Thermostat-assisted Continuous-tempered Hamiltonian Monte Carlo for Multimodal Posterior Sampling

Rui Luo[†‡], **Yaodong Yang**[†‡], **Jun Wang**[†], **and Yuanyuan Liu**[‡]
[†]Department of Computer Science, University College London
[‡]American International Group Inc.
{r.luo,yaodong.yang,j.wang}@cs.ucl.ac.uk, yuanyuan.liu@aig.com

Abstract

In this paper, we propose a new sampling method named as the thermostat-assisted continuous-tempered Hamiltonian Monte Carlo for multimodal posterior sampling on large datasets. It simulates a noisy system, which is augmented by a coupling tempering variable as well as a set of Nosé-Hoover thermostats. This augmentation is devised to address two main issues of concern: the first is to effectively generate i.i.d. samples from complex multimodal posterior distributions; the second is to adaptively control the system dynamics in the presence of unknown noise that arises from the use of mini-batches. The experiment on synthetic distributions has been performed; the result demonstrates the effectiveness of the proposed method.

1 Hamiltonian formalism for Bayesian posterior sampling

In Bayesian posterior sampling, one's interest typically lies in drawing i.i.d. samples efficiently from the posterior distribution $\rho(\boldsymbol{\theta}|\mathscr{D})$ of the variable of interest $\boldsymbol{\theta} \in \mathbb{R}^D$ given in the dataset $\mathscr{D}$. With the prior distribution $\rho(\boldsymbol{\theta})$ and the likelihood $\mathscr{L}(\boldsymbol{\theta};\mathscr{D})$ with respect to the dataset $\mathscr{D} = \{\boldsymbol{x}_i\}_{k=1}^N$ of $|\mathscr{D}| = N$ independent data points $\boldsymbol{x}_i$, the posterior distribution to be sampled is expressed as

$$\rho(\boldsymbol{\theta}|\mathscr{D}) \propto \rho(\boldsymbol{\theta})\mathscr{L}(\boldsymbol{\theta};\mathscr{D}) = \rho(\boldsymbol{\theta})\prod_i^N \ell(\boldsymbol{\theta};\boldsymbol{x}_i), \text{ with the likelihood per data point } \ell(\boldsymbol{\theta};\boldsymbol{x}_i). \quad (1)$$

As the random walk effect within both Metropolis algorithms [MRR+53, Has70] and Gibbs samplers [GG84] impairs sampling efficiency; Duane et al. [DKPR87] leveraged Hamiltonian dynamics for mitigation of such inefficiency in sampling and proposed the hybrid/Hamiltonian Monte Carlo (HMC) algorithm. In a typical setting of HMC [N+11], the variable of interest $\boldsymbol{\theta} \in \mathscr{C}$ is interpreted as the position of a physical system, or the system configuration, in the configuration space $\mathscr{C} \subset \mathbb{R}^D$ — the vector space spanned from the collection of all $\boldsymbol{\theta}$'s possible assignments; another variable $\boldsymbol{p}_\theta \in \mathbb{R}^D$ has been introduced as the conjugate momentum with respect to the position $\boldsymbol{\theta}$ to describe the rate of change in the latter. A matrix $\boldsymbol{M}_\theta = \mathrm{diag}[m_{\theta_i}]$ is defined to specify the mass of the configuration $\boldsymbol{\theta}$, which can be utilised for preconditioning. With all those definitions, a particular system, uniquely determined by the system state, i.e. the collection of position and momentum, has been established for HMC, where Hamiltonian dynamics applies — it governs the time evolution of system state.

To enable HMC to sample the target posterior, it is essential to introduce the concept of the canonical distribution. The theory of statistical physics states that the possible states Γ of a physical system with energy function $H(\Gamma)$ are distributed as a particular distribution, known as the canonical distribution

$$\pi(\Gamma;T) = \frac{1}{Z_\Gamma(T)} e^{-\frac{H(\Gamma)}{T}}, \text{ with the normalising constant } Z_\Gamma(T) = \sum_\Gamma e^{-\frac{H(\Gamma)}{T}}, \quad (2)$$

NIPS 2017 Workshop: *Advances in Approximate Bayesian Inference*, Long Beach, CA, USA.

where T denotes a predefined parameter representing the absolute temperature for the system with the Boltzmann constant k_B absorbed. In the established system for HMC, the energy function $H(\Gamma)$, referred to as the Hamiltonian of the system, is essentially the sum of potential and kinetic energy; the potential $U(\boldsymbol{\theta})$ is then defined as the negative logarithm of the posterior $\rho(\boldsymbol{\theta}|\mathscr{D})$ in Eq. (1):

$$U(\boldsymbol{\theta}) = -\log \rho(\boldsymbol{\theta}|\mathscr{D}) = -\log \rho(\boldsymbol{\theta}) - \sum_{i=1}^{N} \log \ell(\boldsymbol{\theta}; \boldsymbol{x}_i) - \text{const}, \tag{3}$$

which links the target posterior to the established system. In particular, the posterior $\rho(\boldsymbol{\theta}|\mathscr{D})$ equals to the marginal distribution w.r.t. $\boldsymbol{\theta}$ at the unity temperature $\pi(\boldsymbol{\theta}; T = 1) \propto e^{-U(\boldsymbol{\theta})}$ (cf. Eq. (2)).

HMC performs two alternative steps in each iteration: one simulates the established system with Hamiltonian dynamics, where the guided trajectory of state $\Gamma(t)$ passes through states on the contour of the Hamiltonian $H(\Gamma(t_0))$ specified by the initial state $\Gamma(t_0)$; the other involves resampling of momentum, by which the Hamiltonian is altered and the contour is switched from one to another.

In many complex systems, the configuration space $\mathscr{C}$ is divided into multiple sub regions by high energy barriers [LP02], where the probability for a trajectory of state to leap over an energy barrier of height ΔU is proportional to $e^{-\Delta U/T}$ [HO96]. It means that the high energy barriers will necessarily hinder efficient exploration of the entire configuration space: the trajectory tends to get trapped into some of the sub regions due to those barriers. One may notice that the probability to escape $e^{-\Delta U/T}$ increases as the temperature T rises, which forms the foundation of the methodology of tempering. Simulated tempering [MP92] and parallel tempering [ED05] are two typical tempering methods that aim to enhance the efficiency of sampling. Simulated tempering, also known as method of expanded ensembles [LMSVV92], performs a random walk on a ladder of temperature with discrete rungs, accompanied by simulation of the original system; parallel tempering, also referred to as replica-exchange Monte Carlo [SW86], simulates multiple replicas of the original system in parallel, each at a different temperature, and randomly exchanges configurations between systems. Both methods operate on discrete sets of temperature and rely on Metropolis updates at each iteration.

Recent advances in physics [GL15] and chemistry [LM16] suggest a systematic approach to continuous tempering with extended system: the original system for HMC is extended and coupled with additional degrees of freedom, namely the tempering variable $\xi \in \mathbb{R}$ of mass m_ξ and the conjugate momentum $p_\xi \in \mathbb{R}$, which control the effective temperature for the original system via Hamiltonian dynamics in a continuous fashion. The (extended) Hamiltonian of the extended system reads

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

where $\lambda(\xi) \in [\lambda_{min}, \lambda_{max}]$ denotes a coupling function that transforms the tempering variable to a multiplier of temperature such that the effective temperature for the original system $T\lambda(\xi)$ would vary; its domain $\mathrm{dom}\lambda(\xi) = [\xi_{min}, \xi_{max}]$ is a finite interval regulated by a confining potential $W(\xi)$.

To solve the system dynamics, i.e. the equation system of motion, from the extended Hamiltonian in Eq. (4), it requires to calculate both the potential $U(\boldsymbol{\theta})$ and its gradient $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$, where the entire dataset $\mathscr{D} = \{\boldsymbol{x}_i\}$ is involved as presented in Eq. (3). For simplicity, we denote the negative gradient of potential as $\boldsymbol{f}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \rho(\boldsymbol{\theta}) + \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \log \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)$, which is essentially an analogy to the conservative force induced by potential in physical systems. Provided a very large dataset $\mathscr{D}$, calculations of both the potential $U(\boldsymbol{\theta})$ and the induced force $\boldsymbol{f}(\boldsymbol{\theta})$ become computationally inefficient or even infeasible; we instead consider the noisy approximations $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ calculated on small subsets of the entire dataset $\mathscr{S} = \{\boldsymbol{x}_{i_k}\}_{k=1}^{S} \subset \mathscr{D}$, referred to as mini-batches of size $|\mathscr{S}| = S \ll N$, to approximate the exact values (cf. Eq. (3)):

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

where the indices i_k in the subset $\mathscr{S}$ are uniformly sampled from the index set $\{i\}_1^N$. Clearly, $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ are unbiased estimators of $U(\boldsymbol{\theta})$ and $\boldsymbol{f}(\boldsymbol{\theta})$. Since the data points $\boldsymbol{x}_i \in \mathscr{D}$ are independent from each other by definition, the approximations $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ are essentially sums of S i.i.d. random variables, where the Central Limit Theorem applies: the approximations asymptotically converge to Gaussian variables such as $\tilde{U}(\boldsymbol{\theta}) \sim \mathscr{N}(U(\boldsymbol{\theta}), V_U(\boldsymbol{\theta}))$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta}) \sim \mathscr{N}(\boldsymbol{f}(\boldsymbol{\theta}), \boldsymbol{V}_f(\boldsymbol{\theta}))$.

With the presence of noise, the system dynamics become a set of stochastic differential equations (SDEs). Chen et al. [CFG14] showed that the SDEs involving noisy approximations will lead to deviation from the distribution of interest; in their method, i.e. the Stochastic Gradient Hamiltonian Monte Carlo (SGHMC), the noise within stochastic gradient is estimated via Fisher information matrix, and its influence is mitigated by Langevin dynamics with properly selected coefficients. The method of Ding et al. [DFB+14], i.e. the Stochastic Gradient Nosé-Hoover Thermostat (SGNHT), instead leverages the Nosé-Hoover thermostat [Nos84, Hoo85] to adaptively dissipate the waste heat injected by noisy gradient and hence, retains the distribution of interest.

2 Thermostat-assisted Continuous-tempered Hamiltonian Monte Carlo

In this section, we propose a sampling method called the thermostat-assisted continuous-tempered Hamiltonian Monte Carlo (TACT-HMC), where we introduce a set of independent Nosé-Hoover thermostats into the extended Hamiltonian in Eq. (4) and couple those thermostats with the degrees of freedom of both the configuration $\boldsymbol{\theta}$ and the tempering variable ξ. With a proper coupling setting that is compatible with the tempering scheme, the thermostats can effectively recognise and automatically neutralise the approximation noise that arises from the use of mini-batches for calculations of both the potential and its gradient, and the distribution of interest is thus preserved for sampling.

We define the system dynamics of the noisy system by incorporating dynamics of Nosé-Hoover thermostats with pure Hamiltonian dynamics derived from Eq. (4):

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

where the variables $\boldsymbol{S}_\theta$ and s_ξ denote the Nosé-Hoover thermostats coupled to $\boldsymbol{\theta}$ and ξ; in particular, $\boldsymbol{S}_\theta = [s_\theta^{\langle i,j\rangle}]$ is a $D \times D$ matrix with the (i,j)-th component $s_\theta^{\langle i,j\rangle}$ associated with the pair $(p_{\theta_i}, p_{\theta_j})$; the constants $Q_\theta^{\langle i,j\rangle}$ and Q_ξ are thermal inertia corresponding to $s_\theta^{\langle i,j\rangle}$ and s_ξ, respectively. Intuitively, the thermostats $\boldsymbol{S}_\theta$ and s_ξ act as negative feedback controllers on the momenta $\boldsymbol{p}_\theta$ and p_ξ, which adaptively neutralise the influence of the noisy approximations $\tilde{f}(\boldsymbol{\theta})$ and $\tilde{U}(\boldsymbol{\theta})$. For instance, when the quantity p_ξ^2/m_ξ exceeds the reference value T, the associated thermostat s_ξ increases, resulting in a greater friction term $-s_\xi p_\xi$, which in turn reduces the magnitude of momentum p_ξ. We define the diffusion coefficients $b_\xi(\boldsymbol{\theta}) = \frac{1}{2}V_U(\boldsymbol{\theta})\,\mathrm{d}t$ and $\boldsymbol{B}_\theta(\boldsymbol{\theta}) = [b_\theta^{\langle i,j\rangle}(\boldsymbol{\theta})] = \frac{1}{2}\boldsymbol{V}_f(\boldsymbol{\theta})\,\mathrm{d}t$ for the derivation of the main theorem, which relates the continuous-time Fokker-Planck equation with the discrete-time evaluation of potential and gradient. With all elements assembled, the main theorem is presented as

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

$$\pi_{eq}(\Gamma, \boldsymbol{S}_\theta, s_\xi; T) \propto e^{-\left[H(\Gamma) + \frac{1}{2Q_\xi}\left(s_\xi - \frac{b_\xi(\boldsymbol{\theta})}{Tm_\xi}\right)^2 + \sum_{i,j}\frac{1}{2Q_\theta^{\langle i,j\rangle}}\left(s_\theta^{\langle i,j\rangle} - \frac{b_\theta^{\langle i,j\rangle}(\boldsymbol{\theta})}{Tm_{\theta_j}}\right)^2\right]\Big/T}, \tag{6}$$

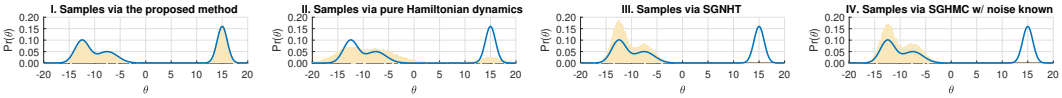
where $\Gamma = (\boldsymbol{\theta}, \xi, \boldsymbol{p}_\theta, p_\xi)$ denotes the joint state of the extended Hamiltonian as defined in Eq. (4).

Proof. The statement is followed by the application of the Fokker-Planck equation to Eq. (5). □

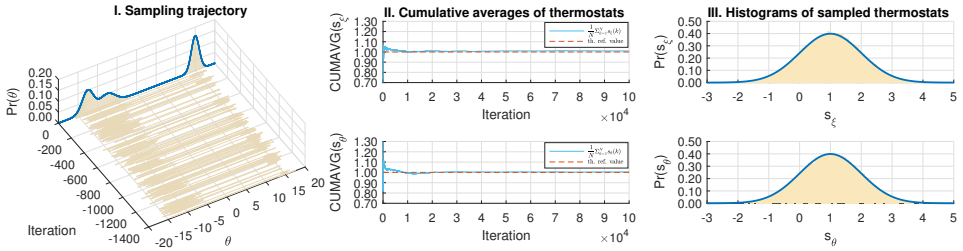
The main theorem states that, when the system reaches equilibrium, the states are distributed as Eq. (6), where the unknown noise is separated from the extended Hamiltonian $H(\Gamma)$ by thermostats. We can thus marginalise the thermostats $\boldsymbol{S}_\theta$ and s_ξ to get rid of the noise and then obtain the canonical distribution w.r.t. $H(\Gamma)$ defined in Eq. (2). As we have been looking for the marginal distribution w.r.t. $\boldsymbol{\theta}$ at a proper temperature such that the target posterior can be recovered, we choose specific values of the tempering variable $\xi = \xi^*$ in such a way that the effective temperature for the original system is fixed at $T\lambda(\xi^*) = 1$. With marginalisation on the canonical distribution with respect to the momenta $\boldsymbol{p}_\theta$ and p_ξ, the posterior $\rho(\boldsymbol{\theta}|\mathscr{D})$ can be formulated as

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

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



(a) Histograms of samples drawn by different methods, with the target distributions indicated by blue curves.



(b) *Left*: Sampling trajectory of TACT-HMC, indicating a robust mixing property; *Middle*: Cumulative averages of thermostats, showing fast convergence to the theoretical reference values drawn by red lines; *Right*: Histograms of sampled thermostats, presenting good fits to the theoretical distributions depicted by blue curves.

Figure 1: Result of experiment on sampling a 1D test distribution.

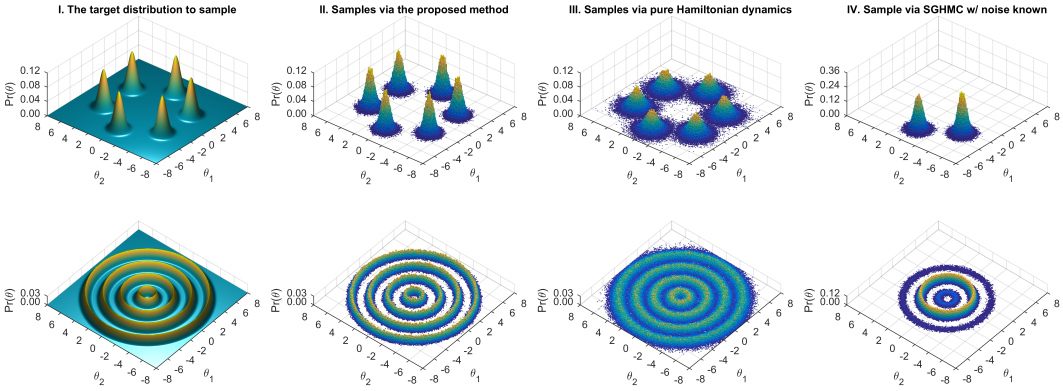


Figure 2: Result of experiments on sampling two 2D test distributions. Column 1 presents the target distributions: (1) a mixture of 6 Gaussians in Row 1 and (2) a composition of 4 concentric rings in Row 2. The sampled histograms are illustrated from Column 2 to 4 with each corresponding to a different sampling method: TACT-HMC, pure Hamiltonian dynamics, and SGHMC, respectively.

3 Experiment

To demonstrate the effectiveness of TACT-HMC, we performed experiments on three synthetic 1D/2D distributions. Three baselines were compared, namely SGNHT [DFB+14], SGHMC [CFG14] given the knowledge of noise, and pure Hamiltonian dynamics derived from Eq. (4) without thermostats.

Figure 1 summarises the result of experiment on sampling a mixture of three 1D Gaussians. In Fig. 1a, the histogram sampled by TACT-HMC is compared with those by baselines: only TACT-HMC sampled correctly from the distribution of interest; pure Hamiltonian dynamics was heavily affected by the noisy approximations of potential and its gradient, which results in a spread histogram; both SGNHT and SGHMC have got trapped by the potential barrier and hence failed to explore the entire configuration space. Details of the sampling trajectory and the properties of sampled thermostats are presented in Fig. 1b, which agree with the theoretical results and hence verify the correctness.

Figure 2 illustrates the sampled histograms of two 2D test distributions by different methods: correct samples were drawn by TACT-HMC in both cases whereas the samplings of baselines were impaired by either the approximation noise or the high energy barriers as found in the previous 1D experiment.

Disclaimer

This is not AIG work product and the paper expresses the views of the authors only in their individual capacity and not those of American International Group, Inc. or its affiliate entities.

References

[CFG14] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

[DFB+14] 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.

[DKPR87] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[ED05] David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[GL15] Gianpaolo Gobbo and Benedict J Leimkuhler. Extended hamiltonian approach to continuous tempering. *Physical Review E*, 91(6):061301, 2015.

[Has70] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HO96] Ulrich HE Hansmann and Yuko Okamoto. Monte carlo simulations in generalized ensemble: Multicanonical algorithm versus simulated tempering. *Physical Review E*, 54(5):5863, 1996.

[Hoo85] William G Hoover. Canonical dynamics: equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[LM16] 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.

[LMSVV92] AP Lyubartsev, AA Martsinovski, SV Shevkunov, and PN Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *The Journal of chemical physics*, 96(3):1776–1783, 1992.

[LP02] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[MP92] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

[MRR+53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[N+11] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

[Nos84] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[SW86] Robert H Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.