Non-reversible Parallel Tempering for Deep Posterior Approximation

Wei Deng[*1] [2], **Qian Zhang**[*1], **Qi Feng**[*3], **Faming Liang** [1], **Guang Lin** [1]

[1] Purdue University, West Lafayette, IN
[2] Morgan Stanley, New York, NY
[3] University of Michigan, Ann Arbor, MI
weideng056@gmail.com, guanglin@purdue.edu

Abstract

Parallel tempering (PT), also known as replica exchange, is the go-to workhorse for simulations of multi-modal distributions. The key to the success of PT is to adopt efficient swap schemes. The popular deterministic even-odd (DEO) scheme exploits the non-reversibility property and has successfully reduced the communication cost from quadratic to linear given the sufficiently many P chains. However, such an innovation largely disappears in big data due to the limited chains and few bias-corrected swaps. To handle this issue, we generalize the DEO scheme to promote non-reversibility and propose a few solutions to tackle the underlying bias caused by the geometric stopping time. Notably, in big data scenarios, we obtain a nearly linear communication cost based on the optimal window size. In addition, we also adopt stochastic gradient descent (SGD) with large and constant learning rates as exploration kernels. Such a user-friendly nature enables us to conduct approximation tasks for complex posteriors without much tuning costs.

Introduction

Langevin diffusion is a standard sampling algorithm that follows a stochastic differential equation

$$d\boldsymbol{\beta}_t = -\nabla U(\boldsymbol{\beta}_t)dt + \sqrt{2\tau}d\boldsymbol{W}_t,$$

where $\boldsymbol{\beta}_t \in \mathbb{R}^d$, $U(\cdot)$ is the energy function $U(\cdot)$, $\boldsymbol{W}_t \in \mathbb{R}^d$ is a Brownian motion, and τ is the temperature. The diffusion process converges to a stationary distribution $\pi(\boldsymbol{\beta}) \propto e^{-\frac{U(\boldsymbol{\beta})}{\tau}}$ and setting $\tau = 1$ yields a Bayesian posterior. A convex $U(\cdot)$ leads to a rapid convergence (Dalalyan 2017); however, a non-convex $U(\cdot)$ inevitably slows down the mixing rate (Raginsky, Rakhlin, and Telgarsky 2017; Deng et al. 2022; Deng, Lin, and Liang 2022). To accelerate simulations, replica exchange Langevin diffusion (reLD) proposes to include a high-temperature particle $\boldsymbol{\beta}_t^{(P)}$, where $P \in \mathbb{N}^+ \setminus \{1\}$, for *exploration*. Meanwhile, a low-temperature particle $\boldsymbol{\beta}_t^{(1)}$ is presented for *exploitation*:

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

where $\tau^{(P)} > \tau^{(1)}$ and $\boldsymbol{W}_t^{(P)}$ is independent of $\boldsymbol{W}_t^{(1)}$. To promote more explorations for the low-temperature particle, the particles at the position $(\beta^{(1)}, \beta^{(P)}) \in \mathbb{R}^{2d}$ swap with a probability $aS(\beta^{(1)}, \beta^{(P)})$, where

$$S(\beta^{(1)}, \beta^{(P)}) = 1 \wedge e^{\left(\frac{1}{\tau^{(1)}} - \frac{1}{\tau^{(P)}}\right)\left(U(\beta^{(1)}) - U(\beta^{(P)})\right)}, \tag{2}$$

and $a \in (0, \infty)$ is the swap intensity. To be specific, the conditional swap rate at time t follows that

$$\begin{aligned} &\mathbb{P}(\boldsymbol{\beta}_{t+dt} = (\beta^{(P)}, \beta^{(1)}) | \boldsymbol{\beta}_t = (\beta^{(1)}, \beta^{(P)})) \\ &\quad = aS(\beta^{(1)}, \beta^{(P)})dt. \end{aligned}$$

In the longtime limit, the Markov jump process converges to the joint distribution $\pi(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(P)}) \propto e^{-\frac{U(\boldsymbol{\beta}^{(1)})}{\tau^{(1)}} - \frac{U(\boldsymbol{\beta}^{(P)})}{\tau^{(P)}}}$, where the marginals are denote by $\pi^{(1)}(\boldsymbol{\beta}) \propto e^{-\frac{U(\boldsymbol{\beta})}{\tau^{(1)}}}$ and $\pi^{(P)}(\boldsymbol{\beta}) \propto e^{-\frac{U(\boldsymbol{\beta})}{\tau^{(P)}}}$.

Preliminaries

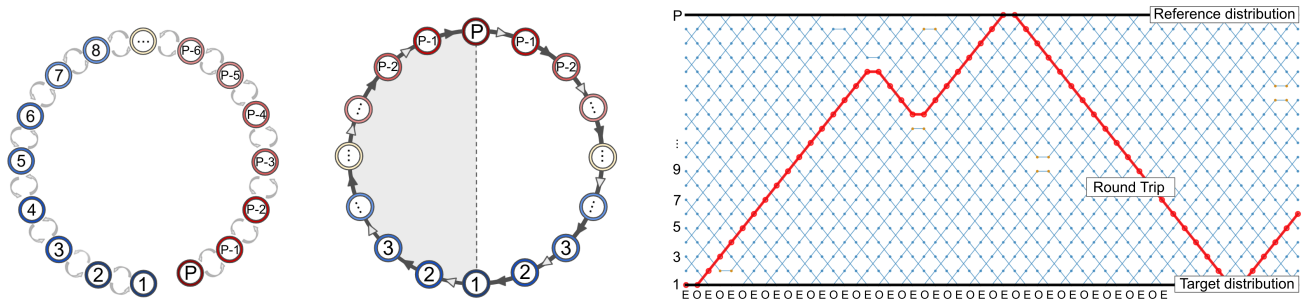
Sufficient explorations require a large $\tau^{(P)}$, which leads to limited accelerations due to a *small overlap* between $\pi^{(1)}$ and $\pi^{(P)}$. To tackle this issue, one can bring in multiple particles with temperatures $(\tau^{(2)}, \cdots, \tau^{(P-1)})$, where $\tau^{(1)} < \tau^{(2)} < \cdots < \tau^{(P)}$, to hollow out "tunnels". To maintain feasibility, numerous schemes are presented to select candidate pairs to attempt the swaps.

APE The all-pairs exchange (APE) attempts to swap arbitrary pair of chains (Brenner et al. 2007; Lingenheil et al. 2009), however, such a method requires a swap time (see definition in section A.5 (appendix)) of $O(P^3)$ and may not be user-friendly in practice.

ADJ In addition to swap arbitrary pairs, one can also swap *adjacent* (ADJ) pairs iteratively from $(1, 2)$, $(2, 3)$, to $(P - 1, P)$ under the Metropolis rule. Despite the convenience, the *sequential nature* requires to wait for exchange information from previous exchanges, which only works well with a small number of chains and has greatly limited its extension to a distributed context.

SEO The stochastic even-odd (SEO) scheme first divides the adjacent pairs $\{(p-1, p) | p = 2, \cdots, P\}$ into E and O, where E and O denote even and odd pairs of forms

[*]These authors contributed equally.



(a) Reversibility (b) Non-reversibility (c) Non-reversible chains

Figure 1: (Non-)Reversibility. (a): a reversible index that takes $O(P^2)$ time to communicate; (b): a linear non-reversible index moves along a periodic orbit.

$(2p-1, 2p)$ and $(2p, 2p+1)$, respectively. Then, SEO randomly picks E or O pairs with an equal chance in each iteration to attempt the swaps. Notably, it can be conducted *simultaneously* without waiting from other chains. The scheme yields a reversible process (see Figure 1(a)), however, the gains in overcoming the sequential obstacle don't offset the $O(P^2)$ *round trip time* and SEO is still not effective enough.

DEO The deterministic even-odd (DEO) scheme instead attempts to swap even (E) pairs at even (E) iterations and odd (O) pairs at odd (O) iterations alternatingly[†] (Okabe et al. 2001). The asymmetric manner was later interpreted as a non-reversible PT (Syed et al. 2021) and an ideal index process follows a periodic orbit, as shown in Figure 1(b). With a large swap rate, Figure 1(c) shows how the scheme yields an almost straight path and a linear round trip time can be expected.

Equi-acceptance The power of PT hinges on maximizing the number of round trips, which is equivalent to minimizing $\sum_{p=1}^{P-1} \frac{1}{1-r_p}$ (Nadler and Hansmann 2007a), where r_p denotes the rejection rate for the chain pair $(p, p+1)$. Moreover, $\sum_{p=1}^{P-1} r_p$ converges to a fixed barrier Λ as $P \to \infty$ (Predescu, Predescu, and Ciobanu 2004; Syed et al. 2021). Applying Lagrange multiplies to the constrained optimization problem leads to $r_1 = r_2 = \cdots = r_{P-1} := r$, where r is the *equi-rejection rate*. In general, a quadratic round trip time is required for ADJ and SEO due to the reversible indexes. By contrast, DEO only yields a *linear round trip* time in terms of P as $P \to \infty$ (Syed et al. 2021).

Optimal Non-reversible Scheme for PT

The linear round trip time is appealing for maximizing the algorithmic potential, however, such an advance only occurs given sufficiently many chains. In *non-asymptotic settings* with limited chains, a pearl of wisdom is to avoid frequent swaps (Dupuis et al. 2012) and to keep the average acceptance rate from 20% to 40% (Kone and Kofke 2005; Lingenheil et al. 2009; Atchadé, Roberts, and Rosenthal 2011). Most importantly, the acceptance rates are severely reduced in big data due to the bias-corrected swaps associated with stochastic energies (Deng et al. 2020), see details in section A.1 (appendix) . As such, maintaining low rejection rates in big data becomes quite challenging and the *issue of quadratic costs* still exists.

Generalized DEO Scheme

Continuing the equi-acceptance settings, we see in Figure 2(a) that the probability for the blue particle to move upward 2 steps to maintain the same momentum after a pair of even and odd iterations is $(1-r)^2$. As such, with a large equi-rejection rate r, the blue particle often makes little progress (Figure 2(b-d)). To handle this issue, the key is to propose small enough rejection rates to track the periodic orbit in Figure 1(b). Instead of pursuing excessive amount of chains, *we resort to a different solution by introducing the generalized even and odd iterations* E_W and O_W, where $W \in \mathbb{N}^+$, $E_W = \{\lfloor \frac{k}{W} \rfloor \bmod 2 = 0 | k = 1, 2, \cdots, \infty\}$ and $O_W = \{\lfloor \frac{k}{W} \rfloor \bmod 2 = 1 | k = 1, 2, \cdots, \infty\}$. Now, we present the generalized DEO scheme with a window size W as follows and refer to it as DEO$_W$: [§]

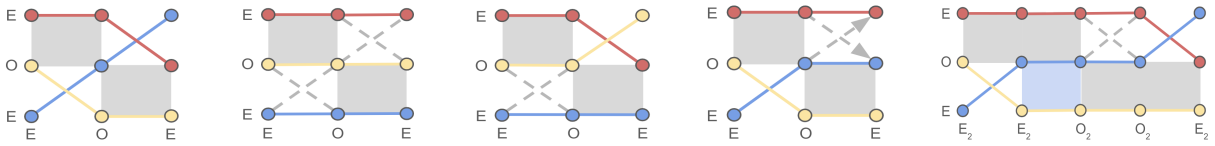
- Attempt to swap E (or O) pairs at E_W (or O_W) iterations.
- Allow *at most one* swap at E_W (or O_W) iterations.

As shown in Figure 2(e), the blue particle has a larger chance of $(1-r^2)^2$ to move upward 2 steps given $W = 2$ instead of $(1-r)^2$ when $W = 1$, although the window number is also halved. Such a trade-off inspires to analyze the round trip based on a window of size W.

How to alleviate the bias Although allowing at most one swap introduces the geometric stopping of swaps and affects the target distribution (see section 2 of (Gerber, Shiu, and Yang 2015)), the bias can be much alleviated empirically by introducing a window-wise correction term. Moreover, it becomes rather mild when the energy estimators have a large variance. Check section C.2 (appendix) for the details. For tasks without high-accuracy demands, we propose to ignore the correction term in practice following Li et al. (2016) to

[†] E (O) shown in iterations means even (odd) iterations and denotes even (odd) pairs for chain indexes.

[§] The generalized DEO with the optimal window size is denoted by DEO$_\star$.



(a) DEO (b) Bad case 1 (c) Bad case 2 (d) Bad case 3 (e) DEO_2

Figure 2: Illustration of DEO and DEO_2. (a): an ideal DEO scheme; (b-d): failed DEO swaps given a large r; (e): how DEO_2 tackles the issue. The x-axis and y-axis denote (generalized) E (or O) iterations and E (or O) pairs, respectively. The dashed line denotes no swap; the gray areas are frozen to refuse swapping odd pairs at even iterations (or vice versa); the blue area freezes swap attempts.

facilitate the round trip analysis and promote more tractable explorations.

Analysis of Round Trip Time

To bring sufficient interactions between the reference distribution $\pi^{(P)}$ and the target distribution $\pi^{(1)}$, we expect to minimize the expected round trip time T (defined in section A.5 (appendix)) to ensure both efficient exploitation and explorations. The non-Markovian nature of the index process makes the analysis challenging. To facilitate the analysis, we treat swap indicators as independent Bernoulli variables following Syed et al. (2021). Combining the Markov property, we estimate the expected round trip time $\mathbb{E}[T]$ as follows:

Lemma 1. *Under the stationary and weak independence assumptions B1 and B2 in section B (appendix) , for P ($P \geq 2$) chains with window size W ($W \geq 1$) and rejection rates $\{r_p\}_{p=1}^{P-1}$, we have*

$$\mathbb{E}[T] = 2WP + \underbrace{2WP \sum_{p=1}^{P-1} \frac{r_p^W}{1 - r_p^W}}_{\textit{Quadratic term in } P}. \tag{3}$$

The proof in section B.1 shows that $\mathbb{E}[T]$ increases as we adopt larger number of chains P and rejection rates $\{r_p\}_{p=1}^{P-1}$. In such a case, the round trip rate $\frac{P}{\mathbb{E}[T]}$ is also maximized by the key renewal theorem. In particular, applying $W = 1$ recovers the vanilla DEO scheme.

Analysis of Optimal Window Size

By Lemma 1, we observe a potential to remove the quadratic term given an appropriate W. Such a fact motivates us to study the optimal W to achieve the best efficiency. Under the equi-acceptance settings, by treating W as a continuous variable and taking the derivative with respect to W, we have

$$\frac{\partial}{\partial W}\mathbb{E}[T] = \frac{2P}{(1-r^W)^2}\Big\{(1-r^W)^2 + (P-1)r^W(1-r^W + W\log r)\Big\}, \tag{4}$$

where r is the equi-rejection rate for adjacent chains. Define $x := r^W \in (0,1)$, where $W = \log_r(x) = \frac{\log x}{\log r}$. The following analysis hinges on the study of the solution of $g(x) = (1-x)^2 + (P-1)x(1-x+\log(x)) = 0$. By analyzing the growth of derivatives and boundary values, we can identify the *uniqueness* of the solution. Then, we proceed to verify that $\frac{1}{P\log P}$ yields an approximation such that $g(\frac{1}{P\log P}) = -\frac{\log(\log P)}{\log P} + O\left(\frac{1}{\log P}\right) \to 0$ as $P \to \infty$. In the end, we have

Theorem 1. *Under Assumptions B1 (Stationarity) and B2 (Weak independence) based on equi-acceptance settings, if $P = 2, 3$, the minimal round trip time is achieved when $W = 1$. If $P \geq 4$, with the optimal window size $W_\star \approx \left\lceil \frac{\log P + \log\log P}{-\log r} \right\rceil$, where $\lceil\cdot\rceil$ is the ceiling function. The round trip time follows $O(\frac{P\log P}{-\log r})$.*

The result yields a remarkable round trip time of $O(P\log P)$ by setting the optimal $W_\star$. By contrast, the vanilla DEO only leads to a longer time of $O(P^2)$ [§]. Denoting by $\text{DEO}_\star$ the generalized DEO with the optimal window size $W_\star$, we summarize the popular swap schemes in Table 1, where $\text{DEO}_\star$ performs the best among all the three criteria. We acknowledge that the assumptions are inevitably strong to simplify the analysis (Syed et al. 2021) due to the intractable index process. Empirically, assumption B1 approximately holds after a sufficient burn-in period; more in-depth discussions on the robustness of assumption B2 have also been evaluated in section 7.2 of Syed et al. (2021).

Discussions on the Optimal Number of Chains

Note that in practice given P parallel chains, a large P leads to a smaller equi-rejection rate r. As such, we can further obtain a crude estimate of the optimal P to minimize the round trip time.

Corollary 1. *Under Assumptions B1-B4 and C1 with equi-acceptance and the optimal window, the optimal chains follow that $P_\star > \min_p \frac{\sigma_p}{3\tau^{(p)}}\log(\frac{\tau^{(P)}}{\tau^{(1)}})$, where σ_p is defined in Eq.(14) (appendix) .*

The assumptions and proof are postponed in section B.3 (appendix) . In mini-batch settings, insufficient chains may lead to few effective swaps for accelerations; by contrast, introducing too many chains may be too costly in terms of the round trip time. This is different from the conclusion in full-batch settings, where Syed et al. (2021) suggested running the vanilla DEO scheme with as many chains as possible to

[§] By Taylor expansion, given a large rejection rate r, $-\log(r) = 1 - r$, which means $\frac{1}{-\log(r)} = O(\frac{r}{1-r})$.

	ROUND TRIP TIME (NON-ASYMPTOTIC)	ROUND TRIP TIME (ASYMPTOTIC)	SWAP TIME
ADJ	$O(P^2)$ (NADLER AND HANSMANN 2007B)	$O(P^2)$ (NADLER AND HANSMANN 2007B)	$O(P)$
SEO	$O(P^2)$ (SYED ET AL. 2021)	$O(P^2)$ (SYED ET AL. 2021)	$O(1)$
DEO	$O(P^2)$ (SYED ET AL. 2021)	$O(P)$ (SYED ET AL. 2021)	$O(1)$
DEO$_\star$	$O(P\log P)$	$O(P)$	$O(1)$

Table 1: Round trip time and swap time for different schemes. Notably, non-asymptotic refers to cases with large rejection rates due to a limited number of chains; asymptotic occurs given sufficiently many chains such that rejection rates are close to 0. The APE scheme requires an expensive swap time of $O(P^3)$ and is not compared.

Algorithm 1: Non-reversible parallel tempering with SGD-based exploration kernels (DEO$_\star$-SGD).

Input Number of chains $P \geq 3$, boundary learning rates $\eta^{(1)}$ and $\eta^{(P)}$, target swap rate $\mathbb{S}$.
Input Optimal window size $W := \left\lceil \frac{\log P + \log\log P}{-\log(1-\mathbb{S})} \right\rceil$.
for $k = 0$ to K **do**
  $\boldsymbol{\beta}_{k+1} \sim \mathcal{T}_\eta(\boldsymbol{\beta}_k)$ following Eq.(6) ▷ Sampling phase
  $\mathcal{P} = \left\{\forall p \in \{1, 2, \cdots, P\} : p \bmod 2 = \lfloor \frac{k}{W} \rfloor \bmod 2\right\}$.
  for $p = 1, 2$ to $P-1$ **do**
    $\mathcal{A}^{(p)} := 1_{\widetilde{U}(\boldsymbol{\beta}_{k+1}^{(p+1)}) + \mathbb{C}_k < \widetilde{U}(\boldsymbol{\beta}_{k+1}^{(p)})}$
    if k mod W=0 **then**
      Open: $\mathcal{G}^{(p)} = 1$. ▷ Open the gate to allow swaps
    end if
    if $p \in \mathcal{P}$ and $\mathcal{G}^{(p)}$ and $\mathcal{A}^{(p)}$ **then**
      Swap: $\boldsymbol{\beta}_{k+1}^{(p)}$ and $\boldsymbol{\beta}_{k+1}^{(p+1)}$. ▷ Communication phase
      Freeze: $\mathcal{G}^{(p)} = 0$. ▷ Close the gate to refuse swaps
    end if
    if $p > 1$ **then**
      Update learning rate following Eq.(11)
    end if
  end for
  Correction: $\mathbb{C}_{k+1} = \mathbb{C}_k + \gamma_k \left(\frac{1}{P-1} \sum_{p=1}^{P-1} \mathcal{A}^{(p)} - \mathbb{S} \right)$.
end for
Output Target models $\{\boldsymbol{\beta}_k^{(1)}\}_{k=1}^K$.

yield a small enough equi-rejection rate r to maintain the non-reversibility.

Cutoff phenomenon On the one hand, when we only afford at most P chains, where $P < P_\star$, a large equi-rejection rate r is inevitable and DEO$_\star$ is preferred over DEO; on the other hand, the rejection rate r goes to 0 when $P \gg P_\star$ and DEO$_\star$ recovers the DEO scheme.

In section B.4 (appendix) , we show $P_\star$ is in the *order of thousands* for the CIFAR100 example, which is hard to achieve due to the limited budget and further motivates us to adopt finite chains with a target swap rate $\mathbb{S}$ to balance between acceleration and accuracy.

User-friendly Approximate Explorations

Despite the asymptotic correctness, stochastic gradient Langevin dynamics (Welling and Teh 2011) (SGLD) only works well given *small enough learning rates* and fails in explorative purposes (Ahn, Korattikara, and Welling 2012). A large learning rate, however, leads to excessive stochastic gradient noise and ends up with a crude approximation. As such, similarly to Izmailov et al. (2018); Zhang et al. (2020), we only adopt SGLD for exploitations.

Efficient explorations not only require a high temperature but also prefer a large learning rate. Such a demand inspires us to consider SGD with a constant learning rate η as the exploration component

$$\begin{aligned} \boldsymbol{\beta}_{k+1} &= \boldsymbol{\beta}_k - \eta \nabla \widetilde{U}(\boldsymbol{\beta}_k) \\ &= \boldsymbol{\beta}_k - \eta \nabla U(\boldsymbol{\beta}_k) - \sqrt{2\eta(\frac{\eta}{2})}\varepsilon(\boldsymbol{\beta}_k), \end{aligned} \tag{5}$$

where $\widetilde{U}(\cdot)$ is the unbiased energy estimate of $U(\cdot)$ and $\varepsilon(\boldsymbol{\beta}_k) \in \mathbb{R}^d$ is the stochastic gradient noise with mean 0. Under mild normality assumptions on ε (Mandt, Hoffman, and Blei 2017; Chen et al. 2020), $\boldsymbol{\beta}_k$ converges approximately to an invariant distribution, where the underlying *temperature linearly depends on the learning rate* η. Motivated by this fact, we propose an approximate transition kernel $\mathcal{T}_\eta$ with P parallel *SGD runs* based on different learning rates

$$\begin{aligned} &\textbf{Exploration:} \begin{cases} \boldsymbol{\beta}_{k+1}^{(P)} = \boldsymbol{\beta}_k^{(P)} - \eta^{(P)} \nabla \widetilde{U}(\boldsymbol{\beta}_k^{(P)}), \\ \cdots \\ \boldsymbol{\beta}_{k+1}^{(2)} = \boldsymbol{\beta}_k^{(2)} - \eta^{(2)} \nabla \widetilde{U}(\boldsymbol{\beta}_k^{(2)}), \end{cases} \\ &\textbf{Exploitation:} \quad \boldsymbol{\beta}_{k+1}^{(1)} = \boldsymbol{\beta}_k^{(1)} - \eta^{(1)} \nabla \widetilde{U}(\boldsymbol{\beta}_k^{(1)}) \ + \overbrace{\Xi_k}^{\text{optional}}, \end{aligned} \tag{6}$$

where $\eta^{(1)} < \eta^{(2)} < \cdots < \eta^{(P)}$, $\Xi_k \sim \mathcal{N}(0, 2\eta^{(1)}\tau^{(1)})$, and $\tau^{(1)}$ is the target temperature.

Since there exists an optimal learning rate for SGD to estimate the desired distribution through Laplace approximation (Mandt, Hoffman, and Blei 2017), the exploitation kernel can be also replaced with SGD based on constant learning rates if the accuracy demand is not high. Regarding the validity of adopting different learning rates for parallel tempering, we leave discussions to section A.2 (appendix) .

Approximation Analysis

Moreover, the stochastic gradient noise exploits the Fisher information (Ahn, Korattikara, and Welling 2012) and yields convergence potential to wide optima with good generalizations (Berthier, Bach, and Gaillard 2020). Despite the implementation convenience, the inclusion of SGDs has made the temperature variable inaccessible, rendering a difficulty in implementing the Metropolis rule Eq.(2). To tackle this issue,

we utilize the randomness in stochastic energies and propose a *deterministic swap condition* for the approximate kernel $\mathcal{T}_\eta$ in Eq.(6)

Deterministic swap condition:

$$\begin{aligned}\text{If } \widetilde{U}(\boldsymbol{\beta}^{(p+1)}) + \mathbb{C} < \widetilde{U}(\boldsymbol{\beta}^{(p)}) \\ (\boldsymbol{\beta}^{(p)}, \boldsymbol{\beta}^{(p+1)}) \to (\boldsymbol{\beta}^{(p+1)}, \boldsymbol{\beta}^{(p)}),\end{aligned} \tag{7}$$

where $p \in \{1, 2, \cdots, P-1\}$, $\mathbb{C} > 0$ is a correction buffer to approximate the Metropolis rule Eq.(2).

Lemma 2. *Assume the energy normality assumption (C1), then for any fixed $\partial U_p := U(\boldsymbol{\beta}^{(p)}) - U(\boldsymbol{\beta}^{(p+1)})$, there exists an optimal $\mathbb{C}_\star \in (0, (\frac{1}{\tau^{(p)}} - \frac{1}{\tau^{(p+1)}})\sigma_p^2]$ that perfectly approximates the random event $\widetilde{S}(\boldsymbol{\beta}^{(p)}, \boldsymbol{\beta}^{(p+1)}) > u$, where σ_p defined in Eq.(14) (appendix) and $u \sim$ Unif $[0, 1]$.*

The proof is postponed in section C.1 (appendix) , which paves the way for the guarantee that a *deterministic swap condition* may replace the Metropolis rule Eq.(2) for approximations. In addition, the normality assumption can be naturally extended to the asymptotic normality assumption (Quiroz et al. 2019; Deng et al. 2021) given large enough batch sizes. Admittedly, the approximation error still exists for different ∂U_p. By the mean-value theorem, there exists a tunable $\mathbb{C}$ to optimize the overall approximation. Further invoking the central limit theorem such that $\varepsilon(\cdot)$ in Eq.(5) approximates a Gaussian distribution, we can expect a reasonable approximation for the SGD-based exploration kernels (Mandt, Hoffman, and Blei 2017).

Theorem 2. *Consider the exact transition kernel $\mathcal{T}$ and the proposed approximate kernel $\mathcal{T}_\eta$, which yield stationary distributions π and π_η, respectively. Under smoothness (C2) and dissipativity assumptions (C3), $\mathcal{T}$ satisfies the geometric ergodicity such that there is a contraction constant $\rho \in [0, 1)$ for any distribution μ:*

$$\|\mu\mathcal{T} - \pi\|_{TV} \leq \rho\|\mu - \pi\|_{TV},$$

where $\|\cdot\|_{TV}$ is the total variation (TV) distance. Moreover, assume that $\varepsilon(\cdot) \sim \mathcal{N}(0, \mathcal{M})$ for some positive definite matrix $\mathcal{M}$ (C4) (Mandt, Hoffman, and Blei 2017), then there is a uniform upper bound of the one step error between $\mathcal{T}$ and $\mathcal{T}_\eta$ such that

$$\|\mu\mathcal{T} - \mu\mathcal{T}_\eta\|_{TV} \leq \Delta_{\max}, \forall\mu,$$

where $\Delta_{\max} \geq 0$ is a constant. Eventually, the TV distance between π and π_η is bounded by

$$\|\pi - \pi_\eta\|_{TV} \leq \frac{\Delta_{\max}}{1-\rho}.$$

The proof is postponed to section C.2 (appendix) . The SGD-based exploration kernels *no longer require to fine-tune the temperatures* directly and naturally inherits the empirical successes of SGD in large-scale deep learning tasks. The inaccessible Metropolis rule Eq.(2) is approximated via the *deterministic swap condition* Eq.(7) and leads to robust approximations by *tuning* $\boldsymbol{\eta} = (\eta^{(1)}, \cdots, \eta^{(P)})$ *and* $\mathbb{C}$.

In addition, our proposed algorithm for posterior approximation also relates to non-convex optimization. For detailed discussions, we refer interested readers to section A.4 (appendix) .

Equi-acceptance Parallel Tempering

Stochastic approximation (SA) is a standard method to achieve equi-acceptance (Atchadé, Roberts, and Rosenthal 2011), however, implementing this idea with fixed $\eta^{(1)}$ and $\eta^{(P)}$ is rather non-trivial. Motivated by the linear relation between learning rate and temperature, we propose to adaptively *optimize the learning rates* to achieve equi-acceptance in a user-friendly manner. Further by the geometric temperature spacing commonly adopted by practitioners (Kofke 2002; Earl and Deem 2005; Syed et al. 2021), we adopt the following scheme

$$\partial \log(\upsilon_t^{(p)}) = h^{(p)}(\upsilon_t^{(p)}), \tag{8}$$

where $p \in \{1, 2, \cdots, P-1\}$, $\upsilon_t^{(p)} = \eta_t^{(p+1)} - \eta_t^{(p)}$, $h^{(p)}(\upsilon_t^{(p)}) := \int H^{(p)}(\upsilon_k^{(p)}, \boldsymbol{\beta})\pi^{(p,p+1)}(d\boldsymbol{\beta})$ is the mean-field function, $\pi^{(p,p+1)}$ is the joint invariant distribution for the p-th and $p+1$-th processes. In particular, $H^{(p)}(\upsilon_k^{(p)}, \boldsymbol{\beta}) = 1_{\widetilde{U}(\boldsymbol{\beta}^{(p+1)}) + \mathbb{C} < \widetilde{U}(\boldsymbol{\beta}^{(p)})} - \mathbb{S}$ is the random-field function to approximate $h^{(p)}(\upsilon_k^{(p)})$ with limited perturbations, $\upsilon_k^{(p)}$ [†] implicitly affects the distribution of the indicator function, and $\mathbb{S}$ is the target swap rate. Now consider stochastic approximation of Eq.(8), we have

$$\log(\upsilon_{k+1}^{(p)}) = \log(\upsilon_k^{(p)}) + \gamma_k H^{(p)}(\upsilon_k^{(p)}, \boldsymbol{\beta}_k), \tag{9}$$

where γ_k is the step size. Reformulating Eq.(9), we have

$$\upsilon_{k+1}^{(p)} = \max(0, \upsilon_k^{(p)})e^{\gamma_k H(\upsilon_k^{(p)})},$$

where the $\max$ operator is conducted explicitly to ensure the sequence of learning rates is non-decreasing. This means given fixed boundary learning rates (temperatures) $\eta_k^{(p-1)}$ and $\eta_k^{(p+1)}$, applying $\eta^{(p)} = \eta^{(p-1)} + \upsilon^{(p)}$ and $\eta^{(p)} = \eta^{(p+1)} - \upsilon^{(p+1)}$ for $p \in \{2, 3, \cdots, P-1\}$ lead to

$$\begin{aligned}\eta_{k+1}^{(p)} &= \underbrace{\eta_k^{(p-1)} + \max(0, \upsilon_k^{(p)})e^{\gamma_k H(\upsilon_k^{(p)})}}_{\text{forward sequence}} \\ &= \underbrace{\eta_k^{(p+1)} - \max(0, \upsilon_k^{(p+1)})e^{\gamma_k H(\upsilon_k^{(p+1)})}}_{\text{backward sequence}}.\end{aligned} \tag{10}$$

Adaptive learning rates (temperatures) Now given a fixed $\eta^{(1)}$, the sequence $\eta^{(2)}, \eta^{(3)}, \cdots, \eta^{(P)}$ can be approximated iteratively via the forward sequence of (10); conversely, given a fixed $\eta^{(P)}$, the backward sequence $\eta^{(P-1)}, \eta^{(P-2)}, \cdots, \eta^{(1)}$ can be decided reversely as well. Combining the forward and backward sequences, $\eta_{k+1}^{(p)}$ can be approximated via

$$\begin{aligned}\eta_{k+1}^{(p)} := \frac{\eta_k^{(p-1)} + \eta_k^{(p+1)}}{2} + \\ \frac{\max(0, \upsilon_k^{(p)})e^{\gamma_k H(\upsilon_k^{(p)})} - \max(0, \upsilon_k^{(p+1)})e^{\gamma_k H(\upsilon_k^{(p+1)})}}{2},\end{aligned} \tag{11}$$

[†] $\upsilon_t^{(p)}$ denotes a continuous-time diffusion at time t and $\upsilon_k^{(p)}$ is a discrete approximation at iteration k.

which resembles the *binary search* in the SA framework. In particular, the first term is the middle point given boundary learning rates and the second term continues to penalize learning rates that violates the equi-acceptance between pairs $(p-1,p)$ and $(p,p+1)$ until an equilibrium is achieved. This is the first attempt to achieve equi-acceptance given two fixed boundary values to our best knowledge. By contrast, Syed et al. (2021) proposed to estimate the barrier Λ to determine the temperatures and it easily fails in big data given a finite number of chains and bias-corrected swaps.

Adaptive correction buffers In addition, equi-acceptance does not guarantee a convergence to the desired acceptance rate $\mathbb{S}$. To avoid this issue, we propose to adaptively optimize $\mathbb{C}$ as follows

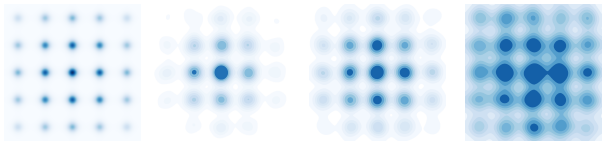
$$\mathbb{C}_{k+1} = \mathbb{C}_k + \gamma_k \left(\frac{1}{P-1} \sum_{p=1}^{P-1} 1_{\widetilde{U}(\boldsymbol{\beta}_{k+1}^{(p+1)}) + \mathbb{C}_k - \widetilde{U}(\boldsymbol{\beta}_{k+1}^{(p)}) < 0} - \mathbb{S} \right). \tag{12}$$

As $k \to \infty$, the threshold and the adaptive learning rates converge to the desired fixed points. Note that setting a uniform $\mathbb{C}$ greatly simplifies the algorithm.Now we refer to the approximate non-reversible parallel tempering algorithm with the DEO$_\star$ scheme and SGD-based exploration kernels as DEO$_\star$-SGD and formally formulate our algorithm in Algorithm 1. Extensions of SGD with a preconditioner (Li et al. 2016) or momentum (Chen, Fox, and Guestrin 2014) to further improve the approximation and efficiency are both straightforward (Mandt, Hoffman, and Blei 2017) and are denoted as DEO$_\star$-pSGD and DEO$_\star$-mSGD, respectively.

Experiments

Simulations of Multi-Modal Distributions

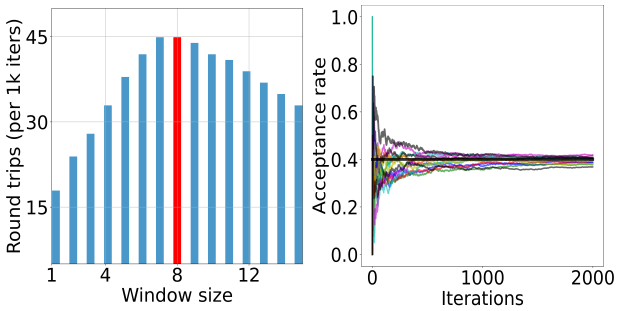
We first simulate the proposed algorithm on a distribution $\pi(\boldsymbol{\beta}) \propto \exp(-U(\boldsymbol{\beta}))$, where $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $U(\boldsymbol{\beta}) = 0.2(\beta_1^2 + \beta_2^2) - 2(\cos(2\pi\beta_1) + \cos(2\pi\beta_2))$. The heat map is shown in Figure 3(a) with 25 modes of different volumes. To mimic big data scenarios, we can only access stochastic gradient $\nabla\widetilde{U}(\boldsymbol{\beta}) = \nabla U(\boldsymbol{\beta}) + 2\mathcal{N}(0, \boldsymbol{I}_{2\times 2})$ and stochastic energy $\widetilde{U}(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) + 2\mathcal{N}(0, I)$.



(a) Truth (b) $\mathbb{S} = 0.2$ (c) $\mathbb{S} = 0.4$ (d) $\mathbb{S} = 0.6$

Figure 3: Study of different target swap rate $\mathbb{S}$ via DEO$_\star$-SGD, where SGLD is the exploitation kernel.

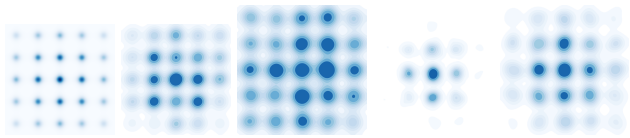
We first run DEO$_\star$-SGD$\times$P16 based on 16 chains and 20,000 iterations. We fix the lowest learning rate 0.003 and the highest learning 0.6 and propose to tune the target swap rate $\mathbb{S}$ for the acceleration-accuracy trade-off. Figure 3 shows that fixing $\mathbb{S} = 0.2$ is too conservative and underestimates the uncertainty on the corners; $\mathbb{S} = 0.6$ results in too many radical swaps and eventually leads to crude estimations; by contrast, $\mathbb{S} = 0.4$ yields the best posterior approximation among the five choices.



(a) Round trips (b) Accept rates

Figure 4: Study of window sizes and acceptance rates.

Next, we select $\mathbb{S} = 0.4$ and study the round trips. We observe in Figure 4(a) that the vanilla DEO only yields 18 round trips every 1,000 iterations; by contrast, slightly increasing W tends to improve the efficiency significantly and the optimal 45 round trips are achieved at $W = 8$, which *matches our theory*. In Figure 4(b), the geometrically initialized learning rates lead to unbalanced acceptance rates in the early phase and some adjacent chains have few swaps, but as the optimization proceeds, the learning rates gradually converge.



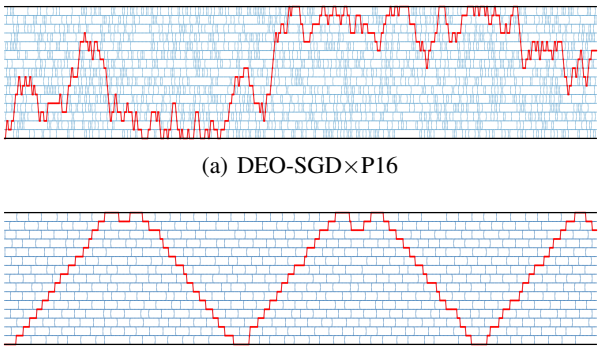
(a) Truth (b) SGLD (c) cycSGLD (d) DEO-SGD (e) DEO$_\star$-SGD

Figure 5: Simulations of the multi-modal distribution through different sampling algorithms. All the algorithms are run based on 16 chains $\times P16$

We compare the proposed algorithm with parallel SGLD based on 20,000 iterations and 16 chains (SGLD$\times$P16); we fix the learning rate 0.003 and a temperature 1. We also run cycSGLD$\times$T16, which denotes a single chain based on 16 times of budget and cosine learning rates (Zhang et al. 2020) of 100 cycles. We see in Figure 5(b) that SGLD$\times$P16 has good explorations but fails to approximate the posterior. Figure 5(c) shows that cycSGLD$\times$T16 explores most of the modes but overestimates some areas occasionally. Figure 5(d) demonstrates the DEO-SGD with 16 chains (DEO-SGD$\times$P16) estimates the uncertainty of the centering 9 modes well but fails to deal with the rest of the modes. As to DEO$_\star$-SGD$\times$P16, the approximation is rather accurate, as shown in Figure 5(e). We also present the index process for both schemes in Figure 6. The vanilla DEO scheme results in volatile paths and a particle takes quite a long time to complete a round trip; by contrast, DEO$_\star$ only conducts at most one cheap swap in a window and yields much more deterministic paths.

MODEL	ResNet20		ResNet32		ResNet56	
	NLL	ACC (%)	NLL	ACC (%)	NLL	ACC (%)
cycSGHMC×T10	8198±59	76.26±0.18	7401±28	78.54±0.15	6460±21	81.78±0.08
cycSWAG×T10	8164±38	76.13±0.21	7389±32	78.62±0.13	6486±29	81.60±0.14
mSGD×P10	7902±64	76.59±0.11	7204±29	79.02±0.09	6553±15	81.49±0.09
DEO-mSGD×P10	7964±23	76.84±0.12	7152±41	79.34±0.15	6534±26	81.72±0.12
DEO$_\star$-mSGD×P10	**7741±67**	**77.37±0.16**	**7019±35**	**79.54±0.12**	**6439±32**	**82.02±0.15**

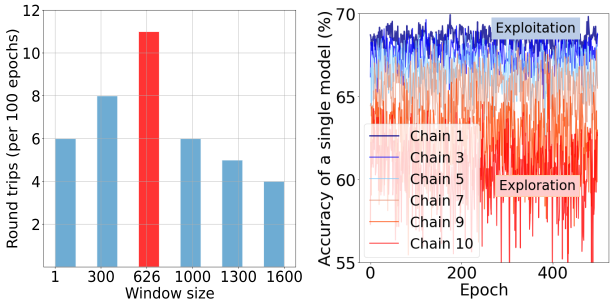
Table 2: Posterior approximation and optimization on CIFAR100 via $10\times$ budget.



(a) DEO-SGD×P16

(b) DEO$_\star$-SGD×P16

Figure 6: Dynamics of the index process. The red path denotes the round trip path for a particle.



(a) Round trips (b) Accuracies

Figure 7: Study of window sizes and accuracies on ResNet20.

Posterior Approximation for Image Data

Next, we conduct experiments on computer vision tasks. We choose ResNet20, ResNet32, and ResNet56 and train the models on CIFAR100. We report negative log likelihood (NLL) and test accuracy (ACC). For each model, we first pre-train 10 fixed models via 300 epochs and then run algorithms based on momentum SGD (mSGD) for 500 epochs with 10 parallel chains and denote it by DEO$_\star$-mSGD×P10. We fix the lowest and highest learning rates as 0.005 and 0.02, respectively. For a fair comparison, we also include the baseline DEO-mSGD×P10 with the same setup except that the window size is 1; the standard ensemble mSGD×P10 is also included with a learning rate of 0.005. In addition, we include two baselines based on a single long chain, i.e. we run stochastic gradient Hamiltonian Monte Carlo 5000 epochs with cyclical learning rates and 50 cycles (Zhang et al. 2020) and refer to it as cycSGHMC×T10; we run SWAG×T10 (Maddox et al. 2019) under similar setups.

In particular for DEO$_\star$-mSGD×P10, we tune the target swap rate $\mathbb{S}$ and find an optimum at $\mathbb{S} = 0.005$. We compare our proposed algorithm with the four baselines and observe in Table 2 that mSGD×P10 can easily obtain competitive results simply through model ensemble, which outperforms cycSGHMC×T10 and cycSWAG×T10 on ResNet20 and ResNet32 models and perform the worst among the five methods on ResNet56; DEO-mSGD×P10 itself is already a pretty powerful algorithm, however, DEO$_\star$-mSGD×P10 consistently outperforms the vanilla alternative.

To analyze why the proposed scheme performs well, we study the round trips in Figure 7(a) and find that the theoretical optimal window obtains around 11 round trips every 100 epochs, which is almost 2 times as much as DEO. In Figure 7(b), we observe that the smallest learning rate obtains the highest accuracy (blue) for exploitations, while the largest learning rate yields decent explorations (red).

Appendix We refer readers to the full version at https://arxiv.org/abs/2211.10837.

Code The code is released to https://github.com/WayneDW/Non-reversible-Parallel-Tempering-for-Deep-Posterior-Approximation for reproduction.

Conclusion

In this paper, we show how to adapt the multiple-chain parallel tempering algorithm to big data problems. Given a limited budget of parallel chains in big data, we show the standard non-reversible DEO scheme leads to an expensive quadratic communication cost with respect to the number of chains. To tackle that issue, we propose a generalized DEO scheme to achieve larger swap rates (window-wise) with mild costs. By sacrificing a mild accuracy in big data, we prove the existence of an optimal window size to encourage *deterministic paths* and obtain in a significant *acceleration of* $O(\frac{P}{\log P})$ *times*. For a user-friendly purpose, we also propose a deterministic swap condition to interact with SGD-based exploration kernels. A crude bias analysis is provided to facilitate the understanding of the extensions.

Acknowledgments

Lin and Deng would like to acknowledge the support from National Science Foundation (DMS-2053746, DMS-1555072, and DMS-1736364), Brookhaven National Laboratory Subcontract (382247), and U.S. Department of Energy Office of Science Advanced Scientific Comput-ing Research (DE-SC0021142). Liang's research is supported in part by the grant DMS-2015498 from National Science Foundation and the grants R01- GM117597 and R01-GM126089 from National Institutes of Health.

References

Ahn, S.; Korattikara, A.; and Welling, M. 2012. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *ICML*.

Atchadé, Y. F.; Roberts, G. O.; and Rosenthal, J. S. 2011. Towards Optimal Scaling of Metropolis-coupled Markov Chain Monte Carlo. *Statistics and Computing*.

Berthier, R.; Bach, F.; and Gaillard, P. 2020. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model. In *NeurIPS*.

Brenner, P.; Sweet, C. R.; VonHandorf, D.; and Izaguirre, J. A. 2007. Accelerating the Replica Exchange Method through an Efficient All-pairs Exchange. *The Journal of Chemical Physics*, 126: 074103.

Chen, T.; Fox, E. B.; and Guestrin, C. 2014. Stochastic Gradient Hamiltonian Monte Carlo. In *ICML*.

Chen, X.; Lee, J. D.; Tong, X. T.; and Zhang, Y. 2020. Statistical Inference for Model Parameters in Stochastic Gradient Descent. *Annals of Statistics*, 48(1): 251–273.

Dalalyan, A. S. 2017. Theoretical Guarantees for Approximate Sampling from Smooth and Log-concave Densities. *JRSS-B*.

Deng, W.; Feng, Q.; Gao, L.; Liang, F.; and Lin, G. 2020. Non-Convex Learning via Replica Exchange Stochastic Gradient MCMC. In *ICML*.

Deng, W.; Feng, Q.; Karagiannis, G.; Lin, G.; and Liang, F. 2021. Accelerating Convergence of Replica Exchange Stochastic Gradient MCMC via Variance Reduction. In *ICLR*.

Deng, W.; Liang, S.; Hao, B.; Lin, G.; and Liang, F. 2022. Interacting Contour Stochastic Gradient Langevin Dynamics. In *ICLR*.

Deng, W.; Lin, G.; and Liang, F. 2022. An Adaptively Weighted Stochastic Gradient MCMC Algorithm for Monte Carlo Simulation and Global Optimization. *Statistics and Computing*, 32–58.

Dupuis, P.; Liu, Y.; Plattner, N.; and Doll, J. D. 2012. On the Infinite Swapping Limit for Parallel Tempering. *SIAM J. Multiscale Modeling & Simulation*, 10.

Earl, D. J.; and Deem, M. W. 2005. Parallel Tempering: Theory, Applications, and New Perspectives. *Phys. Chem. Chem. Phys.*, 7: 3910–3916.

Gerber, H. U.; Shiu, E. S. W.; and Yang, H. 2015. Geometric Stopping of a Random Walk and Its Applications to Valuing Equity-linked Death Benefits. *Insurance: Mathematics and Economics*, 64: 313–325.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. In *UAI*.

Kofke, D. A. 2002. On the Acceptance Probability of Replica-Exchange Monte Carlo Trials. *The Journal of Chemical Physics*, 117.

Kone, A.; and Kofke, D. A. 2005. Selection of Temperature Intervals for Parallel-tempering Simulations. *The Journal of Chemical Physics*, 122: 206101.

Li, C.; Chen, C.; Carlson, D.; and Carin, L. 2016. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. In *AAAI*, 1788–1794.

Lingenheil, M.; Denschlag, R.; Mathias, G.; and Tavan, P. 2009. Efficiency of Exchange Schemes in Replica Exchange. *Chemical Physics Letters*, 478: 80–84.

Maddox, W.; Garipov, T.; Izmailov, P.; Vetrov, D.; and Wilson, A. G. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *NeurIPS*.

Mandt, S.; Hoffman, M. D.; and Blei, D. M. 2017. Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*.

Nadler, W.; and Hansmann, U. H. E. 2007a. Dynamics and Optimal Number of Replicas in Parallel Tempering Simulations. *Phys. Rev. E*, 76: 065701.

Nadler, W.; and Hansmann, U. H. E. 2007b. Generalized Ensemble and Tempering Simulations: A Unified View. *Phys. Rev. E*, 75: 026109.

Okabe, T.; Kawata, M.; Okamoto, Y.; and Mikami, M. 2001. Replica Exchange Monte Carlo Method for the Isobaric–isothermal Ensemble. *Chemical Physics Letters*, 335: 435–439.

Predescu, C.; Predescu, M.; and Ciobanu, C. V. 2004. The Incomplete Beta Function Law for Parallel Pempering Sampling of Classical Canonical Systems. *Chemical Physics Letters*, 120: 4119–4128.

Quiroz, M.; Kohn, R.; Villani, M.; and Tran, M.-N. 2019. Speeding Up MCMC by Efficient Data Subsampling. *Journal of the American Statistical Association*.

Raginsky, M.; Rakhlin, A.; and Telgarsky, M. 2017. Nonconvex Learning via Stochastic Gradient Langevin Dynamics: a Nonasymptotic Analysis. In *COLT*.

Syed, S.; Bouchard-Côté, A.; Deligiannidis, G.; and Doucet, A. 2021. Non-Reversible Parallel Tempering: a Scalable Highly Parallel MCMC scheme. *JRSS-B*.

Welling, M.; and Teh, Y. W. 2011. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *ICML*, 681–688.

Zhang, R.; Li, C.; Zhang, J.; Chen, C.; and Wilson, A. G. 2020. Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning. In *ICLR*.