfloor\eta})d\boldsymbol{W}_s\right\|$$

We first square both sides and take expectation, then apply the Burkholder-Davis-Gundy inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}[\sup_{0\leq t\leq T} \|\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}_t^{\eta}\|^2] &\leq 2\mathbb{E}\left[\left(\int_0^T \|\nabla G(\boldsymbol{\beta}_s) - \nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\|ds\right)^2 + \sup_{0\leq t\leq T}\left\|\int_0^t (\Sigma_s - \widetilde{\Sigma}^{\eta}_{\lfloor s/\eta\rfloor\eta})d\boldsymbol{W}_s\right\|^2\right] \\
&\leq \underbrace{2T\mathbb{E}\left[\int_0^T \|\nabla G(\boldsymbol{\beta}_s) - \nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\|^2 ds\right]}_{\mathcal{I}} + \underbrace{8\mathbb{E}\left[\int_0^T \|\Sigma_s - \widetilde{\Sigma}^{\eta}_{\lfloor s/\eta\rfloor\eta}\|^2 ds\right]}_{\mathcal{J}}
\end{aligned} \tag{8}$$

**Estimate of stochastic gradient:** For the first term $\mathcal{I}$, by using the inequality

$$\|a+b+c\|^2 \leq 3(\|a\|^2+\|b\|^2+\|c\|^2),$$

we get

$$\begin{aligned}\mathcal{I} =& 2T\mathbb{E}\left[\int_0^T \left\|\left(\nabla G(\boldsymbol{\beta}_s)-\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_s)\right)+\left(\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_s)-\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\right)+\left(\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})-\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\right)\right\|^2 ds\right]\\ \leq & \underbrace{6T\mathbb{E}\left[\int_0^T \|\nabla G(\boldsymbol{\beta}_s)-\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_s)\|^2 ds\right]}_{\mathcal{I}_1}+\underbrace{6T\mathbb{E}\left[\int_0^T \|\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_s)-\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\|^2 ds\right]}_{\mathcal{I}_2}\\ & +\underbrace{6T\mathbb{E}\left[\int_0^T \|\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})-\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta})\|^2 ds\right]}_{\mathcal{I}_3}\\ \leq & \mathcal{I}_1+\mathcal{I}_2+\mathcal{I}_3.\end{aligned} \tag{9}$$

By using the smoothness assumption 1, we first estimate

$$\mathcal{I}_1 \leq 6TC^2\mathbb{E}\left[\int_0^T \|\boldsymbol{\beta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_s\|^2 ds\right].$$

By applying the smoothness assumption 1 and discretization scheme, we can further estimate

$$\begin{aligned}\mathcal{I}_2 &\leq 6TC^2\mathbb{E}\left[\int_0^T \|\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta}\|^2 ds\right]\\ &\leq 6TC^2\sum_{k=0}^{\lfloor T/\eta\rfloor}\mathbb{E}\left[\int_{k\eta}^{(k+1)\eta}\|\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta}\|^2 ds\right]\\ &\leq 6TC^2\sum_{k=0}^{\lfloor T/\eta\rfloor}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\sup_{k\eta\leq s<(k+1)\eta}\|\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta}\|^2\right]ds\end{aligned} \tag{10}$$

For $\forall\ k\in\mathbb{N}$ and $s\in[k\eta,(k+1)\eta)$, we have

$$\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta}=\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta}=-\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})\cdot(s-k\eta)+\widetilde{\Sigma}^{\eta}_{k\eta}\int_{k\eta}^{s} d\boldsymbol{W}_r$$

which indeed implies

$$\sup_{k\eta\leq s<(k+1)\eta}\|\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta}\|\leq\|\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})\|(s-k\eta)+\sup_{k\eta\leq s<(k+1)\eta}\|\widetilde{\Sigma}^{\eta}_{k\eta}\int_{k\eta}^{s} d\boldsymbol{W}_r\|$$

Similar to the estimate (8), square both sides and take expectation, then apply the Burkholder-Davis-Gundy inequality, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{k\eta\leq s<(k+1)\eta}\|\widetilde{\boldsymbol{\beta}}^{\eta}_s-\widetilde{\boldsymbol{\beta}}^{\eta}_{\lfloor s/\eta\rfloor\eta}\|^2\right] &\leq 2\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})\|^2(s-k\eta)^2]+8\sum_{j=1}^{2d}\mathbb{E}\left[\left(\widetilde{\Sigma}^{\eta}_{k\eta}(j)\langle\int_{k\eta}^{\cdot} d\boldsymbol{W}_r\rangle_s^{1/2}\right)^2\right]\\ &\leq 2(s-k\eta)^2\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})\|^2]+32d\tau_2(s-k\eta),\end{aligned}$$

where the last inequality follows from the fact that $\widetilde{\Sigma}^{\eta}_{k\eta}$ is a diagonal matrix with diagonal elements $\sqrt{2\tau_1}$ or $\sqrt{2\tau_2}$. For the first term in the above inequality, we further have

$$\begin{aligned}2(s-k\eta)^2\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})\|^2] &= 2(s-k\eta)^2\mathbb{E}[\|(\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})+\boldsymbol{\phi}_k)\|^2]\\ &\leq 4\eta^2\mathbb{E}[\|\nabla G(\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta})-\nabla G(\boldsymbol{\beta}^*)\|^2+\|\boldsymbol{\phi}_k\|^2]\\ &\leq 8C^2\eta^2\mathbb{E}[\|\widetilde{\boldsymbol{\beta}}^{\eta}_{k\eta}\|^2+\|\boldsymbol{\beta}^*\|^2]+4\eta^2\mathbb{E}[\|\boldsymbol{\phi}_k\|^2],\end{aligned}$$

where the first inequality follows from the separation of the noise from the stochastic gradient and the choice of stationary point $\boldsymbol{\beta}^*$ of $G(\cdot)$ with $\nabla G(\boldsymbol{\beta}^*) = 0$, and $\boldsymbol{\phi}_k$ is the stochastic noise in the gradient at step $k$. Thus, combining the above two parts and integrate $\mathbb{E}\left[\sup_{k\eta\le s<(k+1)\eta}\|\widetilde{\boldsymbol{\beta}}_s^\eta - \widetilde{\boldsymbol{\beta}}_{k\eta}^\eta\|^2\right]$ on the time interval $[k\eta,(k+1)\eta)$, we obtain the following bound

$$\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\sup_{k\eta\le s<(k+1)\eta} \|\widetilde{\boldsymbol{\beta}}_s^\eta - \widetilde{\boldsymbol{\beta}}_{k\eta}^\eta\|^2\right] ds \le 8C^2\eta^3 \left(\sup_{k\ge 0}\mathbb{E}[\|\widetilde{\boldsymbol{\beta}}^\eta_{k\eta}\|^2 + \|\boldsymbol{\beta}^*\|^2]\right) + 4\eta^3\max_k \mathbb{E}[\|\boldsymbol{\phi}_k\|^2] + 32d\tau_2\eta^2 \tag{11}$$

By plugging the estimate (11) into estimate (10), we obtain the following estimates when $\eta \le 1$,

$$\begin{aligned} \mathcal{I}_2 &\le 6TC^2(1+T/\eta)\left[8C^2\eta^3\left(\sup_{k\ge 0}\mathbb{E}[\|\widetilde{\boldsymbol{\beta}}^\eta_{k\eta}\|^2 + \|\boldsymbol{\beta}^*\|^2]\right) + 4\eta^3\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2] + 32d\tau_2\eta^2\right] \\ &\le \tilde{\delta}_1(d,\tau_2,T,C,a,b)\eta + 24TC^2(1+T)\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2], \end{aligned} \tag{12}$$

where $\tilde{\delta}_1(d,\tau_2,T,C,a,b)$ is a constant depending on $d,\tau_2,T,C,a$ and $b$. Note that the above inequality requires a result on the bounded second moment of $\sup_{k\ge 0}\mathbb{E}[\|\widetilde{\boldsymbol{\beta}}^\eta_{k\eta}\|^2]$, and this is proved in Lemma $C.2$ in Chen et al. (2019) when we choose the stepzise $\eta \in (0, a/C^2)$. We are now left to estimate the term $\mathcal{I}_3$ and we have

$$\begin{aligned} \mathcal{I}_3 &\le 6T\sum_{k=0}^{\lfloor T/\eta\rfloor}\mathbb{E}\left[\int_{k\eta}^{(k+1)\eta}\|\nabla G(\widetilde{\boldsymbol{\beta}}^\eta_{k\eta}) - \nabla\widetilde{G}(\widetilde{\boldsymbol{\beta}}^\eta_{k\eta})\|^2 ds\right] \\ &\le 6T(1+T/\eta)\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]\eta \\ &\le 6T(1+T)\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]. \end{aligned} \tag{13}$$

Combing all the estimates of $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{I}_3$, we obtain

$$\begin{aligned} \mathcal{I} \le & \underbrace{6TC^2\int_0^T\mathbb{E}\left[\sup_{0\le s\le T}\|\boldsymbol{\beta}_s - \widetilde{\boldsymbol{\beta}}^\eta_s\|^2\right]ds}_{\mathcal{I}_1} + \underbrace{\tilde{\delta}_1(d,\tau_2,T,C,a,b)\eta + 24TC^2(1+T)\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]}_{\mathcal{I}_2} \\ & \underbrace{+\,6T(1+T)\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]}_{\mathcal{I}_3}. \end{aligned} \tag{14}$$

**Estimate of stochastic diffusion:** For the second term $\mathcal{J}$, we have

$$\begin{aligned} \mathcal{J} &= 8\mathbb{E}\left[\int_0^T\|\Sigma_s(j) - \widetilde{\Sigma}_{\lfloor s/\eta\rfloor\eta}(j)\|^2 ds\right] \\ &\le 8\sum_{j=1}^{2d}\sum_{k=0}^{\lfloor T/\eta\rfloor}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\Sigma_s(j) - \widetilde{\Sigma}^\eta_{k\eta}(j)\|^2\right]ds \\ &\le 8\sum_{j=1}^{2d}\sum_{k=0}^{\lfloor T/\eta\rfloor}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\Sigma_s(j) - \Sigma^\eta_{k\eta}(j) + \Sigma^\eta_{k\eta}(j) - \widetilde{\Sigma}^\eta_{k\eta}(j)\|^2\right]ds \\ &\le 16\sum_{j=1}^{2d}\sum_{k=0}^{\lfloor T/\eta\rfloor}\left[\underbrace{\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\Sigma_s(j) - \Sigma^\eta_{k\eta}(j)\|^2\right]ds}_{\mathcal{J}_1} + \underbrace{\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\Sigma^\eta_{k\eta}(j) - \widetilde{\Sigma}^\eta_{k\eta}(j)\|^2\right]ds}_{\mathcal{J}_2}\right] \end{aligned}. \tag{15}$$

where $\Sigma^\eta_{k\eta}$ is the temperature matrix for the continuous-time interpolation of $\{\boldsymbol{\beta}^\eta(k)\}_{k\ge 1}$, which is similar to (6) without noise generated from mini-batch settings and is defined as below

$$\boldsymbol{\beta}^\eta_t = \boldsymbol{\beta}_0 - \int_0^t\nabla G(\boldsymbol{\beta}^\eta_{k\eta})ds + \int_0^t\Sigma^\eta_{k\eta}d\boldsymbol{W}_s. \tag{16}$$

We estimate $\mathcal{J}_1$ first, considering that $\Sigma_s$ and $\Sigma^{\eta}_{\lfloor s/\eta \rfloor \eta}$ are both diagonal matrices, we have

$$\begin{aligned}
\mathcal{J}_1 &= 4(\sqrt{\tau_2}-\sqrt{\tau_1})^2 \int_{k\eta}^{(k+1)\eta} \mathbb{P}(\Sigma_s(j) \neq \Sigma^{\eta}_{k\eta}(j)) ds \\
&= 4(\sqrt{\tau_2}-\sqrt{\tau_1})^2 \mathbb{E}\left[\int_{k\eta}^{(k+1)\eta} \mathbb{P}(\Sigma_s(j) \neq \Sigma^{\eta}_{k\eta}(j) | \boldsymbol{\beta}^{\eta}_{k\eta}) ds\right] \\
&= 4(\sqrt{\tau_2}-\sqrt{\tau_1})^2 r \int_{k\eta}^{(k+1)\eta} [(s-k\eta) + \mathcal{R}(s-k\eta)] ds \\
&\leq \tilde{\delta}_2(r, \tau_1, \tau_2)\eta^2,
\end{aligned}$$

where $\tilde{\delta}_2(r, \tau_1, \tau_2) = 4(\sqrt{\tau_2}-\sqrt{\tau_1})^2 r$, and the equality follows from the fact that the conditional probability $\mathbb{P}(\Sigma_s(j) \neq \Sigma^{\eta}_{k\eta}(j)|\boldsymbol{\beta}^{\eta}_{k\eta}) = rS(\boldsymbol{\beta}^{\eta(1)}_{k\eta}, \boldsymbol{\beta}^{\eta(2)}_{k\eta}) \cdot (s-\eta) + r\mathcal{R}(s-k\eta)$. Here $\mathcal{R}(s-k\eta)$ denotes the higher remainder with respect to $s-k\eta$. The estimate of $\mathcal{J}_1$ without stochastic gradient for the Langevin diffusion is first obtained in Chen et al. (2019), we however present here again for reader's convenience. As for the second term $\mathcal{J}_2$, it follows that

$$\begin{aligned}
\mathcal{J}_2 &= 4(\sqrt{\tau_2}-\sqrt{\tau_1})^2 \int_{k\eta}^{(k+1)\eta} \mathbb{P}(\Sigma_{k\eta}(j) \neq \widetilde{\Sigma}_{k\eta}(j)) ds \\
&= 4(\sqrt{\tau_2}-\sqrt{\tau_1})^2 r\eta \mathbb{E}\left[\left|S(\boldsymbol{\beta}^{\eta(1)}_{k\eta}, \boldsymbol{\beta}^{\eta(2)}_{k\eta}) - \tilde{S}(\widetilde{\boldsymbol{\beta}}^{\eta(1)}_{k\eta}, \widetilde{\boldsymbol{\beta}}^{\eta(2)}_{k\eta})\right|\right] \\
&\leq \tilde{\delta}_2(r, \tau_1, \tau_2)\eta \sqrt{\mathbb{E}\left[\left|S(\boldsymbol{\beta}^{\eta(1)}_{k\eta}, \boldsymbol{\beta}^{\eta(2)}_{k\eta}) - \tilde{S}(\widetilde{\boldsymbol{\beta}}^{\eta(1)}_{k\eta}, \widetilde{\boldsymbol{\beta}}^{\eta(2)}_{k\eta})\right|^2\right]} \\
&\leq \tilde{\delta}_2(r, \tau_1, \tau_2)\eta \sqrt{\mathbb{E}\left[|\psi_k|^2\right]},
\end{aligned} \tag{17}$$

where $\psi_k$ is the noise in the swapping rate. Thus, one concludes the following estimates combing $\mathcal{I}$ and $\mathcal{J}$.

$$\begin{aligned}
\mathbb{E}[\sup_{0\leq t\leq T} \|\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}^{\eta}_t\|^2] \leq & \underbrace{6TC^2 \int_0^T \mathbb{E}\left[\sup_{0\leq s\leq T} \|\boldsymbol{\beta}_s - \widetilde{\boldsymbol{\beta}}^{\eta}_s\|^2\right] ds}_{\mathcal{I}_1} + \underbrace{\tilde{\delta}_1(d, \tau_2, T, C, a, b)\eta + 24TC^2(\eta+T)\max_k \mathbb{E}[\|\boldsymbol{\phi}_k\|^2]}_{\mathcal{I}_2} \\
& + \underbrace{6T(1+T)\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]}_{\mathcal{I}_3} + \underbrace{32d(1+T)\tilde{\delta}_2(r, \tau_1, \tau_2)\left(\eta + \max_k \sqrt{\mathbb{E}\left[|\psi_k|^2\right]}\right)}_{\mathcal{J}}.
\end{aligned} \tag{18}$$

Apply Gronwall's inequality to the function

$$t \mapsto \mathbb{E}\left[\sup_{0\leq u\leq t} \|\boldsymbol{\beta}_u - \widetilde{\boldsymbol{\beta}}^{\eta}_u\|^2\right],$$

and deduce that

$$\mathbb{E}[\sup_{0\leq t\leq T} \|\boldsymbol{\beta}_t - \widetilde{\boldsymbol{\beta}}^{\eta}_t\|^2] \leq D_1\eta + D_2 \max_k \mathbb{E}[\|\boldsymbol{\phi}_k\|^2] + D_3 \max_k \sqrt{\mathbb{E}\left[|\psi_k|^2\right]}, \tag{19}$$

where $D_1$ is a constant depending on $\tau_1, \tau_2, d, T, C, a, b$; $D_2$ depends on $T$ and $C$; $D_3$ depends on $r, d, T$ and $C$.

■

### 1.4. Exponential decay of Wasserstein distance in continuous-time

We proceed to quantify the evolution of the 2-Wasserstein distance between $\nu_t$ and $\pi$. We first consider the ordinary Langevin diffusion without swaps and derive the log-Sobolev inequality (LSI). Then we extend LSI to reLD and obtain the exponential decay of the relative entropy. Finally, we derive the exponential decay of the 2-Wasserstein distance.

In order to distinguish from the replica exchange Langevin diffusion $\boldsymbol{\beta}_t$ defined in (4), we call it $\hat{\boldsymbol{\beta}}_t$ which follows,

$$d\hat{\boldsymbol{\beta}}_t = -\nabla G(\hat{\boldsymbol{\beta}}_t)dt + \Sigma_t d\boldsymbol{W}_t. \tag{20}$$

where $\Sigma_t \in \mathbb{R}^{2d\times 2d}$ is a diagonal matrix with the form $\begin{pmatrix} \sqrt{2\tau_1}\mathbf{I}_d & 0 \\ 0 & \sqrt{2\tau_2}\mathbf{I}_d \end{pmatrix}$. The process $\hat{\boldsymbol{\beta}}_t$ is a Markov diffusion process with infinitesimal generator $\mathcal{L}$ in the following form, for $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}^d$,

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

Note that since matrix $\Sigma_t$ is a non-degenerate diagonal matrix, operator $\mathcal{L}$ is an elliptic diffusion operator. According to the smoothness assumption (1), we have that $\nabla^2 G \geq -C\mathbf{I}_{2d}$, where $C > 0$, the unique invariant measure $\pi$ associate with the underlying diffusion process satisfies the Poincare inequality and LSI with the Dirichlet form given as follows,

$$\mathcal{E}(f) = \int \Big(\tau_1 \|\nabla_{x_1} f\|^2 + \tau_2 \|\nabla_{x_2} f\|^2\Big) d\pi(x_1,x_2), \qquad f \in \mathcal{C}_0^2(\mathbb{R}^{2d}). \tag{21}$$

In this elliptic case with $G$ being convex, the proof for LSI follows from standard Bakry-Emery calculus (Bakry and émery, 1985). Since, we are dealing with the non-convex function $G$, we are particularly interested in the case of $\nabla^2 G \geq -C\mathbf{I}_{2d}$. To obtain a Poincaré inequality for invariant measure $\pi$, Chen et al. (2019) adapted an argument from Bakry et al. (2008) and Raginsky et al. (2017) by constructing an appropriate Lyapunov function for the replica exchange diffusion without swapping $\hat{\boldsymbol{\beta}}_t$. Denote $\nu_t$ as the distribution associated with the diffusion process $\{\hat{\boldsymbol{\beta}}_t\}_{t\geq 0}$, which is absolutely continuous with respect to $\pi$. It is a direct consequence of the aforementioned results that the following log-Sobolev inequality holds.

**Lemma 2** (LSI for Langevin Diffusion)**.** *Under assumptions* (1) *and* (2)*, we have the following log-Sobolev inequality for invariant measure $\pi$, for some constant $c_{LS} > 0$,*

$$D(\nu_t||\pi) \leq 2c_{LS}\mathcal{E}(\sqrt{\frac{d\nu_t}{d\pi}}).$$

*where $D(\nu_t||\pi) = \int d\nu_t \log \frac{d\nu_t}{d\pi}$ denotes the relative entropy and the Dirichlet form $\mathcal{E}(\cdot)$ is defined in* (21).

**Proof**

According to Cattiaux et al. (2010), the sufficient conditions to establish LSI are:

1. There exists some constant $C \geq 0$, such that $\nabla^2 G \succcurlyeq -CI_{2d}$.

2. $\pi$ satisfies a Poincaré inequality with constant $c_p$, namely, for all probability measures $\nu \ll \pi$, $\chi^2(\nu||\pi) \leq c_p\mathcal{E}(\sqrt{\frac{d\nu_t}{d\pi}})$, where $\chi^2(\nu||\pi) := \|\frac{d\nu}{d\pi} - 1\|^2$ is the $\chi^2$ divergence between $\nu$ and $\pi$.

3. There exists a $\mathcal{C}^2$ Lyapunov function $V : \mathbb{R}^{2d} \to [1,\infty)$ such that $\frac{\mathcal{L}V(x_1,x_2)}{V(x_1,x_2)} \leq \kappa - \gamma(\|x_1\|^2 + \|x_2\|^2)$ for all $(x_1,x_2) \in \mathbb{R}^{2d}$ and some $\kappa, \gamma > 0$.

Note that the first condition on the Hessian is obtained from the smoothness assumption (1). Moreover, the Poincaré inequality in the second condition is derived from Lemma C.1 in Chen et al. (2019) given assumptions (1) and (2). Finally, to verify the third condition, we follow Raginsky et al. (2017) and construct the Lyapunov function $V(x_1, x_2) :=$

$\exp\left\{a/4 \cdot \left(\frac{\|x_1\|^2}{\tau_1} + \frac{\|x_2\|^2}{\tau_2}\right)\right\}$. From the dissipitive assumption 2, $V(x_1, x_2)$ satisfies the third condition because

$$\begin{aligned}\mathcal{L}(V(x_1, x_2)) &= \left(\frac{a}{2\tau_1} + \frac{a}{2\tau_2} + \frac{a^2}{4\tau_1^2}\|x_1\|^2 + \frac{a^2}{4\tau_2^2}\|x_2\|^2 - \frac{a}{2\tau_1^2}\langle x_1, \nabla G(x_1) - \frac{a}{2\tau_2^2}\langle x_1, \nabla G(x_2)\rangle\right) V(x_1, x_2) \\ &\leq \left(\frac{a}{2\tau_1} + \frac{a}{2\tau_2} + \frac{ab}{2\tau_1^2} + \frac{ab}{2\tau_2^2} - \frac{a^2}{4\tau_1^2}\|x_1\|^2 - \frac{a^2}{4\tau_2^2}\|x_2\|^2\right) V(x_1, x_2) \\ &\leq \left(\kappa - \gamma(\|x_1\|^2 + \|x_2\|^2)\right) V(x_1, x_2),\end{aligned} \tag{22}$$

where $\kappa = \frac{a}{2\tau_1} + \frac{a}{2\tau_2} + \frac{ab}{2\tau_1^2} + \frac{ab}{2\tau_2^2}$, and $\gamma = \frac{a^2}{4\tau_1^2} \wedge \frac{a^2}{4\tau_2^2}$. [2] Therefore, the invariant measure $\pi$ satisfies a LSI with the constant

$$c_{\text{LS}} = c_1 + (c_2 + 2)c_p, \tag{23}$$

where $c_1 = \frac{2C}{\gamma} + \frac{2}{C}$ and $c_2 = \frac{2C}{\gamma}\left(\kappa + \gamma\int_{\mathbb{R}^{2d}}(\|x_1\|^2 + \|x_2\|^2)\pi(dx_1 dx_2)\right)$.

■

We are now ready to prove the log-Sobolev inequality for invariant measure associated with the replica exchange Langevin diffusion (4). We use a similar idea from Chen et al. (2019) where they prove the Poincaré inequality for the invariant measure associated with the replica exchange Langevin diffusion (4) by analyzing the corresponding Dirichlet form. In particular, a larger Dirichlet form ensures a smaller log-Sobolev constant and hence results in a faster convergence in the relative entropy and Wasserstein distance.

**Lemma 3** (Accelerated exponential decay of $\mathcal{W}_2$)**.** *Under assumptions* (1) *and* (2)*, we have that the replica exchange Langevin diffusion converges exponentially fast to the invariant distribution* $\pi$*:*

$$\mathcal{W}_2(\nu_t, \pi) \leq D_0 e^{-k\eta(1+\delta_S)/c_{LS}}, \tag{24}$$

*where* $D_0 = \sqrt{2c_{LS}D(\nu_0||\pi)}$, $\delta_S := \inf_{t>0}\frac{\mathcal{E}_S(\sqrt{\frac{d\nu_t}{d\pi}})}{\mathcal{E}(\sqrt{\frac{d\nu_t}{d\pi}})} - 1$ *is a non-negative constant depending on the swapping rate* $S(\cdot, \cdot)$ *and obtains* $0$ *only if* $S(\cdot, \cdot) = 0$.

**Proof** Consider the infinitesimal generator associated with the diffusion process (4), denoted as $\mathcal{L}_S$, contains an extra term arising from the temperature swapping. The operator $\mathcal{L}_S$ in this particular case, indeed, has the following form

$$\mathcal{L}_S = \mathcal{L} + S(x_1, x_2) \cdot (f(x_2, x_1) - f(x_1, x_2)). \tag{25}$$

According to Theorem 3.3 (Chen et al., 2019), the Dirichlet form associated with operator $\mathcal{L}_S$ under the invariant measure $\pi$ has the form

$$\mathcal{E}_S(f) = \mathcal{E}(f) + \underbrace{\frac{1}{2}\int S(x_1, x_2) \cdot (f(x_2, x_1) - f(x_1, x_2))^2 d\pi(x_1, x_2)}_{\text{acceleration}}, \ f \in \mathcal{C}_0^2(\mathbb{R}^{2d}), \tag{26}$$

where $f$ corresponds to $\frac{d\nu_t}{d\pi(x_1, x_2)}$, and the asymmetry of $\frac{\nu_t}{\pi(x_1, x_2)}$ is critical in the acceleration effect (Chen et al., 2019). Given two different temperatures $\tau_1$ and $\tau_2$, a non-trivial distribution $\pi$ and function $f$, the swapping rate $S(x_1, x_2)$ is positive for almost any $x_1, x_2 \in \mathbb{R}^d$. As a result, the Dirichlet form associated with $\mathcal{L}_S$ is strictly larger than $\mathcal{L}$. Therefore, there exists a constant $\delta_S > 0$ depending on $S(x_1, x_2)$, such that $\delta_S = \inf_{t>0}\frac{\mathcal{E}_S(\sqrt{\frac{d\nu_t}{d\pi}})}{\mathcal{E}(\sqrt{\frac{d\nu_t}{d\pi}})} - 1$. From Lemma 2, we have

$$D(\nu_t||\pi) \leq 2c_{\text{LS}}\mathcal{E}(\sqrt{\frac{d\nu_t}{d\pi}}) \leq 2c_{\text{LS}}\sup_t \frac{\mathcal{E}(\sqrt{\frac{d\nu_t}{d\pi}})}{\mathcal{E}_S(\sqrt{\frac{d\nu_t}{d\pi}})}\mathcal{E}_S(\sqrt{\frac{d\nu_t}{d\pi}}) = 2\frac{c_{\text{LS}}}{1+\delta_S}\mathcal{E}_S(\sqrt{\frac{d\nu_t}{d\pi}}). \tag{27}$$

[2] $a \wedge b$ denotes $\min\{a, b\}$.

Thus, we obtain the following log-Sobolev inequality for the unique invariant measure $\pi$ associated with replica exchange Langevin diffusion $\{\boldsymbol{\beta}_t\}_{t\geq 0}$ and its corresponding Dirichlet form $\mathcal{E}_S(\cdot)$. In particular, the LSI constant $\frac{c_{\mathrm{LS}}}{1+\delta_S}$ in replica exchange Langevin diffusion with swapping rate $S(\cdot,\cdot)>0$ is strictly smaller than the LSI constant $c_{\mathrm{LS}}$ in the replica exchange Langevin diffusion with swapping rate $S(\cdot,\cdot)=0$. By the exponential decay in entropy (Bakry et al., 2014)[Theorem 5.2.1] and the tight log-Sobolev inequality in Lemma 2, we get that, for any $t\in[k\eta,(k+1)\eta)$,

$$D(\nu_t||\pi)\leq D(\nu_0||\pi)e^{-2t(1+\delta_S)/c_{\mathrm{LS}}}\leq D(\mu_0||\pi)e^{-2k\eta(1+\delta_S)/c_{\mathrm{LS}}}. \tag{28}$$

Finally, we can estimate the term $\mathcal{W}_2(\nu_t,\pi)$ by the Otto-Villani theorem (Bakry et al., 2014)[Theorem 9.6.1],

$$\mathcal{W}_2(\nu_t,\pi)\leq\sqrt{2c_{\mathrm{LS}}D(\nu_t||\pi)}\leq\sqrt{2c_{\mathrm{LS}}D(\mu_0||\pi)}e^{-k\eta(1+\delta_S)/c_{\mathrm{LS}}}. \tag{29}$$

■

### 1.5. Summary: Convergence of reSGLD

Now that we have all the necessary ingredients in place, we are ready to derive the convergence of the distribution $\mu_k$ to the invariant measure $\pi$ in terms of 2-Wasserstein distance,

**Theorem 1** (Convergence of reSGLD). *Let the assumptions* (1) *and* (2) *hold. For the unique invariant measure* $\pi$ *associated with the Markov diffusion process* (4) *and the distribution* $\{\mu_k\}_{k\geq 0}$ *associated with the discrete dynamics* $\{\widetilde{\boldsymbol{\beta}}^{\eta}(k)\}_{k\geq 1}$*, we have the following estimates, for* $0\leq k\in\mathbb{N}^+$ *and the learning rate* $\eta$ *satisfying* $0<\eta<1\wedge a/C^2$*,*

$$\mathcal{W}_2(\mu_k,\pi)\leq D_0e^{-k\eta(1+\delta_S)/c_{LS}}+\sqrt{\delta_1\eta+\delta_2\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]+\delta_3\max_k\sqrt{\mathbb{E}\left[|\psi_k|^2\right]}} \tag{30}$$

*where* $D_0=\sqrt{2c_{LS}D(\mu_0||\pi)}$*,* $\delta_S:=\min_k\frac{\mathcal{E}_S(\sqrt{\frac{d\mu_k}{d\pi}})}{\mathcal{E}(\sqrt{\frac{d\mu_k}{d\pi}})}-1$ *is a non-negative constant depending on the swapping rate* $S(\cdot,\cdot)$ *and obtains the minimum zero only if* $S(\cdot,\cdot)=0$*.*

**Proof** We reduce the estimates into the following two terms by using the triangle inequality,

$$\mathcal{W}_2(\mu_k,\pi)\leq\mathcal{W}_2(\mu_k,\nu_t)+\mathcal{W}_2(\nu_t,\pi),\qquad t\in[k\eta,(k+1)\eta). \tag{31}$$

The first term $\mathcal{W}_2(\mu_k,\nu_t)$ follows from the analysis of discretization error in Lemma.1. Recall the very definition of the $\mathcal{W}_2(\cdot,\cdot)$ distance defined in (3). Thus, in order to control the distance $\mathcal{W}_2(\mu_k,\nu_t)$, $t\in[k\eta,(k+1)\eta)$, we need to consider the diffusion process whose law give $\mu_k$ and $\nu_t$, respectively. Indeed, it is obvious that $\nu_t=\mathcal{L}(\boldsymbol{\beta}_t)$ for $t\in[k\eta,(k+1)\eta)$. For the other measure $\mu_k$, it follows that $\mu_k=\tilde{\nu}_{k\eta}$ for $t=k\eta$, where $\tilde{\nu}_{k\eta}=\mathcal{L}(\widetilde{\boldsymbol{\beta}}_t^{\eta})$ is the probability measure associated with the continuous interpolation of reSGLD (5). By Lemma.1, we have that for $k\in\mathbb{N}$ and $t\in[k\eta,(k+1)\eta)$,

$$\mathcal{W}_2(\mu_k,\nu_t)=\mathcal{W}_2(\tilde{\nu}_{k\eta},\nu_t)\leq\sqrt{\mathbb{E}[\sup_{0\leq s\leq t}\|\boldsymbol{\beta}_s-\widetilde{\boldsymbol{\beta}}_s^{\eta}\|^2]}\leq\sqrt{\delta_1\eta+\delta_2\max_k\mathbb{E}[\|\boldsymbol{\phi}_k\|^2]+\delta_3\max_k\sqrt{\mathbb{E}\left[|\psi_k|^2\right]}}, \tag{32}$$

Recall from the accelerated exponential decay of replica exchange Langevin diffusion in Lemma.3, we have

$$\mathcal{W}_2(\nu_t,\pi)\leq\sqrt{2c_{\mathrm{LS}}D(\nu_0||\pi)}e^{-k\eta(1+\delta_S)/c_{\mathrm{LS}}}=\sqrt{2c_{\mathrm{LS}}D(\mu_0||\pi)}e^{-k\eta(1+\delta_S)/c_{\mathrm{LS}}}. \tag{33}$$

Combing the above two estimates completes the proof. ■

## 2. Hyper-parameter Setting for Bayesian GANs

In the semi-supervised learning tasks, we fine-tune the hyper-parameters for Bayesian GANs and report them in Table 1. In particular, $N_s$ is the number of labeled data; $\eta^{(1)}$ and $\eta^{(2)}$ are the learning rates for the low-temperature chain and high-temperature chain, respectively; $\tau_1$ and $\tau_2$ are the temperatures; $\hat{F}$ is the correction factor, which often yields several swaps. In addition, the learning rates also follow a truncated exponential decay, for example, $\eta_k^{(1)} = \left(0.05 \vee e^{-\frac{k}{800}}\right) \eta^{(1)}$ and $\eta_k^{(2)} = \left(0.05 \vee e^{-\frac{k}{800}}\right) \eta^{(2)}$, where $k$ is the number of iterations.

| Dataset | $N_s$ | $\eta^{(1)}$ | $\eta^{(2)}$ | $\tau_1$ | $\tau_2$ | $\hat{F}$ |
|---|---|---|---|---|---|---|
| CIFAR10 | $2000 \sim 3500$ | 4.5e-4 | 7.0e-4 | 0.01 | 1 | 3.0e5 |
| | $4000 \sim 5000$ | 4.5e-4 | 7.0e-4 | 0.01 | 1 | 2.0e5 |
| CIFAR100 | $2000 \sim 2500$ | 5.0e-4 | 7.5e-4 | 0.04 | 1 | 1.0e4 |
| | $3000 \sim 3500$ | 5.0e-4 | 7.5e-4 | 0.02 | 1 | 2.5e4 |
| | $4000 \sim 5000$ | 5.0e-4 | 7.5e-4 | 0.01 | 1 | 5.0e4 |
| SVHN | $2000 \sim 4000$ | 4.5e-3 | 5.0e-3 | 0.01 | 1 | 8.0e4 |
| | $4500 \sim 5000$ | 4.5e-3 | 7.0e-3 | 0.01 | 1 | 8.0e4 |

*Table 1.* Hyper-parameter setting of Bayesian GANs for Semi-Supervised Learning experiments.

## References

Dominique Bakry and Michel émery. Diffusions Hypercontractives. *Séminaire de Probabilités XIX 1983/84*, pages 177–206, 1985.

Dominique Bakry, Patrick Cattiaux F. Barthe, and Arnaud Guillin. A Simple Proof of the Poincaré Inequality for A Large Class of Probability Measures. *Electron. Comm. Probab.*, 13:60–66, 2008.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. Analysis and Geometry of Markov Diffusion Operators. *Springer*, 2014.

Patrick Cattiaux, Arnaud Guillin, and Li-Ming Wu. A Note on Talagrand's Transportation Inequality and Logarithmic Sobolev Inequality. *Prob. Theory and Rel. Fields*, 148:285–334, 2010.

Yi Chen, Jinglin Chen, Jing Dong, Jian Peng, and Zhaoran Wang. Accelerating Nonconvex Learning via Replica Exchange Langevin Diffusion. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2019.

Paul Dupuis, Yufei Liu, Nuria Plattner, and J. D. Doll. On the Infinite Swapping Limit for Parallel Tempering. *SIAM J. Multiscale Modeling & Simulation*, 10, 2012.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex Learning via Stochastic Gradient Langevin Dynamics: a Nonasymptotic Analysis. In *Proc. of Conference on Learning Theory (COLT)*, June 2017.

Issei Sato and Hiroshi Nakagawa. Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Itô Process. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 982–990, 2014.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

George Yin and Chao Zhu. *Hybrid Switching Diffusions: Properties and Applications*. Springer, 2010.