Appendix of “Replica-Exchange Nosé-Hoover Dynamics for Bayesian Learning on Large Datasets”

Anonymous Author(s)

A Replica Exchange

Figure 1 illustrates the runtime trajectories of five replicas, in which two subroutines are executed in an alternating scheme. In the following subsections, theoretical bases will be established.

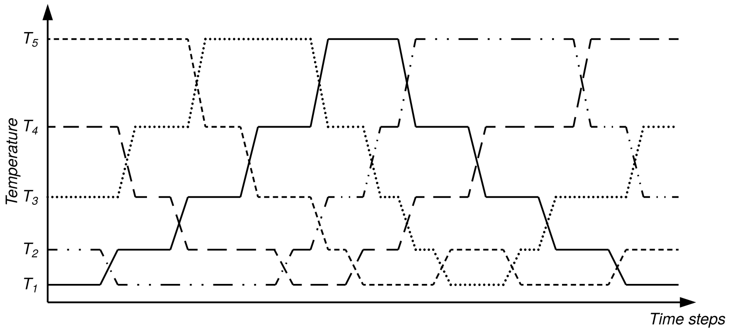

Figure 1: A schematic illustration of the replica-exchange protocol. Lines describe 5 trajectories of dynamics of replicas at different temperatures: horizontal segments represent parallel evolution while intersections is replica exchange.

B Hermite polynomials and the derivatives

Table 1 lists the first 3 logistic derivatives of odd orders. For higher orders, a recursive routine [4] is developed for fast computation.

Table 1: An example of Hermite polynomials and the derivatives of $g(z) = 1/[1 + e^{-z}]$.

order	$H_n\left[u = \sigma^2/4\lambda\right]$	$q^{(2n)}_{\mathcal{L}}(z)$ in terms of g
$n = 0$	1	$g - g^2$
$n = 1$	$2u$	$g - 7g^2 + 12g^3 - 6g^4$
$n = 2$	$4u^2 - 2$	$g - 31g^2 + 180g^3 - 390g^4 + 360g^5 - 120g^6$

C Empirical results of Gaussianity test on stochastic gradient when training ResNet

Figure 2 illustrates the trajectories of percentage of Gaussian elements within the stochastic gradients during training ResNet-20; each curve represents a different block of ResNet-20. It is clear that the percentage of Gaussian elements within each blocks are higher than 90%, which indicates the Gaussianity assumption for ResNet are appropriate hypothesis.

D Effective potentials of replicas at different temperatures

Figure 3 shows the effective potentials of replicas at different temperatures; the corresponding $\pi_j(\theta_j)$ are then illustrated. It becomes clear that when climbing the increasing ladder of temperatures $\{T_j\}$, $\pi_j(\theta_j)$ moves gradually towards a flat histogram.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

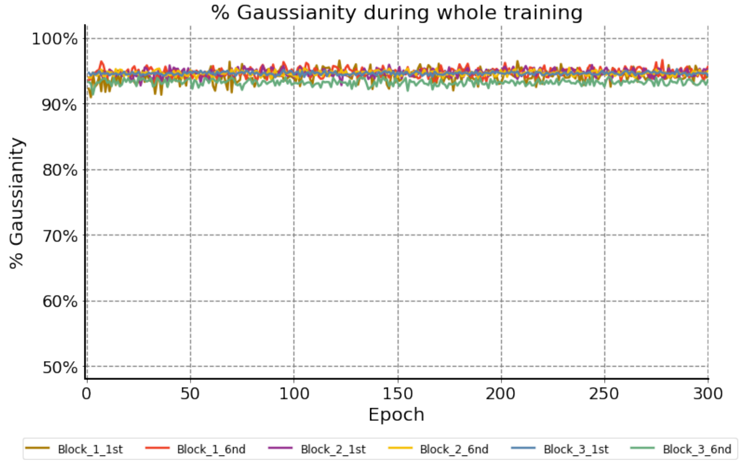

Figure 2: Gaussianity of each layer in ResNet-20 during training epochs.

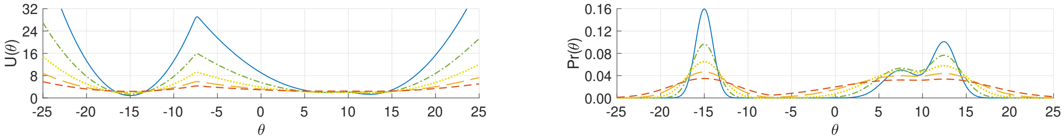

Figure 3: (*colored*) The *left* plot shows the effective potentials for 5 replicas at different temperatures. As temperature rises, the energy barrier at -7 reduces, which facilitates the passage. The *right* gives the marginal distributions, moving towards flattened histograms during tempering. The blue curves is the real potential (*left*) and thus the true posterior (*right*) at $T = 1$.

E Proof of Theorem 1

Proof. We prove the existence of the invariant distributions. The uniqueness follows as a consequence of the assumption on ergodicity.

The Nosé-Hoover dynamics in (3) defines a system of stochastic differential equations, which governs the time evolution of state in a probabilistic way from a microscopic perspective. On the other hand, consider the entire ensemble, *i.e.* the collection of all possible states, its evolution can be characterized statistically from a macroscopic point of view through the time evolution of state distribution $\pi_j(\Gamma_j, t)$. The Fokker-Planck equation [6] translates the stochastic dynamics of state into the differential equation

$$\begin{aligned} \dot{\pi}_j(\Gamma_j, t) = & -\partial_{\theta_j}^\top \left[p_j \pi_j\right] - \partial_{p_j}^\top \left[f(\theta_j)\pi_j\right] + \partial_{p_j}^\top \left[\xi_j p_j \pi_j\right] \\ & - \partial_{\xi_j} \left[(p_j^\top p_j - T_j d)\pi_j\right] + \partial_{p_j}^\top \left[B \partial_{p_j} \pi_j\right], \end{aligned} \tag{1}$$

which can be solved deterministically or even analytically; the invariant distributions can be indicated by $\dot{\pi}_j(\Gamma_j, t) = 0$.

We presume that the invariant distribution of ξ is separable from that of θ_j and p_j so that $\pi_j(\Gamma_j) = \pi_j(\xi_j)\pi_j(\theta_j, p_j)$. For the marginal distribution $\pi_j(\theta_j, p_j)$, we consider the typical Boltzmann distribution for the Hamiltonian system (θ_j, p_j) with the potential U and an additive quadratic kinetic energy $p_j^\top p_j/2$ (we presume all replicas have unit masses) as is defined for our system:

$$\pi_j(\theta_j, p_j) \propto \exp\Big[- \big[U(\theta_j) + p_j^\top p_j/2\big] \Big/ T_j \Big]. \tag{2}$$

When solving $\dot{\pi}_j(\Gamma_j, t) = 0$, the Boltzmann $\pi_j(\theta_j, p_j)$ in (2) results in the Hamiltonian dynamics [5]; the first and second terms in (1) therefore cancel with each other. The resulting equation *w.r.t.* $\pi_j(\xi_j)$ is simplified as

$$\frac{1}{\pi_j(\xi_j)} \frac{\mathrm{d}\pi_j(\xi_j)}{\mathrm{d}\xi_j} = -\frac{1}{T_j}\left[\xi_j - \frac{B}{T_j}\right],$$

which gives the unique solution up to a normalizing constant

$$\pi_j(\xi) \propto \exp\left[- \frac{(\xi_j - {}^{B}\!/\!{}_{T_j})^2}{2T_j} \right]. \tag{3}$$

Combining two marginal distributions in (2) and (3), the joint distribution of state is obtained as in (4), which is invariant by construction. ☐

F Well-Tempered Ensemble for replica reduction

In this section, we discuss an "optional" device, the *Well-tempered Ensemble* (WTE) [1], for RENHD. WTE is important, albeit not indispensable, for its use of enhancing the memory efficiency of RENHD by reducing the number of replicas for real-world applications, especially for deep neural networks.

In learning very deep neural networks, it might be the case that the parameters grows to hundreds of millions, or even billions. As the efficiency of RENHD relies on the chance of successful exchanges, and the latter is a function of (potential) energy differences: in our case, it resembles the logistic function $g(\Delta E_{jk})$. For a pair of replicas (j,k), a greater overlap of the energy distributions $\pi_j(E)$ and $\pi_k(E)$ will lead to a better chance on the exchange between θ_j and θ_k. However, observations reveal that the overlap will decrease in the rate of ${}^{1}\!/\!{}_{\sqrt{d}}$ when the system size d (*i.e.* the dimension for $\theta \in \mathbb{R}^d$) increases [3]. Therefore, to retain a constant acceptance probability, the number of replicas needs to increase in $\sqrt{d}$. For very large systems, the amount of replicas might be prohibitively large.

WTE manages to reduce the number of replicas by enlarging the energy overlap of replicas. It constructs and then maintains for each replica j a time-dependent biasing potential $A_j^\gamma(E,t)$ with $\gamma > 1$ denoting the *tempering factor*, which is a predefined constant defining the increase of energy overlaps by WTE. Figure 4 illustrates the effect of deploying WTE on a demo model with Gaussian energy distributions; the overlap of energy distributions (of adjacent replicas) is substantially enlarged.

The time evolution of the biasing potential A_j^γ in WTE is defined by

$$\frac{\mathrm{d}A_j^\gamma(E,t)}{\mathrm{d}t} = h \exp\left[- \frac{A_j^\gamma(E,t)}{(\gamma-1)T_j} \right] \cdot \delta\big[E - U(\theta_j(t))\big], \tag{4}$$

where $\theta_j(t)$ indicates the trajectory of θ_j at time t, h is a constant determining the learning rate of A_j^γ, and $\delta[\cdot]$ denotes the Dirac delta function. As γ is a constant, we hereafter omit it for simplicity. Intuitively, the dynamics (4) gradually builts up a $1d$ landscape A_j for replicas j, with the coordinates being the energy E, at a rate of h. The way it determines where to make such incremental changes is by calculating the potential U at the current configuration $\theta_j(t)$.

It has been shown that $A_j(E,t)$ converges asymptotically [2]. With $A_j(E) := A_j(E, t \to \infty)$, the augmented potential can be defined as $V_j(\theta_j) := U(\theta_j) + A_j(U(\theta_j))$ and the tempered energy distribution reads

$$\begin{aligned} \tilde{\pi}_j^A(E) &\propto \int \delta\big[E - U(\theta_j)\big] e^{-V(\theta_j)/T_j}\, \mathrm{d}\theta_j \\ &= \left(\int \delta\big[E - U(\theta_j)\big]\, \mathrm{d}\theta_j \right) e^{-\left[E + A_j(E)\right]/T_j}. \end{aligned} \tag{5}$$

Theorem 1 ([1])**.** *The energy distribution* (5) *of the WTE-augmented replica j with converged A_j satisfies*

$$\tilde{\pi}_j^A(E) \propto \big[\pi_j(E)\big]^{1/\gamma}, \tag{6}$$

indicating that the fluctuation **var**$[E]$ w.r.t. $\tilde{\pi}_j^A$ *is effectively amplified by a factor* γ.

Proof. We firstly recall equation (5). We define the integral in the last equity as the temperature-independent density of states, formulated as

$$N_j(E) := \int \delta[E - U(\theta_j)]\, \mathrm{d}\theta_j \tag{7}$$

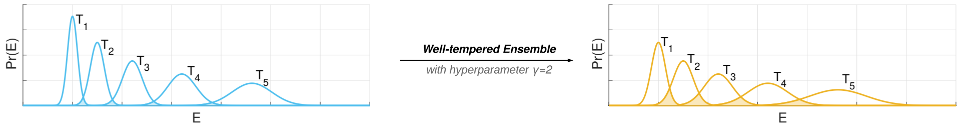

Figure 4: Effect of deploying WTE on a set of 5 replicas at different temperatures. On the *left* depicts the real histograms of replicas' energy distributions while the *right* shows their tempered counterparts. With WTE enabled, the energy overlap (*shaded*) of adjacent replicas is greatly enlarged, leading to a better chance for successful exchange and thus a higher efficiency.

Algorithm 1 Replica-exchange Nosé-Hoover dynamics with Well-tempered Ensemble

```
function NHDYNAMICS($\{\theta_j\}, \{A_j[\cdot]\}, \{T_j\}$, model, $\mathscr{D}, |\mathcal{S}|_{\text{nhd}}, N, \epsilon, c, \gamma, h, \Delta$)
                              ▷ NHD length $N$; $\epsilon, c$ in (20); $\gamma, h$ in (4); $\Delta$ for quantizing $A_j$
    for all $\{j\}$ do                                              ▷ all $j$ running in parallel
        $v_j \sim \mathcal{N}(0, T_j\epsilon)$ and $s_j \leftarrow c/T_j$                      ▷ resetting auxiliary variables
        for $n$ = RANGE($1, N$) do
            $\mathcal{S} \leftarrow$ NEXTBATCH( $\mathscr{D}, |\mathcal{S}|_{\text{nhd}}$ )                 ▷ fetching new mini-batch
            $E_j \leftarrow$ model.FORWARD( $\theta_j, \mathcal{S}$ )                        ▷ $E_j := U(\theta_j)$
            $\tilde{f}_j \leftarrow$ model.BACKWARD( $\theta_j, \mathcal{S}$ )             ▷ evaluating mini-batch gradient
            $i \leftarrow$ QUANTIZE($E_j$)                        ▷ indexing $A_j[i]$ for quantized $E_j$
            $\mathrm{d}A_j \leftarrow \big[A_j[i+1] - A_j[i]\big]/\Delta$                  ▷ approximating $\mathrm{d}A_j(E)/\mathrm{d}E$
            $\mathrm{d}V_j \leftarrow [1 + \mathrm{d}A_j]\tilde{f}_j$                            ▷ $V_j := U(\theta_j) + A_j(U(\theta_j))$
            $v_j \leftarrow v_j + \mathrm{d}V_j\epsilon - s_j v_j + \mathcal{N}(0, 2c\epsilon)$       ▷ additional Gaussian noise added
            $\theta_j \leftarrow \theta_j + v_j$ and $s_j \leftarrow s_j + \big[v_j^\top v_j/d - T_j\epsilon\big]$        ▷ simulating NHD in (3)
            $A_j[i] \leftarrow A_j[i] + h\exp\big[-A_j[i]/(\gamma-1)T_j\big]$               ▷ updating $A_j[\cdot]$ cf. (11)
    return $\{\theta_j\}, \{A_j[\cdot]\}$

MAIN:
$\{\theta_j\} \leftarrow$ RANDN() and $\{A_j[\cdot]\} \leftarrow$ ZEROS()                               ▷ initialization
args $\leftarrow \big(|\mathcal{S}|_{\text{nhd}}, N, \epsilon, c, \gamma, h, \Delta\big)$                         ▷ packing arguments
loop
    $\{\theta_j\}, \{A_j[\cdot]\} \leftarrow$ NHDYNAMICS($\{\theta_j\}, \{A_j[\cdot]\}, \{T_j\}$, model, $\mathscr{D}$, args)
    $\{(j,k)\} \leftarrow$ RAND()                                  ▷ sampling a set of replicas to swap
    for all $\{(j,k)\}$ do
        EXCHANGE($\theta_j, \theta_k$, model, $\mathscr{D}, |\mathcal{S}|_{\text{re}}, \sigma_*^2, \lambda, \tilde{q}_{\mathscr{C}}$)                 ▷ recall Algorithm 1
        if $\theta_j$ and $\theta_k$ exchanged then swap $A_j[\cdot]$ and $A_k[\cdot]$
    samples $\leftarrow \big[$samples, $\theta_0\big]$                       ▷ reweighting needed using (9) or (10)
```

such that the tempered energy distribution is re-written as

$$\tilde{\pi}_j^A(E) \propto N_j(E)\, e^{-\left[E+A_j(E)\right]/T_j}.$$

As stated by [1], the equilibrium of biasing potential $A_j^\gamma(U) := A_j^\gamma(U, t \to \infty)$ can be formulated as

$$\begin{aligned}
A_j^\gamma(E) &= -\frac{(\gamma-1)}{\gamma}\cdot\big[-T_j\log\pi_j(E)\big] \\
&= -\frac{(\gamma-1)}{\gamma}\cdot T_j\log\Big[N_j(E)e^{-E/T_j}\Big]^{-1} + \text{const} \\
&= -\frac{(\gamma-1)}{\gamma}\cdot\big[E - T_j\log N(E)\big] + \text{const}\ ,
\end{aligned}\tag{8}$$

After WTE has converged, the actual potential is essentially the superposition $U(\theta_j) + A_j^\gamma(U(\theta_j))$ of the biasing potential and the original unbiased one. With (7) and (8), the energy distribution reads

$$\begin{aligned} \pi_j^A(E) &\propto \int \delta\big[E - U(\theta_j)\big] e^{-\big[U(\theta_j)+A_j^\gamma(U(\theta_j))\big]/T_j}\, \mathrm{d}\theta_j \\ &= \left[\int \delta\big[E - U(\theta_j)\big]\, \mathrm{d}\theta_j\right] \exp\left[-\frac{E + A_j^\gamma(E)}{T_j}\right] \\ &= N(E) \exp\left[-\frac{E + (\gamma - 1)T_j \log N(E)}{\gamma T_j}\right] \\ &= \big[N(E) e^{-E/T_j}\big]^{1/\gamma} = \big[\pi_j(E)\big]^{1/\gamma}, \end{aligned}$$

which would give an approximately same average $\langle E \rangle$ as in the canonical ensemble but with the fluctuation $\mathbf{var}[E]$ amplified by a factor of γ.

□

An intuitive example can be obtained when the energy distribution is Gaussian, *i.e.* $\pi_j(E) \propto e^{-(E-\langle E\rangle)^2/2T_j}$, the well-tempered distribution is also Gaussian with the exactly same average but larger variance $\pi_j^A(E) = \big[\pi_j(E)\big]^{1/\gamma} \propto e^{-(E-\langle E\rangle)^2/2\gamma T_j}$.

The marginal distribution of θ_j for the WTE-augmented replica j is then modified as (*cf.* (**??**))

$$\begin{aligned} \tilde{\pi}_j^A(\theta_j) &\propto e^{-V_j(\theta_j)/T_j} \\ &= \exp\left[-\frac{U(\theta_j) + A_j(U(\theta_j))}{T_j}\right] \propto \pi_j(\theta_j)\, e^{-A_j(U(\theta_j))/T_j}, \end{aligned} \tag{9}$$

which deviates from the concerned marginal distribution $\pi_j(\theta_j)$ in (**??**) by a factor $e^{-A_j(U(\theta_j))/T_j}$. A re-weighting procedure needs to be conducted by simply implementing importance sampling with the same factor. In practical scenarios where WTE is deployed, large models, *e.g.* deep neural networks, often involve; it is usually the canonical average of some function $r(\theta_j)$, *i.e.* its Monte Carlo integration *w.r.t.* $\pi_j(\theta_j)$, rather than the posterior distribution $\rho(\theta|\mathcal{D}) \equiv \pi(\theta \,|\, T = 1)$ itself that really matters. For that average, we can readily evaluate it in a simple and unbiased way derived from (9):

$$\langle r(\theta_j)\rangle_{\pi_j} = \frac{\left\langle r(\theta_j) e^{A_j(U(\theta_j))/T_j} \right\rangle_{\tilde{\pi}_j^A}}{\left\langle e^{A_j(U(\theta_j))/T_j} \right\rangle_{\tilde{\pi}_j^A}}, \quad \text{with samples from } \tilde{\pi}_j^A, \tag{10}$$

where the biasing potential $A_j(U(\theta_j))$ can be evaluated on the fly during the simulation.

Now we devise WTE's update rule for replica j by setting an array to restore the biasing potential A_j. Given the granularity Δ, the energy E is quantized; each segment is then associated to one of the cells in that array. A_j is evaluated for all quantized E, with the values registered in the corresponding cells. Time is discretized $t \longrightarrow n\Delta t$ using the same steps; the differential equation (4) is hence converted into

$$A_j[E; n] \leftarrow A_j[E_j; n-1] + h\, \delta_{E, E_j^{(n)}} \exp\left[-\frac{A_j[E_j^{(n)}; n-1]}{(\gamma - 1)T_j}\right], \tag{11}$$

where the learning rate h controls the size of increments, $\delta_{E,E_j^{(n)}}$ defines the Kronecker delta function in the quantized E while $E_j^{(n)} := U(\theta_j(n\Delta t))$ denoting the potential energy evaluated at the n-th step. By initializing $A_j[E; 0] \equiv 0$, the biasing potential is adaptively accumulated through the simulation. Algorithm 1 provides a procedural description of RENHD with WTE deployed..

References

[1] Massimiliano Bonomi and Michele Parrinello. Enhanced sampling in the well-tempered ensemble. *Physical review letters*, 104(19):190601, 2010.

[2] James F Dama, Michele Parrinello, and Gregory A Voth. Well-tempered metadynamics converges asymptotically. *Physical review letters*, 112(24):240602, 2014.

[3] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[4] Ali A Minai and Ronald D Williams. On the derivatives of the sigmoid. *Neural Networks*, 6(6):845–853, 1993.

[5] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

[6] H. Risken and H. Haken. *The Fokker-Planck Equation: Methods of Solution and Applications Second Edition*. Springer, 1989.