Parallel-tempered Stochastic Gradient Hamiltonian Monte Carlo for Approximate Multimodal Posterior Sampling

Rui Luo*, Qiang Zhang*, and Yuanyuan Liu
American International Group Inc.

Methodology

For the target posterior $\pi(\theta|\mathcal{D})$, we establish a physical system with potential energy

$$U(\theta) = -\log \pi(\theta|\mathcal{D}) = -\log \pi(\theta) - \sum_{x\in\mathcal{D}} \log \ell(\theta; x) - \text{const}\,. \tag{1}$$

An increasing ladder $\{T_j\}_{j=1}^R$ of temperature is then defined with R rungs. The temperature ranges from $T_1 = 1$ to $T_R = T_{max}$. For each rung j, a replica of the physical system is initialized and the actual potential energy for that replica is scaled as $U(\theta_j)/T_j$.

We simulate all R replicas in a parallel fashion, by incorporating the dynamics of Noé-Hoover thermostats ξ:

$$\frac{\mathrm{d}\theta_j}{\mathrm{d}t} = M^{-1}p_j,\ \ \frac{\mathrm{d}p_j}{\mathrm{d}t} = -\nabla\tilde{U}(\theta_j)/T_j - \xi_j p_j,\ \ \frac{\mathrm{d}\xi_j}{\mathrm{d}t} = \left[p_j^\top M^{-1}p_j - D\right]\Big/Q, \tag{2}$$

where the noisy gradient arises from mini-batches

$$\nabla\tilde{U}(\theta) = -\nabla\log\pi(\theta) - \frac{N}{N_S}\sum_{x\in\mathcal{S}}\nabla\log\ell(\theta; x). \tag{3}$$

We occasionally perform partial exchanges upon the position coordinates θ between two replicas. For the condition of *detailed balance* to be met, the exchange is based on a properly-defined transition probability

$$\alpha[(i,j)\to(j,i)] = \frac{\pi_j(\theta_k)\pi_k(\theta_j)}{\pi_j(\theta_j)\pi_k(\theta_k) + \pi_j(\theta_k)\pi_k(\theta_j)} = \frac{1}{1+e^{-\delta E}},\ \ \text{with}\ \ \delta E = \left[U(\theta_k) - U(\theta_j)\right]\left[(T_k - T_j)/T_jT_k\right], \tag{4}$$

which means the transition probability α is logistic distribution.

As mini-batches are used, the evaluation of $U(\theta)$ is also noisy. According to the CLT, the noisy estimate $\tilde{U}(\theta) \sim \mathcal{N}(U(\theta), \sigma^2)$. It means that once we proceed to a test for replica exchange, we have to conduct some correction to the Gaussian noise; the correction distribution $p_{\mathcal{C}}$ needs to be calculated by deconvolution of logistic distribution with Gaussian.

Invoking the convolution theorem for probability distributions, the problem is converted in the form of characteristic functions

$$p_{\mathcal{C}} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\phi_{\mathcal{L}}(t)}{\phi_{\mathcal{N}_{\sigma^2}}(t)}e^{-ixt}\,\mathrm{d}t,\ \ \text{since}\ \ \phi_{\mathcal{C}} = \phi_{\mathcal{L}}/\phi_{\mathcal{N}_{\sigma^2}}, \tag{5}$$

where $\phi_{\mathcal{L}}$ and $\phi_{\mathcal{N}_{\sigma^2}}$ denote the characteristic functions of the standard logistic variable and Gaussian $\mathcal{N}(0,\sigma^2)$, respectively. Since the logistic distribution has much heavier tails than the Gaussians, the exact solution of $p_{\mathcal{C}}$ does not exist (the integrant on the RHS of Eq. (5) is not integrable). we can only calculate it approximately by introducing the kernel $\psi = e^{-\gamma^2t^4}$ with bandwidth $1/\gamma$:

$$\hat{p}_{\mathcal{C}} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\psi\cdot\frac{\phi_{\mathcal{L}}}{\phi_{\mathcal{N}_{\sigma^2}}}e^{-itx}\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{\psi}{\phi_{\mathcal{N}_{\sigma^2}}}\right]\phi_{\mathcal{L}}e^{-ixt}\,\mathrm{d}t. \tag{6}$$

With the help of Hermite polynomials H_k, we expand the quotient within the brackets of Eq. (6) as

$$\frac{\psi}{\phi_{\mathcal{N}_{\sigma^2}}} = e^{-\gamma^2t^4+\sigma^2t^2/2} = \sum_{k=0}^{\infty}\frac{\gamma^k}{k!}H_k(\sigma^2/4\gamma)t^{2k}. \tag{7}$$

The correction distribution can be approximated via Fourier's differential theorem:

$$\hat{p}_{\mathcal{C}} = \sum_{k=0}^{\infty}\frac{(-1)^k}{k!}H_k(\sigma^2/4\gamma)\gamma^k\left[\frac{1}{2\pi}\int_{-\infty}^{\infty}(-it)^{2k}\phi_{\mathcal{L}}e^{-itx}\,\mathrm{d}t\right] = \sum_{k=0}^{\infty}\frac{(-1)^k}{k!}H_k(\sigma^2/4\gamma)\gamma^k p_{\mathcal{L}}^{(2k)}, \tag{8}$$

where $p_{\mathcal{L}}^{(j)}$ represents the $(j+1)$-th derivative of logistic function, which can be efficiently calculated in a recursive fashion.

Experiment

We perform two sets of experiments on synthetic distributions: the first one is a $1d$ mixture of 4 Gaussians, and the second is a $2d$ Gaussian mixture with 5 modes. The evaluation of potential energy as well as its gradient is perturbed by Gaussian noise with variance $\sigma^2 = 0.25$. In the experiments, we have construct a temperature ladder with $R = 10$ rungs ranging from $T_1 = 1$ to $T_R = 10$, i.e. totally 10 replicas running in parallel. The proposed method is compared with classic HMC and the stochastic gradient version SGNHT.

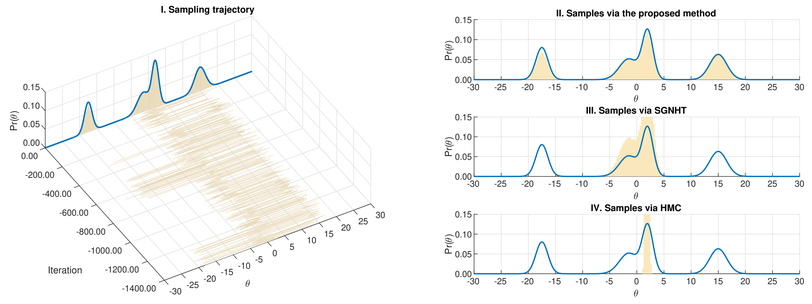

Figure 1: *Left*: Sampling trajectory of our method, indicating a robust mixing property; *Right*: Histograms of samples, presenting the advantage of correct sampling multimodal distribution with the presence of noise.

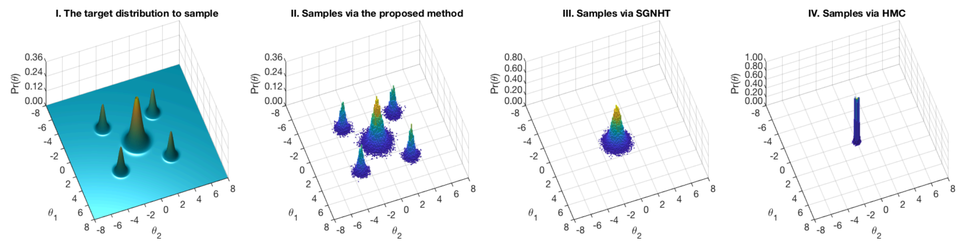

Figure 2: Column 1 shows the target distributions; Column 2 to 4 illustrate the sampled distograms, each by different method.

Discussion. It is demonstrated that, in both synthetic cases, our methods has accurately sampled the multimodal distributions in the presence of gradient noise where the baselines failed: SGNHT managed to control the gradient noise but did not discover the isolated modes whereas HMC is unable to correctly draw samples due to the deviated dynamics. Moreover, the subplot on the *left* of Fig. 1 illustrates the trajectory of the replica at the standard temperature, indicating a good mixing property.

Motivation

To enable *fast* and *robust* sampling of complex posterior distributions with *multiple* modes separated by low probability valleys under the circumstance of *large datasets*.

Contribution

Our contribution is **two**-fold:

1. Fast sampling of multimodal distributions by parallel tempering;
2. Adaptive noise cancellation of mini-batch noise using Nosé-Hoover thermostats;
3. Analytical form of the approximated solution to the mini-batch acceptance test for replica exchange.