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.

Methodology

The (extended) Hamiltonian of the extended system reads

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

With huge dataset, noisy approximations $\tilde{U}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ is applied to approximate the exact potential and gradient:

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

We define the system dynamics of the noisy system by incorporating dynamics of Nosé-Hoover thermostats $\boldsymbol{S}_\theta$ and s_θ with pure Hamiltonian dynamics derived from Eq. (1):

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

The main theorem regarding the invariant distribution of the extended system is presented as

Theorem 1. *The system governed by the dynamics in Eq.* (2) *has the invariant distribution*

$$\pi_{eq}(\Gamma, \boldsymbol{S}_\theta, s_\xi; T) \propto \exp\left\{-\left[H(\Gamma) + \frac{1}{2Q_\xi}\left(s_\xi - \frac{b_\xi(\boldsymbol{\theta})}{Tm_\xi}\right)^2 + \sum_{i,j}\frac{1}{2Q_\theta^{\langle i,j\rangle}}\left(s_\theta^{\langle i,j\rangle} - \frac{b_\theta^{\langle i,j\rangle}(\boldsymbol{\theta})}{Tm_{\theta_j}}\right)^2\right]\Big/T\right\}, \quad (3)$$

where $\Gamma = (\boldsymbol{\theta}, \xi, \boldsymbol{p}_\theta, p_\xi)$ *denotes the joint state of the extended Hamiltonian as in Eq.* (1).

With marginalisation on the invariant distribution in Eq. (3) with respect to the thermostats $\boldsymbol{S}_\theta$ and s_ξ and momenta $\boldsymbol{p}_\theta$ and p_ξ, the posterior $\rho(\boldsymbol{\theta}|\mathscr{D})$ can be recovered as

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

with the tempering variable $\xi = \xi^*$ such that the effective temperature for the original system $T\lambda(\xi^*) = 1$.

Motivation

To enable *fast* sampling of complex posterior distributions with *multiple* modes seperated by low probability valleys provided *large datasets*.

Contribution

Our contribution is **three**-fold:

1. Efficient sampling of multimodal distributions using continuous tempering;
2. Adaptive heat dissipation for mini-batch approximation by thermostating;
3. Systematic integration of tempering and thermstating schemes.

Experiment

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

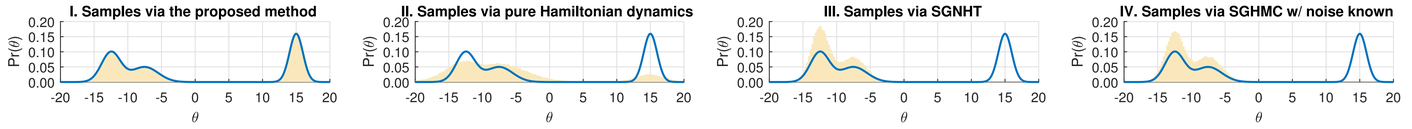

Figure 1: Histograms of samples drawn by different methods; target distributions indicated by blue curves.

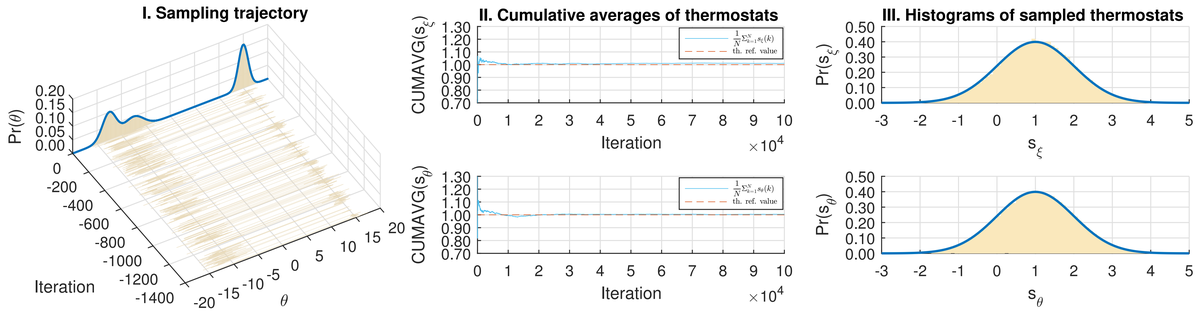

Figure 2: *Left*: Sampling trajectory of TACT-HMC, indicating a robust mixing property; *Middle*: Cumulative averages of thermostats, showing fast convergence to the theoretical reference values drawn by red lines; *Right*: Histograms of sampled thermostats, presenting good fits to the theoretical distributions depicted by blue curves.

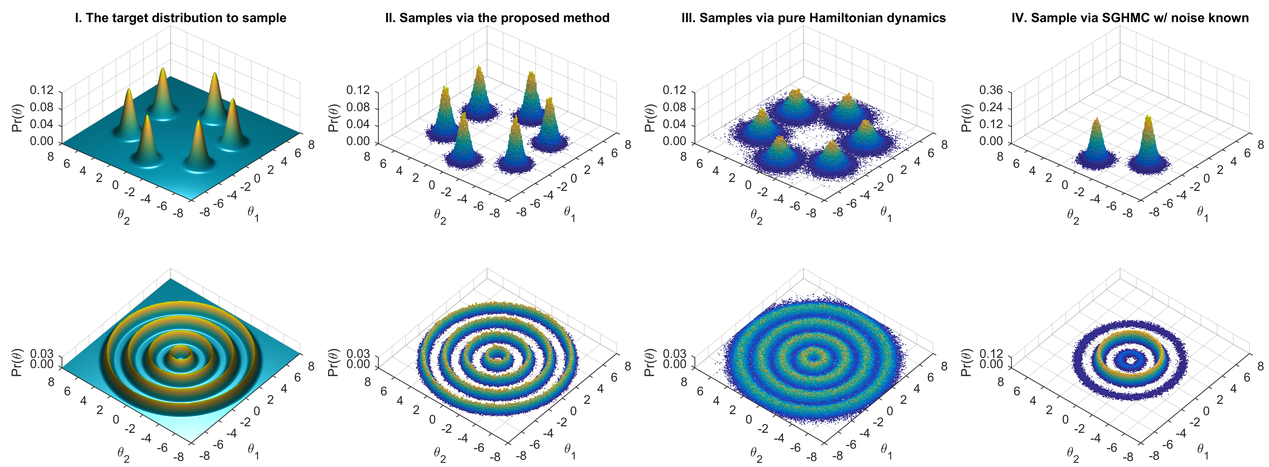

Figure 3: Column 1 shows the target distributions; Column 2 to 4 illustrate the sampled distograms, each by different method.

Discussion. Figure 1 and 3 summarises the result of experiment on sampling 1D/2D multimodal distributions. In Fig. 1 and 3, 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. 2, which agree with the theoretical results and hence verify the correctness.