Thermostat-assisted continuously-tempered Hamiltonian Monte Carlo for Bayesian learning

Rui Luo[1], **Jianhong Wang**[*1], **Yaodong Yang**[*1], **Zhanxing Zhu**[2], **and Jun Wang**[†1]
[1]University College London, [2]Peking University

Abstract

In this paper, we propose a novel sampling method, the thermostat-assisted continuously-tempered Hamiltonian Monte Carlo, for the purpose of multimodal Bayesian learning. It simulates a noisy dynamical system by incorporating both a continuously-varying tempering variable and the Nosé-Hoover thermostats. A significant benefit is that it is not only able to efficiently generate i.i.d. samples when the underlying posterior distributions are multimodal, but also capable of adaptively neutralising the noise arising from the use of mini-batches. While the properties of the approach have been studied using synthetic datasets, our experiments on three real datasets have also shown its performance gains over several strong baselines for Bayesian learning with various types of neural networks plunged in.

1 Introduction

Bayesian learning with Markov chain Monte Carlo (MCMC) methods is appealing for its inborn nature to capture the uncertainty of the learned parameters. However, when the distribution of interest contains multiple modes, efficient exploration across all those modes becomes difficult with many existing sampling techniques [7, 16]. In particular, when the number of modes is large, the "distant" modes could beyond the reach of any closest modes; this would lead to the so-called *pseudo-convergence* [1], where the ergodicity guarantee of MCMC methods breaks.

To make it worse, Bayesian learning on large datasets is typically conducted in an online setting: at each iteration, only a mini-batch of data are used to update the model [24]. While the requirements for computation are substantially reduced, mini-batches introduce noise and hence additional uncertainty into parameters, which makes multimodal posterior sampling even more difficult.

In this paper, we address these issues by proposing a new sampling method named as the thermostat-assisted continuously-tempered Hamiltonian Monte Carlo for multimodal posterior sampling on large datasets. In our model, we introduce a set of Nosé-Hoover thermostats [18, 11] in order to deal with the additional noise from mini-batches. This could effectively dissipate the instabilities arise from the noise so that the distribution of interest can be recovered [12]. We also introduce a novel systematic approach of continuous tempering brought from the recent advances in physics [8] and chemistry [15]: the original system for HMC is extended and coupled with additional degrees of freedom, i.e. the tempering variable and its the conjugate momentum. Continuous tempering is embedded into the original Hamiltonian system in a continuous fashion. As such, it could enhance the sampling efficiency by consistently adapting the temperature to lead the sampling trajectory to escape from the high energy barriers thus decreasing the chances of local trapping. To summarise, our model simulates a noisy system, which is augmented by a coupling tempering variable as well as a set of Nosé-Hoover thermostats. Various experiments are conducted to demonstrate the effectiveness of the new method on sampling complex multimodal distributions with large datasets. Our method consistently outperforms several samplers and optimisers on the accuracy of image classification.

[*]Equal
[†]Correspondence to: j.wang@cs.ucl.ac.uk

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Preliminaries

We review HMC [6] and continuous tempering [8, 15], the two bases of our model, where the former serves as a *de facto* standard for sampling whereas the latter is a state-of-the-art solution to the acceleration of molecular dynamics simulations on complex physical systems.

2.1 Hamiltonian Monte Carlo for posterior sampling

Bayesian posterior sampling aims at efficiently drawing i.i.d. samples from the posterior $\rho(\boldsymbol{\theta}|\mathbb{D})$ of the variable of interest $\boldsymbol{\theta}$ given the dataset $\mathbb{D}$. Provided the prior $\rho(\boldsymbol{\theta})$ and the likelihood $\mathscr{L}(\boldsymbol{\theta};\mathbb{D})$ along with the dataset $\mathbb{D} = \{\boldsymbol{x}_i\}$ of $|\mathbb{D}| = N$ independent datapoints $\boldsymbol{x}_i$, the target posterior to draw samples from can be formulated as

$$\rho(\boldsymbol{\theta}|\mathbb{D}) \propto \rho(\boldsymbol{\theta})\mathscr{L}(\boldsymbol{\theta};\mathbb{D}) = \rho(\boldsymbol{\theta})\prod_{i}^{N}\ell(\boldsymbol{\theta};\boldsymbol{x}_i), \text{ with the likelihood per datapoint } \ell(\boldsymbol{\theta};\boldsymbol{x}_i). \tag{1}$$

In a typical setting of HMC [16], the variable of interest $\boldsymbol{\theta} \in \mathbb{R}^D$ is often referred to as the system configuration in the configuration space $\mathscr{C} \subseteq \mathbb{R}^D$, where $\boldsymbol{\theta}$ can be interpreted as the joint position of all physical objects within a certain mechanical system that is related to the posterior $\rho(\boldsymbol{\theta}|\mathbb{D})$ in Eq. (1), while the configuration space $\mathscr{C}$ is the D-dimensional vector space spanned by all possible $\boldsymbol{\theta}$. An auxiliary variable $\boldsymbol{p}_\theta \in \mathbb{R}^D$ is then introduced as the conjugate momentum w.r.t. $\boldsymbol{\theta}$ to describe its rate of change. The tuple $\boldsymbol{\Gamma} = (\boldsymbol{\theta}, \boldsymbol{p}_\theta)$ represents the state of system that uniquely determines the physical system. A predefined matrix $\boldsymbol{M}_\theta = \text{diag}[m_{\theta_i}]$ that specifies the mass of the objects associated with $\boldsymbol{\theta}$ can be used for preconditioning. The connection between the target posterior and the physical system can be established via the potential $U(\boldsymbol{\theta})$, defined as

$$U(\boldsymbol{\theta}) = -\log\rho(\boldsymbol{\theta}|\mathbb{D}) = -\log\rho(\boldsymbol{\theta}) - \sum_{i=1}^{N}\log\ell(\boldsymbol{\theta};\boldsymbol{x}_i) - \text{const}\,. \tag{2}$$

The energy function $H(\boldsymbol{\Gamma})$ of the physical system, referred to as the Hamiltonian, is essentially sum of the potential in Eq. (2) and the conventional quadratic kinetic energy: $H(\boldsymbol{\Gamma}) = U(\boldsymbol{\theta}) + \boldsymbol{p}_\theta^\top \boldsymbol{M}_\theta^{-1}\boldsymbol{p}_\theta/2$. The Hamiltonian dynamics, i.e. the Hamilton's equations of motion, can be derived by applying the Hamiltonian formalism $[\dot{\boldsymbol{\theta}} = \partial_{\boldsymbol{p}_\theta}H,\ \dot{\boldsymbol{p}}_\theta = -\partial_{\boldsymbol{\theta}}H]$ on $H(\boldsymbol{\Gamma})$, where $\dot{\boldsymbol{\theta}}, \dot{\boldsymbol{p}}_\theta$ denote the time derivatives.

The Hamiltonian dynamics, on one hand, describes the time evolution of system from a microscopic perspective. The principle of statistical physics, on the other hand, states in a macroscopic sense: providing a proper mechanical system with the Hamiltonian $H(\boldsymbol{\Gamma})$, the states $\boldsymbol{\Gamma}$ of the system are distributed as a particular distribution, referred to as the canonical distribution, in the form of

$$\pi(\boldsymbol{\Gamma}) = \frac{1}{Z_\Gamma(T)}e^{-H(\boldsymbol{\Gamma})/T}, \text{ with the normalising constant } Z_\Gamma(T) = \sum_{\boldsymbol{\Gamma}} e^{-H(\boldsymbol{\Gamma})/T}, \tag{3}$$

where T is a predefined parameter that determines the system's absolute temperature. With $T = 1$ and $U(\boldsymbol{\theta})$ in Eq. (2), the canonical distribution in Eq. (3) can be marginalised as the posterior in Eq. (1).

2.2 Continuous tempering

In physical chemistry, continuous tempering [8, 15] is currently a state-of-the-art method to accelerate molecular dynamics simulations by means of continuously and systematically varying the temperature of a physical system. It extends the original system by coupling with additional degrees of freedom – namely the tempering variable $\xi \in \mathbb{R}$ with mass m_ξ and the conjugate momentum $p_\xi \in \mathbb{R}$ – that control the effective temperature of the original system in a continuous fashion via the Hamiltonian dynamics of the extended system. With a suitable choice of coupling function $\lambda(\xi)$ and a compatible confining potential $W(\xi)$, the Hamiltonian of the extended system can be designed as

$$H(\boldsymbol{\Gamma}) = \lambda(\xi)U(\boldsymbol{\theta}) + W(\xi) + \boldsymbol{p}_\theta^\top\boldsymbol{M}_\theta^{-1}\boldsymbol{p}_\theta/2 + p_\xi^2/2m_\xi, \tag{4}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\theta}, \xi, \boldsymbol{p}_\theta, p_\xi)$ denotes the state of the extended system with the position ξ and momentum p_ξ of the tempering variable augmented to the original state, and $\lambda(\xi) \in \mathbb{R}^+$ maps the tempering variable to a multiplier of temperature so that the effective temperature of the original system $T/\lambda(\xi)$ can vary, its domain $\text{dom}\lambda(\xi) \subset \mathbb{R}$ is a finite interval regulated by $W(\xi)$.

3 Thermostat-assisted continuously-tempered Hamiltonian Monte Carlo

In this section, we propose a sampling method, called the thermostat-assisted continuously-tempered Hamiltonian Monte Carlo (TACT-HMC), for multimodal posterior sampling in the presence of unknown noise. TACT-HMC leverages the extended Hamiltonian in Eq. (4) to raise the effective temperature periodically; this efficiently lowers the energy barriers between modes and hence accelerates sampling. Our method also incorporates the Nosé-Hoover thermostats to effectively recognise and automatically neutralise the approximation noise arising from the use of mini-batches.

3.1 System dynamics with the Nosé-Hoover augmentation

In solving for the system dynamics, we apply the Hamiltonian formalism to the extended Hamiltonian in Eq. (4), which requires the potential $U(\boldsymbol{\theta})$ and gradient $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$. We define hereafter the negative gradient of the potential $U(\boldsymbol{\theta})$ as the induced force $\boldsymbol{f}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$. Since the calculation of $U(\boldsymbol{\theta})$ and $\boldsymbol{f}(\boldsymbol{\theta})$ involves the entire dataset $\mathcal{D} = \{\boldsymbol{x}_i\}$, it is computationally expensive or even infeasible to calculate the actual values for large N. Instead, we consider the mini-batch approximations

$$\tilde{U}(\boldsymbol{\theta}) = -\log \rho(\boldsymbol{\theta}) - \frac{N}{S}\sum_{k=1}^{S} \log \ell(\boldsymbol{\theta}; \boldsymbol{x}_{i_k}) \quad \text{and} \quad \tilde{\boldsymbol{f}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \rho(\boldsymbol{\theta}) + \frac{N}{N_S}\sum_{k=1}^{N_S} \nabla_{\boldsymbol{\theta}} \log \ell(\boldsymbol{\theta}; \boldsymbol{x}_{i_k}),$$

where $\boldsymbol{x}_{i_k}$ are datapoints sampled from mini-batches $\mathcal{S} = \{\boldsymbol{x}_{i_k}\} \subset \mathcal{D}$ of size $|\mathcal{S}| = N_S \ll N$. It is clear that $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ are unbiased estimators of $U(\boldsymbol{\theta})$ and $\boldsymbol{f}(\boldsymbol{\theta})$.

As we assume $\boldsymbol{x}_{i_k}$ to be mutually independent, $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ are sums of N_S i.i.d. random variables, where the Central Limit Theorem (CLT) applies; the mini-batch approximations converge to Gaussian variables, i.e. $\tilde{U}(\boldsymbol{\theta}) \sim \mathcal{N}(U(\boldsymbol{\theta}), v_U(\boldsymbol{\theta}))$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{f}(\boldsymbol{\theta}), \boldsymbol{V}_f(\boldsymbol{\theta}))$ with the variance $v_U(\boldsymbol{\theta})$ and $\boldsymbol{V}_f(\boldsymbol{\theta})$. As random variables, $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ will inevitably introduce noise into the system dynamics. We incorporate a set of independent Nosé-Hoover thermostats [18, 11] – apparatuses originally devised for temperature stabilisation in molecular dynamics simulations – to adaptively cancel the effect of noise. The system dynamics with the augmentation of thermostats can be devised as follows

$$\begin{aligned}
&\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \boldsymbol{M}_{\theta}^{-1}\boldsymbol{p}_{\theta}, \qquad \frac{\mathrm{d}\boldsymbol{p}_{\theta}}{\mathrm{d}t} = \lambda(\xi)\tilde{\boldsymbol{f}}(\boldsymbol{\theta}) - \lambda^2(\xi)\boldsymbol{S}_{\theta}\boldsymbol{p}_{\theta}, \qquad \frac{\mathrm{d}s_{\theta}^{\langle i,j\rangle}}{\mathrm{d}t} = \frac{\lambda^2(\xi)}{\kappa_{\theta}^{\langle i,j\rangle}}\left[\frac{p_{\theta_i}p_{\theta_j}}{m_{\theta_i}} - T\delta_{ij}\right], \\
&\frac{\mathrm{d}\xi}{\mathrm{d}t} = \frac{p_{\xi}}{m_{\xi}}, \qquad \frac{\mathrm{d}p_{\xi}}{\mathrm{d}t} = -\lambda'(\xi)\tilde{U}(\boldsymbol{\theta}) - W'(\xi) - \left[\lambda'(\xi)\right]^2 s_{\xi}p_{\xi}, \quad \frac{\mathrm{d}s_{\xi}}{\mathrm{d}t} = \frac{\left[\lambda'(\xi)\right]^2}{\kappa_{\xi}}\left[\frac{p_{\xi}^2}{m_{\xi}} - T\right],
\end{aligned} \tag{5}$$

where $\boldsymbol{S}_{\theta}$ and s_{ξ} denote the Nosé-Hoover thermostats coupled with $\boldsymbol{\theta}$ and ξ, specifically, $\boldsymbol{S}_{\theta} = [s_{\theta}^{\langle i,j\rangle}]$ is a $D \times D$ matrix with the (i,j)-th component $s_{\theta}^{\langle i,j\rangle}$ corresponding to the pair of momenta $(p_{\theta_i}, p_{\theta_j})$, whereas $\kappa_{\theta}^{\langle i,j\rangle}$ and κ_{ξ} are constants that denote the "thermal inertia" corresponding to $s_{\theta}^{\langle i,j\rangle}$ and s_{ξ}. Intuitively, the thermostats $\boldsymbol{S}_{\theta}$ and s_{ξ} serve as negative feedback controllers on the momenta $\boldsymbol{p}_{\theta}$ and p_{ξ}. With the help of thermostats, the noise introduced by mini-batches can be adaptively neutralised, e.g. for the tempering variable ξ, when the quantity p_{ξ}^2/m_{ξ} exceeds the reference T as in Eq. (5), the associated thermostat s_{ξ} will increase, leading to a greater friction $-s_{\xi}p_{\xi}$, which in turn decreases p_{ξ} in the magnitude, resulting in a reduced value of p_{ξ}^2/m_{ξ}. Therefore, the feedback loop is established.

We define the diffusion coefficients $b_U(\boldsymbol{\theta}) \triangleq v_U(\boldsymbol{\theta})\,\mathrm{d}t/2$ and $\boldsymbol{B}_f(\boldsymbol{\theta}) = [b_f^{\langle i,j\rangle}(\boldsymbol{\theta})] \triangleq \boldsymbol{V}_f(\boldsymbol{\theta})\,\mathrm{d}t/2$ such that the variances $v_U(\boldsymbol{\theta})$ and $\boldsymbol{V}_f(\boldsymbol{\theta})$ within the mini-batch approximations evaluated in each of the discrete iterations can be exploited in the Fokker-Planck equation [20] (FPE) defined in continuous time. FPE interprets the microscopic motion of particles, formulated by SDEs, as the macroscopic time evolution of distribution of states, in the form of PDEs. With FPE leveraged, we have the following main theorem to obtain our needed canonical distribution in Eq. (3):

Theorem 1. *The system governed by the dynamics in Eq.* (5) *has the invariant distribution*

$$\pi(\boldsymbol{\Gamma}, \boldsymbol{S}_{\theta}, s_{\xi}) \propto e^{-\left[H(\boldsymbol{\Gamma}) + \left(s_{\xi} - \frac{b_U(\boldsymbol{\theta})}{m_{\xi}T}\right)^2 \kappa_{\xi}/2 + \sum_{i,j}\left(s_{\theta}^{\langle i,j\rangle} - \frac{b_f^{\langle i,j\rangle}(\boldsymbol{\theta})}{m_{\theta_j}T}\right)^2 \kappa_{\theta}^{\langle i,j\rangle}/2\right]/T}, \tag{6}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\theta}, \xi, \boldsymbol{p}_{\theta}, p_{\xi})$ *denotes the extended state as presented in Eq.* (4).

Proof. Recall FPE in its vector form [20]

$$\frac{\partial}{\partial t}\pi(\boldsymbol{x},t) = -\frac{\partial}{\partial \boldsymbol{x}} \cdot \Big[\boldsymbol{\mu}_x(\boldsymbol{x},t)\pi(\boldsymbol{x},t)\Big] + \left[\frac{\partial}{\partial \boldsymbol{x}}\frac{\partial^\top}{\partial \boldsymbol{x}}\right] \cdot \Big[\boldsymbol{B}_x(\boldsymbol{x},t)\pi(\boldsymbol{x},t)\Big], \tag{7}$$

where $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{\Gamma}, \boldsymbol{S}_\theta, s_\xi)$ denotes a vectorisation of the collection of all variables within Eq. (6), $\boldsymbol{\mu}_x$ and $\boldsymbol{B}_x$ are the drift and diffusion terms derived from the dynamics in Eq. (5), and the dot operator $\cdot$ defines the composition of element-wise multiplication and summation.

We substitute the system dynamics in Eq. (5) for the drift and diffusion terms of FPE in Eq. (7). As we assume that the introduced thermostats are mutually independent with each other, the invariant distribution can thus be factorised into marginals as $\pi(\boldsymbol{x}) = \pi_{\boldsymbol{\Gamma}}\pi_{s_\xi}\prod_{i,j}\pi_{s_\theta^{\langle i,j\rangle}}$. It is straightforward to check that the deterministic components with the only dependency on $\boldsymbol{\Gamma}$ cancel with each other exactly. The remains are the stochastic components as well as the deterministic ones that are related to the thermostats $\boldsymbol{S}_\theta$ and s_ξ, which can be formulated as

$$\begin{aligned}\frac{\partial}{\partial t}\pi(\boldsymbol{x},t) = {}& \frac{\partial}{\partial p_\xi}\Big[\big[\lambda'(\xi)\big]^2 s_\xi p_\xi \pi\Big] - \sum_{i,j}\frac{\partial}{\mathrm{d}s_\theta^{\langle i,j\rangle}}\left[\frac{\lambda^2(\xi)}{\kappa_\theta^{\langle i,j\rangle}}\left[\frac{p_{\theta_i}p_{\theta_j}}{m_{\theta_i}} - T\delta_{ij}\right]\pi\right] + \frac{\partial^2}{\partial p_\xi}\Big[\big[\lambda'(\xi)\big]^2 b_U(\boldsymbol{\theta})\pi\Big] \\ & + \frac{\partial}{\partial \boldsymbol{p}_\theta}\cdot\Big[\lambda^2(\xi)\boldsymbol{S}_\theta\boldsymbol{p}_\theta\pi\Big] - \frac{\partial}{\partial s_\xi}\left[\frac{\big[\lambda'(\xi)\big]^2}{\kappa_\xi}\left[\frac{p_\xi^2}{m_\xi} - T\right]\pi\right] + \left[\frac{\partial}{\partial \boldsymbol{p}_\theta}\frac{\partial^\top}{\partial \boldsymbol{p}_\theta}\right]\cdot\Big[\lambda^2(\xi)\boldsymbol{B}_f(\boldsymbol{\theta})\pi\Big].\end{aligned} \tag{8}$$

We solve for the invariant distribution $\pi(\boldsymbol{x})$ by equating Eq. (8) equal to zero. The resulted formulae for the marginals π_{s_ξ} and $\pi_{s_\theta^{\langle i,j\rangle}}$ are obtained under the assumption of factorisation:

$$\frac{1}{\pi_{s_\xi}}\frac{\partial \pi_{s_\xi}}{\partial s_\xi} = -\frac{\kappa_\xi}{T}\left[s_\xi - \frac{b_U(\boldsymbol{\theta})}{m_\xi T}\right] \quad \text{and} \quad \frac{1}{\pi_{s_\theta^{\langle i,j\rangle}}}\frac{\partial \pi_{s_\theta^{\langle i,j\rangle}}}{\partial s_\theta^{\langle i,j\rangle}} = -\frac{\kappa_\theta^{\langle i,j\rangle}}{T}\left[s_\theta^{\langle i,j\rangle} - \frac{b_f^{\langle i,j\rangle}(\boldsymbol{\theta})}{m_{\theta_j}T}\right]. \tag{9}$$

The solution to Eq. (9) is clear: both π_{s_ξ} and $\pi_{s_\theta^{\langle i,j\rangle}}$ are Gaussian distributions that are determined uniquely by the coefficients. The marginals π_{s_ξ} and $\pi_{s_\theta^{\langle i,j\rangle}}$, along with the canonical distribution $\pi_{\boldsymbol{\Gamma}}$ w.r.t. $H(\boldsymbol{\Gamma})$, constitute the invariant distribution defined in Eq. (6). □

Theorem 1 states that, when the system reaches the equilibrium, the states will be distributed as Eq. (6), whereas the unknown noise is separated from the extended Hamiltonian in Eq. (4) by the thermostats. Therefore, we can marginalise both $\boldsymbol{S}_\theta$ and s_ξ out to drop the noise, and then obtain the canonical distribution in Eq. (3). Since we are seeking for the marginal distribution w.r.t. $\boldsymbol{\theta}$ at a proper temperature in order to recover the target posterior, we can choose the tempering variables $\xi = \xi^*$ in such a way that the effective temperature of the original system is fixed at $T/\lambda(\xi^*) = 1$. The posterior $\rho(\boldsymbol{\theta}|\mathcal{D})$ then equals to the marginal distribution w.r.t. $\boldsymbol{\theta}$ given ξ^* and T, obtained by the marginalisation of $\boldsymbol{p}_\theta$ and p_ξ over the canonical distribution

$$\pi(\boldsymbol{\theta}|\xi^*) = \sum_{\boldsymbol{p}_\theta, p_\xi}\pi(\boldsymbol{\Gamma}|\xi^*) = \frac{\sum_{\boldsymbol{p}_\theta,p_\xi} e^{-H(\boldsymbol{\Gamma}|\xi^*)/T}}{\sum_{\boldsymbol{\Gamma}\setminus\xi} e^{-H(\boldsymbol{\Gamma}|\xi^*)/T}} = \frac{e^{-U(\boldsymbol{\theta})}}{\sum_{\boldsymbol{\theta}} e^{-U(\boldsymbol{\theta})}} = \frac{1}{Z_{\boldsymbol{\theta}}(T)}e^{-U(\boldsymbol{\theta})} = \rho(\boldsymbol{\theta}|\mathcal{D}),$$

where $H(\boldsymbol{\Gamma}|\xi^*) = TU(\boldsymbol{\theta}) + W(\xi^*) + \boldsymbol{p}_\theta^\top \boldsymbol{M}_\theta^{-1}\boldsymbol{p}_\theta/2 + p_\xi^2/2m_\xi$ represents the extended Hamiltonian conditioning on the tempering variable $\xi = \xi^*$ at the specific value.

3.2 Tempering enhancement via adaptive biasing force

A necessary condition for the tempering scheme to function properly is that the tempering variable ξ can properly explore the entire domain of the coupling function $\mathrm{dom}\lambda(\xi)$; this ensures the expected variation on the effective temperature during sampling. For complex systems, however, it is often the case that the tempering variable is subject to a strong instantaneous force that prevents ξ from proper exploration of $\mathrm{dom}\lambda(\xi)$ and hence hinders the efficiency of tempering. The adaptive biasing force (ABF) algorithm [3] has emerged as an promising solution to such issues after its inception [4], where it was introduced to address the issue on fast calculation of the free energy of complex chemical and biological systems. Intuitively, ABF maintains and updates an estimate of the average force, i.e. the average of the instantaneous force exerted on the variable of interest. It then applies the estimated average force to that variable in the opposite direction to counteract the instantaneous force and reduce it into small zero-mean fluctuations so that the variable undergoes random walks.

Algorithm 1 Thermostat-assisted continuously-tempering Hamiltonian Monte Carlo

Input: stepsize η_θ, η_ξ; level of injected noise c_θ, c_ξ; thermal inertia $\gamma_\theta, \gamma_\xi$; # of steps for unit interval K

$\boldsymbol{r}_\theta \sim \mathcal{N}(0, \eta_\theta \boldsymbol{I})$ and $r_\xi \sim \mathcal{N}(0, \eta_\xi)$; $(z_\theta, z_\xi) \leftarrow (c_\theta, c_\xi)$
INITIALISE($\boldsymbol{\theta}, \xi,$ `abf`, `samples`)
for $k = 1, 2, 3, \ldots$ **do**
  $\lambda \leftarrow$ LAMBDA(ξ); $\delta\lambda \leftarrow$ LAMBDADERIVATIVE(ξ)
  $z_\xi \leftarrow z_\xi + \delta\lambda^2 \left[r_\xi^2 - \eta_\xi\right] / \gamma_\xi$
  $z_\theta \leftarrow z_\theta + \lambda^2 \left[\boldsymbol{r}_\theta^\top \boldsymbol{r}_{\theta_j} / \dim(\boldsymbol{r}_\theta) - \eta_\theta\right] / \gamma_\theta$
  $\mathcal{S} \leftarrow$ NEXTBATCH($\mathcal{D}, k$); $\delta A \leftarrow$ `abf`[ABFINDEXING(ξ)]
  $\tilde{U} \leftarrow$ MODELFORWARD($\boldsymbol{\theta}, \mathcal{S}$); $\tilde{\boldsymbol{f}} \leftarrow$ MODELBACKWARD($\boldsymbol{\theta}, \mathcal{S}$)
  $r_\xi \leftarrow r_\xi - \delta\lambda \left[\eta_\xi \tilde{U} + \mathcal{N}(0, 2c_\xi\eta_\xi)\right] - \delta\lambda^2 z_\xi r_\xi + \eta_\xi \delta A$
  $\boldsymbol{r}_\theta \leftarrow \boldsymbol{r}_\theta + \lambda \left[\eta_\theta \tilde{\boldsymbol{f}} + \mathcal{N}(0, 2c_\theta\eta_\theta \boldsymbol{I})\right] - \lambda^2 z_\theta \boldsymbol{r}_\theta$
  ABFUPDATE(`abf`, $\xi, \delta\lambda, \tilde{U}, k$)
  $\xi \leftarrow \xi + r_\xi$
  if ISINSIDEWELL(ξ) = `false` **then** ▷ ξ is restricted by the *well* of infinite height.
    $r_\xi \leftarrow -r_\xi$; $\xi \leftarrow \xi + r_\xi$ ▷ ξ bounces back when hitting the wall.
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{r}_\theta$
  if $k = 0 \mod K$ and $\lambda = 0$ **then**
    APPEND(`samples`, $\boldsymbol{\theta}$) ▷ $\boldsymbol{\theta}$ is collected as a new sample in `samples`.
    $\boldsymbol{r}_\theta \sim \mathcal{N}(0, \eta_\theta \boldsymbol{I})$ and $r_\xi \sim \mathcal{N}(0, \eta_\xi)$ ▷ $\boldsymbol{r}_\theta, r_\xi$ is *optionally* resampled.
function ABFUPDATE(`abf`, $\xi, \delta\lambda, \tilde{U}, k$)
  $j \leftarrow$ ABFINDEXING(ξ) ▷ ξ is mapped to the index j of the associated bin.
  `abf`[j] $\leftarrow [1 - 1/k]$`abf`[j] $+ [1/k]\delta\lambda \cdot \tilde{U}$

Formally, the function of free energy w.r.t. ξ is defined by convention in the form of

$$A(\xi) = -T \log \pi(\xi) + \text{const}, \text{ where } \pi(\xi) = \sum_{\boldsymbol{\Gamma} \backslash \xi} \pi(\boldsymbol{\Gamma}) \text{ with the extended state } \boldsymbol{\Gamma} = (\boldsymbol{\theta}, \xi, \boldsymbol{p}_\theta, p_\xi).$$

The equation of p_ξ in Eq. (5) is then augmented with the derivative of $A(\xi)$ such that

$$\mathrm{d}p_\xi \,/\, \mathrm{d}t = -\lambda'(\xi)\tilde{U}(\boldsymbol{\theta}) - W'(\xi) + A'(\xi) - \left[\lambda'(\xi)\right]^2 s_\xi p_\xi, \tag{10}$$

where $A'(\xi)$ is referred to as the adaptive biasing force induced by the free energy as

$$A'(\xi) = -\frac{T}{\pi(\xi)} \frac{\mathrm{d}\pi}{\mathrm{d}\xi} = \frac{\sum_{\boldsymbol{\Gamma} \backslash \xi} \left[\frac{\partial H}{\partial \xi}\right] e^{-H(\boldsymbol{\Gamma})/T}}{\sum_{\boldsymbol{\Gamma} \backslash \xi} e^{-H(\boldsymbol{\Gamma})/T}} \triangleq \left\langle \frac{\partial H}{\partial \xi} \middle| \xi \right\rangle. \tag{11}$$

The brackets $\langle \cdot | \xi \rangle$ denotes the conditional average, i.e. the average on the canonical distribution $\pi(\boldsymbol{\Gamma})$ with ξ held fixed. $A'(\xi)$ is the average of the reversed instantaneous force on ξ. It is proved [14] that ABF converges to the equilibrium at which ξ's free energy landscape is flattened, even though the augmentation in Eq. (10) alters the equations of motion originally defined in Eq. (5).

3.3 Implementation

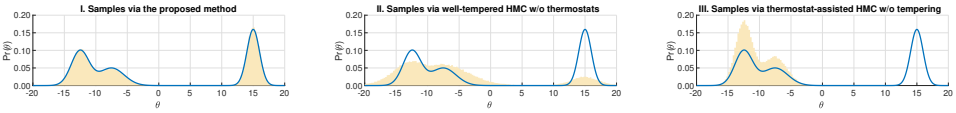
As proved in Theorem 1, the dynamics in Eq. (5) is capable of preserving the correct distribution in the presence of noise. In principle, it requires the thermostat $\boldsymbol{S}_\theta$ to be of size D^2 for the D-dimensional parameter $\boldsymbol{\theta}$; however, we cannot afford this for the complex models in high dimensions. A plausible option to mitigate this issue is to assume homogeneous $\boldsymbol{\theta}$ and isotropic Gaussian noise such that the mass $\boldsymbol{M}_\theta = m_\theta \boldsymbol{I}$ and the variance $\boldsymbol{V}_f(\boldsymbol{\theta}) = v_f(\boldsymbol{\theta})\boldsymbol{I}$; this simplifies the high-dimensional $\boldsymbol{S}_\theta$ to scalar s_θ. The confining potential $W(\xi)$ that determines the range of the tempering variable ξ is implemented as a well of infinite height. When colliding with the boundary of $W(\xi)$, ξ bounces back elastically with the velocity reversed. The Euler's method is then applied such that $\mathrm{d}t \to \Delta t$.

In Eq. (11), the calculation of $A'(\xi)$ involves the ensemble average $\langle \partial H / \partial \xi | \xi \rangle$, hence being intractable. Here we instead calculate the time average $\sum_k \partial H / \partial \xi|_{\xi_k}$, which is equivalent to the ensemble average in the long-time limit under the assumption of ergodicity; it can be readily calculated in a recurrent fashion during sampling. To maintain the runtime estimates, the range of ξ is divided uniformly into J bins of equal length with memory initialised on each of those bins. At each timestep k, ABF determines which bin j the tempering variable $\xi = \xi_k$ is currently located in, and then updates the time average using the record in memory and the current force $\partial H / \partial \xi|_{\xi_k}$ evaluated at ξ_k.

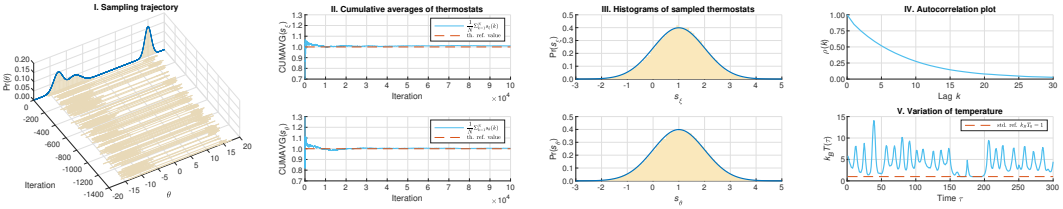
With all components assembled, we establish the TACT-HMC algorithm as Algorithm. 1 with

$$\boldsymbol{r}_\theta = \frac{\boldsymbol{p}_\theta \Delta t}{m_\theta}, \quad r_\xi = \frac{p_\xi \Delta t}{m_\xi}, \quad z_\theta = s_\theta \Delta t, \quad z_\xi = s_\xi \Delta t, \quad \eta_\theta = \frac{\Delta t^2}{m_\theta}, \quad \eta_\xi = \frac{\Delta t^2}{m_\xi}, \quad \gamma_\theta = \frac{\kappa_\theta}{m_\theta D}, \quad \gamma_\xi = \frac{\kappa_\xi}{m_\xi}$$

applied as the change of variables for the convenience of implementation.



(a) Histograms of samples drawn by TACT-HMC and the ablated alternatives, with the target indicated in blue.



(b) I: Sampling trajectory of TACT-HMC, presenting a robust mixing property; II: Cumulative averages of thermostats, indicating fast convergence to the theoretical reference values drawn by red lines; III: Histograms of sampled thermostats, showing a good fit to the theoretical distributions in blue curves; IV: Autocorrelation plot of samples, the decreasing of sample autocorrelation is significantly fast; V: (A snapshot of) the variation on effective system temperature during simulation, with the standard reference of unity temperature marked in red.

Figure 1: Experiment on sampling a $1d$ synthetic distribution.

4 Related work

From the inception of the stochastic gradient Langevin dynamics (SGLD) [24], stochastic methods [21] for Bayesian learning have received increasing attentions. By adding the right amount of noise into the standard stochastic gradient descent algorithm, SGLD manages to properly sample the posterior in a random-walk fashion akin to the Metropolis-adjusted Langevin algorithm (MALA) [22]. Chen et al. [2] extended the study of stochastic methods to HMC, and proposed the stochastic gradient Hamiltonian Monte Carlo (SGHMC). However, they have shown that the noise injected by stochastic gradient could drive the Hamiltonian dynamics to deviate from the correct distribution. SGHMC instead estimates the noise via the Fisher information matrix and compensate it by an additive friction term. It turns out that the friction term can be linked to the momentum within those momentum methods [19, 17, 23] in optimisation. Shortly after SGHMC, Ding et al. [5] came up with idea of incorporating the Nosé-Hoover thermostat [18, 11] into the standard Hamiltonian dynamics and devised the stochastic gradient Nosé-Hoover thermostat (SGNHT). Although proposed for temperature stabilisation in molecular dynamics simulations, the Nosé-Hoover thermostat has been proved, when properly configured, being able to automatically adapt to the unknown noise injected into the Hamiltonian system [12].

Parallel to the aforementioned researches, our work is most related to the method of continuously tempered Langevin dynamics (CTLD) [25], and the continuously-tempered HMC (CTHMC) algorithm [9]. CTHMC extends the Hamiltonian system with an extra Gibbs continuous tempering variate that empowers the dynamics to explore and mix between isolated modes in the distributions of interest. Such tempering variable is updated with a gradient-based HMC dynamics, requiring the information from the full batch data thus is unable to deal with the noise introduced by the mini-batch sampling. Our method however is designed to neutralise the noise by thermostats. CTLD also incorporates similar continuous tempering techniques; furthermore, it embeds the tempering variable into an extended stochastic gradient second-order Langevin dynamics so that the whole system could take into consideration the noise from stochastic approximation. Nonetheless, CTLD targets to serve as an initialiser for the optimisation phase of training deep neural networks. Its focus is to find the "good" wide valleys on the landscape of the objective function to enhance the subsequent gradient-based methods. TACT-HMC combines continuous tempering as well as the the Nosé-Hoover thermostats to tackle both the local trapping issue and the noise introduced by mini-batches, it therefore overcomes the limitations of all listed sampling methods.

5 Experiment

Two sets of experiments are carried out. We first conduct an ablation study with synthetic distributions, where we visualise the system dynamics and validate the efficacy of TACT-HMC. We then evaluate the performance of our method against several strong baselines on three real datasets.

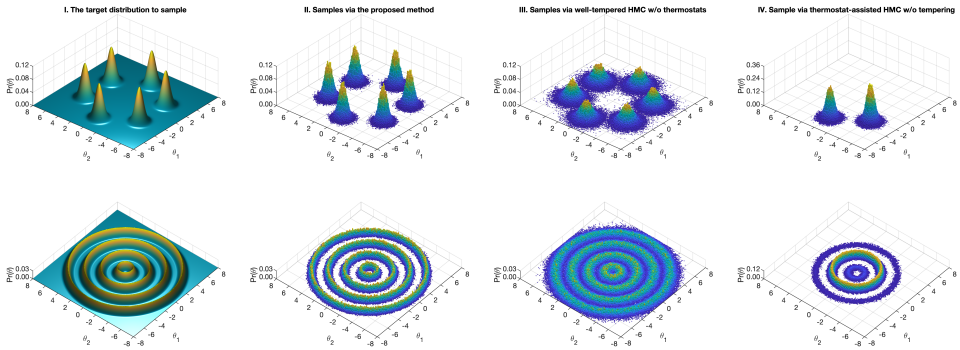


Figure 2: Experiments on sampling two $2d$ synthetic distributions. *Left*: The distributions to sample; *Mid-left*: Histogram of samples drawn by TACT-HMC; *Mid-right*: Histogram by the well-tempered sampler without thermostats; *Right*: Histogram by the thermostat-assisted sampler without tempering.

5.1 Multimodal sampling of synthetic distributions

We run TACT-HMC on three $1d/2d$ synthetic distributions. In the meantime, two ablated alternatives are initiated in parallel with the same setting: one is equipped with thermostats but without tempering for sampling acceleration, the other is well-tempered but with no thermostat for noise control. The distributions are synthesised to contain multiple distant modes; the calculation of gradient is perturbed by Gaussian noise that is unknown to all samplers.

Fig. 1 summarises the result of sampling a mixture of three $1d$ Gaussians. As the figure indicates, only TACT-HMC is capable of correctly sampling from the target. The sampler with no thermostat is heavily influenced by the gradient noise, resulting in a spread histogram; while the one without tempering gets trapped by the energy barriers and hence fails to explore the entire configuration space. The detailed sampling trajectory and the properties of sampled thermostats are illustrated in Fig. 1b, which further justifies the correctness of TACT-HMC. The autocorrelation of samples $\rho(k)$ is calculated and shown in Fig. 1(IV), which decreases monotonically from $\rho(0) = 1$. The effective sample size (ESS) can thus be readily evaluated through the formula

$$ESS = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho(k)}, \quad \text{with } \rho(k) \text{ as the autocorrelation at lag } k.$$

The ESS for TACT-HMC in this $1d$ Gaussian mixture is 2.1096×10^4 out of $n = 10^5$ samples, which is roughly 60.2% of the value for SGHMC and 50.9% of that for SGNHT. We believe that the non-linear interaction between the parameter of interest $\boldsymbol{\theta}$ and the tempering variable ξ via the multiplicative term $\lambda(\xi)U(\boldsymbol{\theta})$ results in a longer autocorrelation time and hence a lower ESS value. We also investigate the variation of the effective system temperature during sampling. A snapshot of the trajectory regarding the effective system temperature is demonstrated in Fig. 1(V): it constantly oscillates between higher and lower temperatures, and returns to the unity temperature occasionally.

We further conduct two $2d$ sampling experiments as shown in Fig. 2. Comparing between columns, we find that TACT-HMC recovers those multiple modes for both distributions while neutralising the influence of the gradient noise; however, the samplings from the ablated alternatives are impaired either by the gradient noise or by the energy barriers as discovered in the $1d$ scenario.

5.2 Bayesian learning on real datasets

Stepping out of the study on the simulation data, we then conduct the task of image classification on three real datasets: EMNIST[3], Fashion-MNIST[4] and CIFAR-10. To ensure the generality of the results, we evaluate on three types of deep neural nets: multilayer perceptrons (MLPs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs). The baselines comprises of two samplers, SGNHT [5] and SGHMC [2]; besides, we also consider the conventional first-order optimisers Adam [13] and Momentum [23].

All baselines are trained to the best on each model. For TACT-HMC, we augment Nosé-Hoover thermostats to the weights and biases respectively, i.e. each layer has two independent scalar variables.

[3] https://www.nist.gov/itl/iad/image-group/emnist-dataset

[4] https://github.com/zalandoresearch/fashion-mnist

Table 1: Result of Bayesian learning experiments on the real datasets

	MLP on EMNIST			RNN on Fashion-MNIST			CNN on CIFAR-10		
% permuted labels	0%	20%	30%	0%	20%	30%	0%	20%	30%
Adam [13]	83.39%	80.27%	80.63%	88.84%	88.35%	88.25%	69.53%	72.39%	71.05%
Momentum [23]	83.95%	82.64%	81.70%	88.66%	88.91%	88.34%	64.25%	65.09%	67.70%
SGHMC [2]	84.53%	82.62%	81.56%	90.25%	88.98%	88.49%	76.44%	73.87%	71.79%
SGNHT [5]	84.48%	82.63%	81.60%	90.18%	89.10%	88.58%	76.60%	73.86%	71.37%
TACT-HMC (Alg. 1)	**84.85**%	**82.95**%	**81.77**%	**90.84**%	**89.61**%	**89.01**%	**78.93**%	**74.88**%	**73.22**%

Each model runs 1000 epochs for training. To demonstrate the robustness of our method, we further randomly permute the training labels at each epoch. The accuracy of classification on test sets are evaluated for all methods with 0%, 20% and 30% of the labels permuted. For the baseline samplers, the final accuracy is calculated from Monte Carlo integration on all sampled models; for the baseline optimisers, they are evaluated directly on test sets after training. The result is summarised in Table. 1.

EMNIST classification with MLP. The MLP herein is a three-layer fully-connected neural network with the hidden layer consisting of 100 neurons. EMNIST Balanced is selected as the dataset, where 47 categories of images are split into a training set of size 112, 800 and a complementary test set of size 18, 800. The batch size is fixed at 128 for all methods of both sampling and training tests. For simplicity, we introduce a tuple of 7 components $[\eta_\theta, \eta_\xi, c_\theta, c_\theta, \gamma_\theta, \gamma_\xi, K]$ to set up TACT-HMC. In this tuple, $[\eta_\theta, c_\theta, \gamma_\theta]$ denote the step size, the level of injected Gaussian noise and the thermal inertia, all w.r.t. the parameter of interest $\boldsymbol{\theta}$; similarly, $[\eta_\xi, c_\xi, \gamma_\xi]$ represent the corresponding quantities for the tempering variable ξ; K defines the number of steps in simulating a unit interval. In this experiment, TACT-HMC is configured as $[0.0015, 0.0015, 0.05, 0.05, 1.0, 1.0, 50]$.

Fashion-MNIST classification with RNN. The RNN contains a LSTM layer [10] as the first layer, with the input/output dimensions to be $28/128$. It takes as the input via scanning a 28×28 image vertically each line of a time. After 28 steps of scanning, the LSTM outputs a representative vector of length 128 into ReLU activation, which is followed by a dense layer of size 64 with ReLU activation. A probabilistic predicting regarding 10 categories is generated through softmax activation in the final layer. The batch size is fixed at 64 for all methods in comparison. The configuration of TACT-HMC in this experiment is set as $[\eta_\theta, \eta_\xi, c_\theta, c_\theta, \gamma_\theta, \gamma_\xi, K] = [0.0012, 0.0012, 0.15, 0.15, 1.0, 1.0, 50]$.

CIFAR-10 classification with CNN. The CNN comprises of four learnable layers: from the bottom to the top, a $2d$ convolutional layer with the kernel of size $3 \times 3 \times 3 \times 16$, another $2d$ convolutional layer with the kernel of size $3 \times 3 \times 16 \times 16$, and two dense layers of size 100 and 10. ReLU activations are inserted between each of those learnable layers. For each convolutional layer, the stride is set to 1×1 and a pooling layer with 2×2 stride is added after the ReLU activation. Softmax function is exploited for the final prediction over 10 categories. The batch size is fixed at 64 for all methods. Here, TACT-HMC's configuration is $[0.0010, 0.0010, 0.10, 0.10, 1.0, 1.0, 50]$.

Discussion. As shown in Table. 1, TACT-HMC outperforms all the four baselines on the accuracy of classification. Specifically, TACT-HMC demonstrates advantages on complicated tasks, e.g. the CIFAR-10 classification with CNN where the model is relatively more complex and the dataset contains more channels. For the RNN task, our method outperforms others with roughly 0.5% on accuracy. The performance on the MLP task is rather limited; we believe the reason for this is that the complexities of both model and data are essentially moderate. When the random permutation is applied to a larger portion of training labels, TACT-HMC still maintains robust performance on the accuracy of classification, even though the landscape of the objective function becomes rougher and the system dynamics gathers more noise.

6 Conclusion

We proposed a novel sampling method named as the thermostat-assisted continuously-tempered Hamiltonian Monte Carlo for multimodal Bayesian learning. It incorporates the continuously-varying tempering variable and the Nosé-Hoover thermostats in simulating the noisy dynamical system. This method is developed for two substantial demands: first, to efficiently generate i.i.d. samples from complex multimodal posterior distributions; second, to adaptively neutralise the noise arising from the use of mini-batches. Extensive experiments have been conducted on both synthetic distributions and real-world datasets. The results validate the efficacy of the tempering and the thermostats techniques we adopted, and show great potentials for being applied to Bayesian learning tasks on complex models such as neural networks.

References

[1] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[2] Tianqi Chen, Emily B Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *ICML*, pages 1683–1691, 2014.

[3] Jeffrey Comer, James C Gumbart, Jérôme Hénin, Tony Lelièvre, Andrew Pohorille, and Christophe Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *The Journal of Physical Chemistry B*, 119(3):1129–1151, 2014.

[4] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20):9169–9183, 2001.

[5] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in neural information processing systems*, pages 3203–3211, 2014.

[6] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[7] Farhan Feroz and MP Hobson. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, 2008.

[8] Gianpaolo Gobbo and Benedict J Leimkuhler. Extended hamiltonian approach to continuous tempering. *Physical Review E*, 91(6):061301, 2015.

[9] Matthew M. Graham and Amos J. Storkey. Continuously tempered hamiltonian monte carlo. In Gal Elidan, Kristian Kersting, and Alexander T. Ihler, editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] William G Hoover. Canonical dynamics: equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[12] Andrew Jones and Ben Leimkuhler. Adaptive stochastic methods for sampling driven molecular systems. *The Journal of chemical physics*, 135(8):084125, 2011.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[14] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155, 2008.

[15] Nicolas Lenner and Gerald Mathias. Continuous tempering molecular dynamics: A deterministic approach to simulated tempering. *Journal of chemical theory and computation*, 12(2):486–498, 2016.

[16] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

[17] Yurii Nesterov. A method of solving a convex programming problem with convergence rate o(1/k2). *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[18] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[19] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

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

[21] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[22] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[23] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[24] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

[25] Nanyang Ye, Zhanxing Zhu, and Rafal Mantiuk. Langevin dynamics with continuous tempering for training deep neural networks. In *Advances in Neural Information Processing Systems*, pages 618–626, 2017.