Neural Sampling in Hierarchical Exponential-family Energy-based Models: Supplementary Information

Xingsi Dong[1]
dxs19980605@pku.edu.cn

Si Wu[1,2]
siwu@pku.edu.cn

1. PKU-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies. School of Psychology and Cognitive Sciences, Peking University.
2. IDG/McGovern Institute for Brain Research. Center of Quantitative Biology, Peking University.

1 Proof of equations in the main text

1.1 The derivative of Kullback–Leibler divergence

Derivation of Eq.(2) in the main text,

$$\nabla_\theta D_{\mathrm{KL}}\left[p_{\text{true}}(\mathbf{x}) \parallel p_\theta(\mathbf{x})\right], \tag{1a}$$

$$= -\mathbb{E}_{\mathbf{x}\sim p_{\text{true}}(\mathbf{x})}\frac{1}{p_\theta(\mathbf{x})}\nabla_\theta p_\theta(\mathbf{x}), \tag{1b}$$

$$= -\mathbb{E}_{\mathbf{x}\sim p_{\text{true}}(\mathbf{x})}\frac{1}{p_\theta(\mathbf{x})}\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}\left[\frac{\nabla_\theta p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right], \tag{1c}$$

$$= -\mathbb{E}_{\mathbf{x}\sim p_{\text{true}}(\mathbf{x})}\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}\left[\frac{\nabla_\theta p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{x},\mathbf{z})}\right], \tag{1d}$$

$$= -\mathbb{E}_{\mathbf{x}\sim p_{\text{true}}(\mathbf{x})}\mathbb{E}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{x})}\left[\nabla_\theta \ln p_\theta(\mathbf{x},\mathbf{z})\right]. \tag{1e}$$

1.2 The derivative of log-partition function

Considering that the sum of the probability equals to 1, we can obtain,

$$1 = \int p(\mathbf{x}_l|\mathbf{x}_{l+1})\mathrm{d}\mathbf{x}_l, \tag{2a}$$

$$= \int \exp\left[\boldsymbol{\eta}_l^T\phi(\mathbf{x}_l) + g(\mathbf{x}_l) - A(\boldsymbol{\eta}_l)\right]\mathrm{d}\mathbf{x}_l, \tag{2b}$$

$$= \frac{\int \exp\left[\boldsymbol{\eta}_l^T\phi(\mathbf{x}_l) + g(\mathbf{x}_l)\right]\mathrm{d}\mathbf{x}_l}{\exp\left[A(\boldsymbol{\eta}_l)\right]}. \tag{2c}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Then, we take the derivative of $\exp\left[A(\boldsymbol{\eta}_l)\right]$ with respect to $\boldsymbol{\eta}_l$, obtaining,

$$A'(\boldsymbol{\eta}_l) = \frac{\left(\exp\left[A(\boldsymbol{\eta}_l)\right]\right)'}{\exp\left[A(\boldsymbol{\eta}_l)\right]}, \tag{3a}$$

$$= \frac{\left(\int \exp\left[\boldsymbol{\eta}_l^T \phi(\mathbf{x}_l) + g(\mathbf{x}_l)\right] \mathrm{d}\mathbf{x}_l\right)'}{\exp\left[A(\boldsymbol{\eta}_l)\right]}, \tag{3b}$$

$$= \frac{\int \phi(\mathbf{x}_l) \exp\left[\boldsymbol{\eta}_l^T \phi(\mathbf{x}_l) + g(\mathbf{x}_l)\right] \mathrm{d}\mathbf{x}_l}{\exp\left[A(\boldsymbol{\eta}_l)\right]}, \tag{3c}$$

$$= \int \phi(\mathbf{x}_l) \exp\left[\boldsymbol{\eta}_l^T \phi(\mathbf{x}_l) + g(\mathbf{x}_l) - A(\boldsymbol{\eta}_l)\right] \mathrm{d}\mathbf{x}_l, \tag{3d}$$

$$= E_{\mathbf{x}_l \sim p_\theta(\mathbf{x}_l|\mathbf{x}_{l+})}\left[\phi(\mathbf{x}_l)\right]. \tag{3e}$$

1.3 Second-order Langevin dynamic

In this section, we provide a proof that the stationary distribution $\tilde{p}(\mathbf{z})$ of the second-order Langevin dynamic, as described in the main text, is equivalent to the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$.

$$\tau_z \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = \nabla_{\mathbf{z}} \ln p_\theta(\mathbf{z}|\mathbf{x}) - \mathbf{v} + \sqrt{2\tau_z}\xi, \tag{4}$$

$$\tau_v \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\frac{m\mathbf{v}}{2} + m\sqrt{2\tau_v}\xi, \tag{5}$$

where ξ is Gaussian white noise satisfying $\langle \xi \xi^T \rangle = \mathbf{I}\delta(t - t')$. Following the anti-symmetric matrix decomposition strategy [1], the above dynamic can be reorganised to,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \mathbf{z} \\ \mathbf{v} \end{pmatrix} = (D + Q)\begin{pmatrix} \nabla_{\mathbf{z}} \ln p_\theta(\mathbf{z}|\mathbf{x}) \\ -\tau_v \mathbf{v}/(4m\tau_z) \end{pmatrix} + \sqrt{2D}\xi, \tag{6}$$

where ξ is Gaussian white noise satisfying $\langle \xi \xi^T \rangle = \mathbf{I}\delta(t - t')$ and

$$D = \begin{pmatrix} I/\tau_z & 2mI\tau_v \\ 2m/\tau_v & 4m^2\tau_z/\tau_v^2 \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & 2mI\tau_v \\ -2m/\tau_v & 0 \end{pmatrix}. \tag{7}$$

The Fokker-Planck equation of the above dynamics is written as,

$$\frac{\partial}{\partial t} p(\mathbf{z}, \mathbf{v}, t) = \nabla^\top \left[(D + Q)\begin{pmatrix} \nabla_{\mathbf{z}} \ln p_\theta(\mathbf{z}|\mathbf{x}) \\ -\tau_v \mathbf{v}/(4m\tau_z) \end{pmatrix} p(\mathbf{z}, \mathbf{v}, t) \right] + \nabla^\top \left[D \nabla p(\mathbf{z}, \mathbf{v}, t) \right]. \tag{8}$$

By utilizing the property,

$$\nabla^\top \left[Q \nabla p(\mathbf{z}, \mathbf{v}, t) \right] = 0, \tag{9}$$

Eq.(8) can be rewritten as,

$$\frac{\partial}{\partial t} p(\mathbf{z}, \mathbf{v}, t) = \nabla^\top \left\{ (D + Q) \left[\begin{pmatrix} \nabla_{\mathbf{z}} \ln p_\theta(\mathbf{z}|\mathbf{x}) \\ -\tau_v \mathbf{v}/(4m\tau_z) \end{pmatrix} p(\mathbf{z}, \mathbf{v}, t) + \nabla p(\mathbf{z}, \mathbf{v}, t) \right] \right\}. \tag{10}$$

By letting the left-hand side of the above equation to be zero, we get that the stationary distribution $\tilde{p}(\mathbf{z}, \mathbf{v})$ satisfy,

$$\begin{pmatrix} \nabla_{\mathbf{z}} \ln p_\theta(\mathbf{z}|\mathbf{x}) \\ -\tau_v \mathbf{v}/(4m\tau_z) \end{pmatrix} p(\mathbf{z}, \mathbf{v}) + \nabla p(\mathbf{z}, \mathbf{v}) = 0, \tag{11}$$

which gives the solution,

$$\tilde{p}(\mathbf{z}, \mathbf{v}) = p_\theta(\mathbf{z}|\mathbf{x}) \mathcal{N}(\mathbf{v}; 0, 4m\tau_z/\tau_v). \tag{12}$$

The marginal stationary distribution of $\mathbf{z}$ is then calculated to be,

$$\tilde{p}(\mathbf{z}) = \int \tilde{p}(\mathbf{z}, \mathbf{v}) \mathrm{d}\mathbf{v} = p_\theta(\mathbf{z}|\mathbf{x}). \tag{13}$$

2 Analysis of Hessian matrix

In this section, we provide a theoretical lower bound for $\lambda_1(H)$ and an upper bound for $\det(H)$ based on the assumption that $\phi(\mathbf{x}_l) = \mathbf{x}_l$ and $g(\mathbf{x}_l) = -\mathbf{x}_l^T\mathbf{x}_l$. Then we demonstrate that HEE-L can only converge to unimodal distributions.

We define the total energy $F_{\mathrm{t}} = -\ln p_\theta(\mathbf{x}_{0:L})$. Considering

$$\frac{\partial^2}{\partial \mathbf{x}_l \partial \mathbf{x}_{l+k}} F_{\mathrm{t}} = \begin{cases} I + \theta_{l-1}^T\theta_{l-1} & , k = 0 \\ -\theta_l & , k = 1 \\ -\theta_{l-1} & , k = -1 \\ 0 & , |k| > 1 \end{cases} \tag{14}$$

therefore, the Hessian matrix $H_{\mathrm{t}} = \Delta_{\mathbf{x}_{0:L}} F_{\mathrm{t}}$ is calculated as,

$$H_{\mathrm{t}} = (I + D)^T(I + D) \tag{15}$$

where

$$D = \begin{pmatrix} 0 & -\theta_0 & & & \\ & 0 & -\theta_1 & & \\ & & \cdot & \cdot & \\ & & & 0 & -\theta_{L-1} \\ & & & & 0 \end{pmatrix} \tag{16}$$

Thus, the lower bound of singular values of $I + D$ is given by (see Theorem 8.13 in [2]),

$$\sigma_1(I + D) \geq \sigma_1(I) - \sigma_{\max}(D) \tag{17}$$

where $\sigma_1(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the smallest and greatest singular value, respectively. Then we assume that θ_l is matrix with sub-gaussian entries, and $n_1 = \cdots = n_L = N/L$ then we have,

$$\begin{aligned} \mathbb{E}\left[\sigma_{\max}(D)\right] &= \mathbb{E}\left[\max_{l \leq L} \sigma_{\max}(\theta_l)\right], \\ &\leq K\sqrt{\ln L}\,\mathbb{E}\left[\sigma_{\max}(\theta_L)\right], && \text{Exercise 2.5.10 in [3]} \\ &\leq K\sqrt{\tfrac{N \ln L}{L}}, && \text{Theorem 4.4.5 in [3]} \end{aligned} \tag{18}$$

where K is a small constant. Thus, the lower bound of $\lambda_1(H_{\mathrm{t}})$ is calculated as

$$\begin{aligned} \lambda_1(H_{\mathrm{t}}) &= \sigma_1^2(I + D) && \text{Eq.}(15) \\ &\geq \left(1 - K\sqrt{\tfrac{N \ln L}{L}}\right)^2 && \text{Eq.}(18) \end{aligned} \tag{19}$$

Thus, the lower bound of $\lambda_1(H)$ is calculated as (Theorem 4.3.15 in [4]),

$$\lambda_1(H) \geq \lambda_1(H_{\mathrm{t}}). \tag{20}$$

And the upper bound of $\det(H)$ is calculated as,

$$\begin{aligned} \det(H) &= \textstyle\prod_{i=1}^N \lambda_i(H) \\ &\leq \textstyle\prod_{i=1}^N \lambda_{i+n_0}(H_{\mathrm{t}}) && \text{Theorem 4.3.15 in [4]} \\ &= 1/\textstyle\prod_{i=1}^{n_0} \lambda_i(H_{\mathrm{t}}) && \text{Utilizing } \det(H_{\mathrm{t}}) = 1 \\ &\leq 1/\left[\lambda_1(H_{\mathrm{t}})\right]^{n_0} \end{aligned} \tag{21}$$

The above analysis reveals that $\lambda_1(H_{\mathrm{t}})$ increase with the decrease of L. Additionally, as the value of $\lambda_1(H_{\mathrm{t}})$ increases, the lower bound of $\lambda_1(H)$ and the upper bound of $\det(H)$ will decrease and increase, respectively.

Then, we prove that the Hessian matrix $H_{\rm o} = -\Delta_{\mathbf{x}_0} \ln p_\theta(\mathbf{x}_0)$ is positive-definite, indicating that HEE-L is capable of approximating only unimodal distributions.

$$
\begin{aligned}
H_{\rm o} &= -\Delta_{\mathbf{x}_0} \ln p_\theta(\mathbf{x}_0), && \text{(22a)}\\
&= -\Delta_{\mathbf{x}_0} \ln \int p_\theta(\mathbf{x}_{0:L}) \mathrm{d}\mathbf{x}_{1:L}, && \text{(22b)}\\
&= \frac{-1}{\left[p_\theta(\mathbf{x}_0)\right]^2} \int \mathrm{d}\mathbf{x}'_{1:L} \int \mathrm{d}\mathbf{x}_{1:L} \left[p_\theta(\mathbf{x}'_{0:L}) \Delta_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) - \nabla_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) \nabla^T_{\mathbf{x}'_0} p_\theta(\mathbf{x}'_{0:L}) \right] && \text{(22c)}\\
&= \frac{-2}{\left[p_\theta(\mathbf{x}_0)\right]^2} \iint \mathrm{d}\mathbf{x}_{1:L} \mathrm{d}\mathbf{x}'_{1:L} \left[P(\mathbf{x}_{0:L}, \mathbf{x}'_{0:L}) + P(\mathbf{x}'_{0:L}, \mathbf{x}_{0:L}) \right], && \text{(22d)}\\
&\succ \frac{-2}{\left[p_\theta(\mathbf{x}_0)\right]^2} \iint \mathrm{d}\mathbf{x}_{1:L} \mathrm{d}\mathbf{x}'_{1:L} \left[P(\mathbf{x}_{0:L}, \mathbf{x}_{0:L}) + P(\mathbf{x}'_{0:L}, \mathbf{x}'_{0:L}) \right], && \text{(22e)}\\
&= \frac{-1}{\left[p_\theta(\mathbf{x}_0)\right]^2} \iint \mathrm{d}\mathbf{x}_{1:L} \mathrm{d}\mathbf{x}'_{1:L} \left[p_\theta(\mathbf{x}_{0:L}) \right]^2 \Delta_{\mathbf{x}_0} \ln p_\theta(\mathbf{x}_{0:L}), && \text{(22f)}\\
&\succ 0 && \text{(22g)}
\end{aligned}
$$

where $P(\mathbf{x}_{0:L}, \mathbf{x}'_{0:L}) = p_\theta(\mathbf{x}'_{0:L}) \Delta_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) - \nabla_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) \nabla^T_{\mathbf{x}'_0} p_\theta(\mathbf{x}'_{0:L})$. In Eq.(22e), we use the property

$$
\begin{aligned}
&\Delta_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) = \Delta_{\mathbf{x}'_0} p_\theta(\mathbf{x}'_{0:L}), && \text{(23a)}\\
&\nabla_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) \nabla^T_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) + \nabla_{\mathbf{x}'_0} p_\theta(\mathbf{x}'_{0:L}) \nabla^T_{\mathbf{x}'_0} p_\theta(\mathbf{x}'_{0:L}) \succ 2 \nabla_{\mathbf{x}_0} p_\theta(\mathbf{x}_{0:L}) \nabla^T_{\mathbf{x}'_0} p_\theta(\mathbf{x}'_{0:L}) && \text{(23b)}
\end{aligned}
$$

In Eq.(22g), we utilize the property that $-\Delta_{\mathbf{x}_0} \ln p_\theta(\mathbf{x}_{0:L})$ is a principal submatrix of $H_{\rm t}$, which is also a positive-definite matrix.

3 Experimental details

Table 1 lists experimental settings for *2D synthetic datasets*, *FashionMNIST* and *CIFAR10*.

Settings	*2D synthetic datasets*	*FashionMNIST*	*CIFAR10*
$f(\mathbf{x}_l)$	$\mathbf{x}_l$	$\mathbf{x}_l$	$\mathbf{x}_l$
$g(\mathbf{x}_l)$	$-\mathbf{x}_l^T \mathbf{x}_l$	$-\mathbf{x}_l^T \mathbf{x}_l$	$-\mathbf{x}_l^T W_l \mathbf{x}_l - \mathrm{Tr}(W_l^T W_l)$
L	5	5	10
n_l	linspace$(10, 10, L)$	linspace$(1000, 100, L)$	linspace$(3000, 100, L)$
τ_z	linspace$(1, 10, L)$	linspace$(1, 10, L)$	linspace$(10, 10, L)$
τ_x	1	1	10
τ_v	$0.5 * \tau_z$	$0.5 * \tau_z$	$0.5 * \tau_z$
m	0.5	0.5	0.5
τ_u	0.1	0.1	0.1
τ_θ	10	10	100
sparse connection	False	False	True (20%)
parameters	400	400K	4M

Table 1: Experimental settings

We conducted all experiments on a GPU (RTX A6000). The inference, learning, and generation dynamics were simulated using the Euler expansion with a step size of $\mathrm{d}t = 0.01\tau_x$. In the inference and learning phase, each data point was presented to the network for $300\tau_x$, with each batch consisting of 30K steps. The *FashionMNIST* model was trained with a batch size of 512, taking approximately 0.5 hours to complete one epoch. The *CIFAR10* model was trained with a batch size of 128, taking approximately 2 hours to complete one epoch. And the parameters W_l and $\mathbf{b}_l$ adopts the gradient method described by Eq.(4) in the main text.

For joint generation, we randomly initialized the neuronal activities and performed the joint generation dynamic for $300\tau_x$. The average value of neuronal activity $\mathbf{x}_0$ in the last $100\tau_x$ was used as the

generated output. For marginal generation, we also randomly initialized the neuronal activities and performed the marginal generation dynamic for $100\tau_x$. The average value of neuronal activity $\mathbf{x}_0$ in the last $50\tau_x$ was used as the generated output.

References

[1] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

[2] Fuzhen Zhang. *Matrix theory: basic results and techniques*. Springer, 2011.

[3] Omiros Papaspiliopoulos. High-dimensional probability: An introduction with applications in data science, 2020.

[4] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.