Variational (Gradient) Estimate of the Score Function in Energy-based Latent Variable Models

Fan Bao[12], **Kun Xu**[1], **Chongxuan Li**[1], **Lanqing Hong**[2], **Jun Zhu**[*1], **Bo Zhang**[1]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, THBI Lab, BNRist Center,
State Key Lab for Intell. Tech. & Sys., Tsinghua University, Beijing, China
[2]Huawei Noah's Ark Lab
bf19@mails.tsinghua.edu.cn,{kunxu.thu, chongxuanli1991}@gmail.com,
honglanqing@huawei.com,{dcszj, dcszb}@tsinghua.edu.cn

Abstract

The learning and evaluation of energy-based latent variable models (EBLVMs) without any structural assumptions are highly challenging, because the true posteriors and the partition functions in such models are generally intractable. This paper presents variational estimates of the score function and its gradient with respect to the model parameters in a general EBLVM, referred to as *VaES* and *VaGES* respectively. The variational posterior is trained to minimize a certain divergence to the true model posterior and the bias in both estimates can be bounded by the divergence theoretically. With a minimal model assumption, VaES and VaGES can be applied to the *kernelized Stein discrepancy* (KSD) and *score matching* (SM)-based methods to learn EBLVMs. Besides, VaES can also be used to estimate the *exact Fisher divergence* between the data and general EBLVMs.

1 Introduction

An energy-based model (EBM) [17] firstly associates an energy $\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v})$ with parameter $\boldsymbol{\theta}$ to each configuration of visible variables $\boldsymbol{v}$, and then normalizes the energy to get a probability density as

$$p_{\boldsymbol{\theta}}(\boldsymbol{v}) = \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})/\mathcal{Z}(\boldsymbol{\theta}) = e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v})}/\mathcal{Z}(\boldsymbol{\theta}),$$

where $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})$ is the unnormalized density and $\mathcal{Z}(\boldsymbol{\theta}) = \int e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v})}\mathrm{d}\boldsymbol{v}$ is called the partition function. As for the learning and evaluation, maximum likelihood estimate (MLE) [9, 31, 36, 24, 25, 2, 3, 7] and noise-contrastive estimation (NCE) [8, 26, 5] are commonly used strategies. However, MLE requires to estimate the partition function and NCE requires to design a proper noise distribution manually. In comparison, Fisher divergence based methods provide an efficient and effective alternative to train [13, 29, 34, 20] and evaluate [6] such models. The calculation of the Fisher divergence is based on the *score function* of an EBM, which is defined as $\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})$. Notably, the score function is independent of the partition function because $\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}) - \cancel{\nabla_{\boldsymbol{v}} \log \mathcal{Z}(\boldsymbol{\theta})} = \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})$, and thereby tractable.

Energy-based latent variable models (EBLVMs) [9, 11, 27, 28, 16, 19, 1] incorporate hidden variables $\boldsymbol{h}$ to EBMs, which increases the model capacity and enables feature extraction. Such a model defines a joint probability density over the visible variables $\boldsymbol{v}$ and the latent variables $\boldsymbol{h}$ as follows

$$p_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h}) = \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})/\mathcal{Z}(\boldsymbol{\theta}) = e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})}/\mathcal{Z}(\boldsymbol{\theta}),$$

where $\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})$ is the energy function with $\boldsymbol{\theta}$ as its parameters, $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})$ is the unnormalized density and $\mathcal{Z}(\boldsymbol{\theta}) = \int e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})}\mathrm{d}\boldsymbol{v}\mathrm{d}\boldsymbol{h}$ is the partition function. At the cost of increasing expressiveness

*Corresponding author.

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

power, the posteriors of EBLVMs $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ are generally intractable without a structural assumption [30, 33]. As a result, commonly used learning and evaluating algorithms in EBMs are not directly applicable to general nonstructural EBLVMs. To learn and evaluate general nonstructural EBLVMs, we propose variational estimates of the score function and its gradient w.r.t. the model parameters in such models, referred to as *VaES* and *VaGES*. Such estimates firstly introduce a tractable variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ to approximate the true one $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$, and then estimate the score function $\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})$ or its gradient $\frac{\partial \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial \boldsymbol{\theta}}$ by Monte Carlo methods based on samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. We show that under some assumptions, the bias introduced by the variational posterior can be bounded by the square root of the KL divergence or the Fisher divergence [15] between the variational posterior and the true posterior. With a minimal model assumption, VaES and VaGES can be applied to the *kernelized Stein discrepancy* (KSD) [22] and *score matching* (SM)-based methods [13, 34, 20] to learn EBLVMs. Besides, VaES can also be used to estimate the *exact Fisher divergence* between the data and general EBLVMs.

2 Method

In this paper, we want to estimate the score function and its gradient w.r.t. the model parameters in an EBLVM. Firstly recall [33] that the score function of an EBLVM can be expressed as

$$\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) = \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} \left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h}) \right], \tag{1}$$

where $p_{\boldsymbol{\theta}}(\boldsymbol{v}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h}) \mathrm{d}\boldsymbol{h}$ is the marginal probability density and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ is the posterior probability density. We then show that the gradient of the score function w.r.t. the model parameters can be decomposed into a term of covariance and a term of expectation:

$$\begin{aligned} \frac{\partial \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial \boldsymbol{\theta}} =& \mathrm{Cov}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})) \\ &+ \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} \left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} \right], \end{aligned} \tag{2}$$

where the covariance term is the covariance between two random vectors $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})$, $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})$ with $\boldsymbol{h} \sim p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ and the expectation term is the expectation of a random matrix $\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})}{\partial \boldsymbol{\theta}}$ with $\boldsymbol{h} \sim p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$. A proof of Eqn. (2) is provided in Appendix A. For simplicity, we denote the covariance term and the expectation term as $c(\boldsymbol{v}; \boldsymbol{\theta})$ and $e(\boldsymbol{v}; \boldsymbol{\theta})$ respectively.

Eqn. (1,2) inspire us with a Monte Carlo estimation of $\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})$ and $\frac{\partial \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial \boldsymbol{\theta}}$ in an EBLVM. However, it needs samples from $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$, which requires the Markov chain Monte Carlo (MCMC) method [23] when only the energy function is accessible, leading to a high time cost. To efficiently get samples, we learn a variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ as an approximation of $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ by

$$\min_{\boldsymbol{\phi}} \mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) || p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})), \tag{3}$$

where $\mathcal{D}$ is the KL divergence when $\boldsymbol{h}$ is discrete and the Fisher divergence when $\boldsymbol{h}$ is continuous and the reasons will be specified in Sec. 2.2. In practice, we often need to estimate Eqn. (1) and Eqn. (2) on a minibatch of data $\boldsymbol{v}_{1:M}$, where M is the batch size. In this case, we don't initialize a new variational parameter $\boldsymbol{\phi}$ for a new minibatch. We only maintain one parameter $\boldsymbol{\phi}$ and update $\boldsymbol{\phi}$ for K times on the minibatch with a prefixed learning rate scheme α by:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \alpha \frac{1}{M} \sum_{i=1}^{M} \nabla_{\boldsymbol{\phi}} \mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}_i) || p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}_i)). \tag{4}$$

2.1 Variational (Gradient) Estimate of the Score Function in EBLVMs

First, we consider estimating the score function. According to Eqn. (1), a naive variational estimate is

$$\frac{1}{L} \sum_{i=1}^{L} \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h}_i), \quad \boldsymbol{h}_i \overset{\text{i.i.d}}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}), \tag{5}$$

where L is the number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. By noticing that

$$\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} \nabla_{\boldsymbol{v}} \log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) = \int q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) \frac{\nabla_{\boldsymbol{v}} q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} \mathrm{d}\boldsymbol{h} = \nabla_{\boldsymbol{v}} \int q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) \mathrm{d}\boldsymbol{h} = \mathtt{0}, \tag{6}$$

we can subtract $r_v \log q_\phi(h|v)$ from the naive estimator without changing the value of the expectation, and the resulting variational estimate of the score function (VaES) is

$$\text{VaES}(v;\theta;\phi) = \frac{1}{L}\sum_{i=1}^{L} r_v \log \frac{p_\theta(v,h_i)}{q_\phi(h_i|v)}, \quad h_i \overset{\text{i.i.d}}{\sim} q_\phi(h|v). \tag{7}$$

When the variational posterior $q_\phi(h|v)$ is equal to the true posterior $p_\theta(h|v)$, $\text{VaES}(v;\theta;\phi) = r_v \log \frac{p_\theta(v,h)}{p_\theta(h|v)} = r_v \log p_\theta(v)$ is a deterministic variable and is exactly equal to the score function.

Then, we consider estimating the gradient of the score function in Eqn. (2). As for the covariance term $c(v;\theta)$, we estimate it with the sample covariance matrix [4], and the resulting variational estimate is

$$\begin{aligned}\hat{c}(v;\theta;\phi) = &\frac{1}{L-1}\sum_{i=1}^{L} r_v \log p_\theta(v,h_i)\frac{\partial \log p_\theta(v,h_i)}{\partial \theta} \\ &- \frac{1}{(L-1)L}\sum_{i=1}^{L} r_v \log p_\theta(v,h_i) \sum_{i=1}^{L} \frac{\partial \log p_\theta(v,h_i)}{\partial \theta}, \quad h_i \overset{\text{i.i.d}}{\sim} q_\phi(h|v).\end{aligned} \tag{8}$$

As for the the expectation term $e(v;\theta)$, we directly apply the Monte Carlo estimate and the resulting variational estimate is

$$\hat{e}(v;\theta;\phi) = \frac{1}{L}\sum_{i=1}^{L} \frac{\partial r_v \log p_\theta(v,h_i)}{\partial \theta}, \quad h_i \overset{\text{i.i.d}}{\sim} q_\phi(h|v). \tag{9}$$

Combing Eqn. (8) and Eqn. (9), the resulting variational gradient estimate of the score function (VaGES) is

$$\text{VaGES}(v;\theta;\phi) = \hat{c}(v;\theta;\phi) + \hat{e}(v;\theta;\phi). \tag{10}$$

Remark: Although Eqn. (9) includes second derivatives, in practice we only need to estimate the product of the second derivatives with vectors and only two backpropagations are required for calculation (since $z^\top \frac{\partial r_v \log p_\theta(v,h_i)}{\partial \theta} = \frac{\partial z^\top r_v \log p_\theta(v,h_i)}{\partial \theta}$).

2.2 Bounding the Bias

Notice that $\text{VaES}(v;\theta;\phi)$ and $\text{VaGES}(v;\theta;\phi)$ are actually estimating $\mathbb{E}_{q_\phi(h|v)}[r_v \log p_\theta(v,h)]$ and

$$\text{Cov}_{q_\phi(h|v)}(r_v \log p_\theta(v,h), r \log p_\theta(v,h)) + \mathbb{E}_{q_\phi(h|v)}\left[\frac{\partial r_v \log p_\theta(v,h)}{\partial \theta}\right]$$

respectively and will introduce some bias. Firstly, we show that when h is discrete, the bias of $\text{VaES}(v;\theta;\phi)$ and $\text{VaGES}(v;\theta;\phi)$ can be bounded by the square root of the KL divergence between $q_\phi(h|v)$ and $p_\theta(h|v)$ under some assumptions on boundedness, as characterized in Theorem 1 and Theorem 2. So in Eqn. (3), we choose S as the KL divergence when h is discrete to learn the variational posterior $q_\phi(h|v)$.

Theorem 1. (h is discrete, VaES, proof in Appendix B) Suppose $r_v \log p_\theta(v,h)$ is bounded w.r.t. v, h and θ, then the bias of $\text{VaES}(v;\theta;\phi)$ can be bounded by the square root of the KL divergence between $q_\phi(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Theorem 2. (h is discrete, VaGES, proof in Appendix B) Suppose $r_v \log p_\theta(v,h)$, $r \log p_\theta(v,h)$ and $\frac{\partial r_v \log p_\theta(v,h)}{\partial \theta}$ are bounded w.r.t. v, h and θ, then the bias of $\text{VaGES}(v;\theta;\phi)$ can be bounded by the square root of the KL divergence between $q_\phi(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Then, we show that under extra assumptions on the Stein regularity (see Def. 2) and boundedness of the Stein factors (see Def. 3), the bias of $\text{VaES}(v;\theta;\phi)$ and $\text{VaGES}(v;\theta;\phi)$ can be bounded by the square root of the Fisher divergence between $q_\phi(h|v)$ and $p_\theta(h|v)$, as characterized in Theorem 3 and Theorem 4. Although the boundedness of the Stein factors have only been verified under some simple cases [18], and haven't been extended to more complex cases, e.g., $p_\theta(v,h)$ is parameterized by a neural network, Theorem 3 and Theorem 4 still inspire us to choose the Fisher divergence in Eqn. (3) when h is continuous to learn the variational posterior $q_\phi(h|v)$.

De nition 1. [18] Suppose p is a probability density de ned on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a function, we de ne g_f^p as a solution of the Stein equation $S_p g = f - E_p f$, where $g : \mathbb{R}^n \to \mathbb{R}^n$ and $S_p g(v) := \nabla_v \log p(v)^\top g(v) + \mathrm{Tr}(\nabla_v g(v))$.

Remark. The solution of the Stein equation exists. For example, let $h = f - E_p f$, then

$$g_1(v) = \frac{1}{p(v)} \int_{-\infty}^{v_1} p(t, v_2, \ldots, v_n) h(t, v_2, \ldots, v_n) dt, \quad g_2(v) = \cdots = g_n(v) = 0$$

is a solution.

De nition 2. Suppose p, q are probability densities de ned on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ is a function, we say f satis es the Stein regular condition w.r.t. p, q iff $\forall i \in \mathbb{Z} \cap [1, m], \lim_{\|v\| \to \infty} q(v) g_{f_i}^p(v) = 0$.

De nition 3. [18] Suppose p, q are probability densities de ned on $\mathbb{R}^n$ and $f_S : \mathbb{R}^n \to \mathbb{R}^m$ is a function satisfying satis es the Stein regular condition w.r.t. p, q, we de ne $\mathcal{F}_f^{p;q} := E_{q(x)} \sum_{i=1}^m \|g_{f_i}^p(x)\|_2^2$, referred to as the Stein factor of f w.r.t. p, q.

Theorem 3. (h is continuous, VaES, proof in Appendix C) Suppose (1) $s(v, \epsilon)$, $\nabla_v \log p(v, h)$ as a function of h satis es the Stein regular condition w.r.t. $p(h|v, \theta)$ and $q(h|v, \phi)$ and (2) the Stein factor of $\nabla_v \log p(v, h)$ as a function of h w.r.t. $p(h|v, \theta), q(h|v, \phi)$ is bounded w.r.t. v, θ and ϕ, then the bias of $\mathrm{VaES}(v, \theta, \phi)$ can be bounded by the square root of the Fisher divergence between $q_\phi(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Theorem 4. (h is continuous, VaGES, proof in Appendix C) Suppose (1) $s(v, \epsilon)$, $\nabla_v \log p(v, h)$, $\nabla_\theta \log p(v, h)$, $\nabla_v \log p(v, h) \frac{\partial \log p(v,h)}{\partial \theta}$ and $\frac{\partial \nabla_v \log p(v,h)}{\partial \theta}$ as functions of h satisfy the Stein regular condition w.r.t. $p(h|v, \theta)$ and $q(h|v, \phi)$ and (2) the Stein factors of $\nabla_v \log p(v, h)$, $\nabla_\theta \log p(v, h)$, $\nabla_v \log p(v, h) \frac{\partial \log p(v,h)}{\partial \theta}$ and $\frac{\partial \nabla_v \log p(v,h)}{\partial \theta}$ as functions of h w.r.t. $p(h|v, \theta), q(h|v, \phi)$ are bounded w.r.t. v, θ and ϕ, (3) $\nabla_v \log p(v, h)$ and $\nabla_\theta \log p(v, h)$ are bounded w.r.t. v, h and θ, then the bias of $\mathrm{VaGES}(v, \theta, \phi)$ can be bounded by the square root of the Fisher divergence between $q_\phi(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

3 Learning EBLVMs

In this section, we show that VaES and VaGES can extend Kernelized Stein Discrepancy (KSD) [22] and methods [3, 4, 20] based on score matching (SM) [13] to learn nonstructural EBLVMs. Besides, we empirically validate our methods and further details of experimental settings can be found in Appendix E.

3.1 Learning EBLVMs with KSD

Method. The KSD between a target probability density $p_t(v)$ and the model probability density $p_\theta(v)$ is de ned as

$$\begin{aligned} \mathrm{KSD}(p_t, p_\theta) = E_{v, v' \sim p_t(v)} \big[& \nabla_v \log p_\theta(v)^\top k(v, v') \nabla_{v'} \log p_\theta(v') + \mathrm{Tr}(\nabla_v \nabla_{v'} k(v, v')) \\ & + \nabla_v k(v, v')^\top \nabla_{v'} \log p_\theta(v') + \nabla_{v'} k(v, v')^\top \nabla_v \log p_\theta(v) \big], \end{aligned} \tag{11}$$

which properly measures the difference between p_t and p_θ under some mild assumptions [22]. To learn an EBLVM from the target density, we can use gradient-based optimization to minimize $\mathrm{KSD}(p_t, p_\theta)$, where the gradient w.r.t. θ is

$$\frac{\partial \mathrm{KSD}(p_t, p_\theta)}{\partial \theta} = 2 E_{v, v' \sim p_t(v)} \Big[k(v, v') \nabla_v \log p_\theta(v) + \nabla_v k(v, v')^\top \frac{\partial \nabla_{v'} \log p_\theta(v')}{\partial \theta} \Big]. \tag{12}$$

Estimating $\nabla_v \log p_\theta(v)$ with VaES and estimating $\frac{\partial \nabla_{v'} \log p_\theta(v')}{\partial \theta}$ with VaGES, the variational stochastic gradient estimate of KSD (VaGES-KSD) is

$$\frac{1}{M} \sum_{i=1}^M \big[k(v_i, v_i') \mathrm{VaES}(v_i, \theta, \phi) + \nabla_v k(v_i, v_i')^\top \mathrm{VaGES}(v_i', \theta, \phi) \big], \tag{13}$$

where the union of $v_{1:M}$ and $v_{1:M}'$ is a minibatch from the target density, and $\mathrm{VaES}(v_i, \theta, \phi)$ and $\mathrm{VaGES}(v_i', \theta, \phi)$ are independent.

Model and Setting. To the best of our knowledge, KSD hasn't been shown feasible to scale up to natural images. So we illustrate the validity of VaGES-KSD by learning Gaussian restricted Boltzmann machines (GRBMs) [35, 10] on the 2-D checkerboard dataset, whose density is shown in Appendix F.1. The energy function of a GRBM is

$$E(v;h;\ ) = \frac{1}{2^{\ 2}}jjv \quad bjj^2 \quad c^{>}h \quad \frac{1}{\ }v^{>}Wh; \tag{14}$$

with learnable parameters $= (\ ;W;b;c)$. Since GRBM has a tractable posterior, we can directly learn it using KSD (see Eqn. (12)). We also compare another baseline where $p_{\ }(v)$ is estimated by importance sampling (i.e., $p_{\ }(v) \quad \frac{1}{L}\sum_{i=1}^{L}\frac{p_{\ }(v;h_i)}{\mathrm{unif}(h_i)}; h_i \overset{\text{i.i.d}}{\sim} \mathrm{unif}(h)$, where unif means the uniform distribution) and we call it IS-KSD. We generate 60,000 samples for training and 10,000 samples for testing. We use the RBF kernel $k(v;v^0) = \exp(\ \frac{jjv \quad v^0jj_2^2}{2^{\ 2}})$ and $= 0:1$.

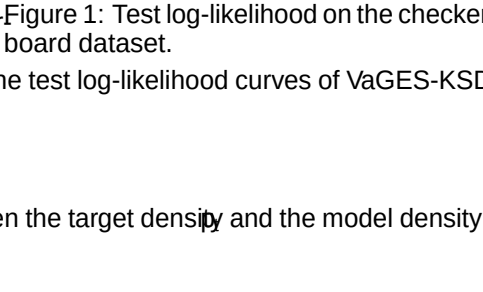

Figure 1: Test log-likelihood on the checkerboard dataset.

Result. As shown in Fig. 1, VaGES-KSD outperforms IS-KSD on all L (i.e., the number of sampled h for each v) and is comparable to the KSD baseline (the test log-likelihood curves of VaGES-KSD are very close to KSD when $L = 5; 10$).

3.2 Learning EBLVMs with Score Matching

Method. Recall that the Fisher divergence [15] between the target density p_t and the model density p is

$$D_F(p_t(v)jjp\ (v)) \ , \ \frac{1}{2}E_{p_t(v)}\ jjr_{\ v}\log p\ (v) \quad r_{\ v}\log p_t(v)jj_2^2\ ; \tag{15}$$

which measures the difference between the score functions of p_t and p . The denoising score matching (DSM) [34] minimizes the Fisher divergence between a perturbed target density and the model density:

$$J_{DSM}(\ ) \ , \ D_F(p_{\ 0}(v)jjp\ (v)) \quad \frac{1}{2}E_{p_t(w)p_{\ 0}(vjw)}jjr_{\ v}\log p_{\ }(v) \quad r_{\ v}\log p_{\ 0}(vjw)jj_2^2; \tag{16}$$

where $p_{\ 0}(v) = \int p_t(w)p_{\ 0}(vjw)dw$ is the perturbed target density, $p_{\ 0}(vjw) = N(vjw;\ _0^2I)$ is the Gaussian perturbation and $_0$ is a fixed noise level. The multiscale denoising score matching (MDSM) [20] is a variant of DSM which uses different levels of noise to learn an EBM in high-dimensional spaces:

$$J_{MDSM}(\ ) \ , \ \frac{1}{2}E_{p_t(w)p(\ )p\ (vjw)}jjr_{\ v}\log p_{\ }(v) \quad r_{\ v}\log p_{\ 0}(vjw)jj_2^2; \tag{17}$$

where $p(\ )$ is a prior distribution over the flexible noise level $$ and $_0$ is a fixed noise level.

We extend the above two score matching methods (Eqn. (16) and Eqn. (17)) to learn EBLVMs. Firstly write Eqn. (16) and Eqn. (17) in a general form for simplicity

$$J(\ ) = \frac{1}{2}E_{p_t(w;v)}\ jjr_{\ v}\log p_{\ }(v) \quad r_{\ v}\log p_{\ 0}(vjw)jj_2^2\ ; \tag{18}$$

where $p_t(w;v)$ is the joint distribution of w and v (specifically, $p_t(w;v) = p_t(w)p_{\ 0}(vjw)$ for Eqn. (16) and $p_t(w;v) = \int p_t(w)p(\ )p\ (vjw)d$ for Eqn. (17)). We use gradient-based optimization to minimize $J(\ )$ and its gradient w.r.t. $$ is

$$\frac{J(\ )}{@\ } = E_{p_t(w;v)}\ (r_{\ v}\log p_{\ }(v) \quad r_{\ v}\log p_{\ 0}(vjw))\frac{@r_{\ v}\log p_{\ }(v)}{@\ }\ : \tag{19}$$

Estimating $\nabla_v \log p_\theta(v)$ with VaES and estimating $\frac{\partial \nabla_{v^0} \log p_\theta(v^0)}{\partial}$ with VaGES, the variational stochastic gradient estimate of the score matching methods (VaGES-SM) is

$$\frac{1}{M}\sum_{i=1}^{M}(\text{VaES}(v_i;\ ;\ ) \quad \nabla_v \log p_{\ 0}(v_i|w_i))\text{VaGES}(v_i;\ ;\ ); \tag{20}$$

where $(w_{1:M}; v_{1:M})$ is a minibatch from $p_t(w; v)$, and $\text{VaES}(v_i;\ ;\ )$ and $\text{VaGES}(v_i^0;\ ;\ )$ are independent. We explicitly denote our methods as VaGES-DSM or VaGES-MDSM according to which score matching objective is used.

Result. To validate the effectiveness of VaGES-SM, we compare VaGES-SM with the state-of-art bi-level score matching (BiSM) [1] for learning EBLVMs, which reformulates Eqn. (16) and Eqn. (17) as bi-level optimization problems (BiDSM and BiMDSM) to learn EBLVMs and uses the gradient unrolling technique to solve the bi-level problems. We compare VaGES-DSM and BiDSM under different settings in GRBMs (see Eqn. (14)) on the Frey face dataset in Fig. 2. For BiSM, M means the number of h sampled for each v; K means the times of updating parameters in the lower level problem given a minibatch; N means the number of steps for gradient unrolling in the higher level problem given a minibatch. Since the meaning of L and K in BiSM is similar to VaGES-SM, we compare them in the same dimension and leave N in another dimension when ploting Fig. 2 (a-d). VaGES-DSM outperforms BiDSM ($N=0$ or $N=2$) and has similar performance with BiDSM ($N=5$). Meanwhile it requires the least memory and the time consuming is similar to BiDSM ($N=2$).

(a) Log-likelihood (b) Fisher divergence (c) Time (s) (d) Memory (MB)

Figure 2: The comparison between VaGES-DSM and BiDSM in GRBMs on the Frey face dataset. The log-likelihood and the Fisher divergence (subtracted by the same unknown constant only relevant to the data) are evaluated on the testing dataset according to the best validation performance. The time is the training time of 2,000 iterations. VaGES-DSM have similar performance with BiDSM ($N=5$), and meanwhile requires less time and memory.

Besides, we compare VaGES-MDSM with BiMDSM in a deep EBLVM with energy function $E(v; h;\ ) = g_3(g_2(g_1(v;\ _1); h);\ _2)$, where g_1 and g_3 are two neural networks, g_2 is an additive coupling layer and h is continuous. We evaluate them on the MNIST dataset. The BiMDSM has better visual quality then VaGES-MDSM. It indicates that there are still challenges to apply VaGES-MDSM to high-dimensional data, which is analyzed in Sec. 5.

(a) BiMDSM (b) VaGES-MDSM ($L=2$) (c) VaGES-MDSM ($L=200$)

Figure 3: Samples from deep EBLVMs trained by BiMDSM and VaGES-MDSM on the MNIST.

4 Evaluating EBLVMs

Method. In this setting, we are given an EBLVM $p(v; h) = e^{E(v;h;\ )}=Z(\ )$ and a set of samples $\{v_i\}_{i=1}^n$ from the target density $p_t(v)$. We want to measure how well the model $p(v)$ approximates

the target density $p_t(v)$, which needs an absolute value representing the difference between p_t and p_θ. This task is more difficult than model comparison, which only needs relative values to compare between models. We use the Fisher divergence of the maximum in [12, 6] to measure how well p_θ approximates p_t:

$$D_F^m(p_t(v)\|p_\theta(v)) = \max_{f\in\mathcal{F}} \mathbb{E}_{p_t(v)}\mathbb{E}_{p(\epsilon)}\left[2\nabla_v \log p_\theta(v)^\top f(v) + 2\nabla_v^\top f(v) - \frac{1}{2}\|f(v)\|_2^2\right], \quad (21)$$

where $\mathcal{F}$ is the set of functions $f : \mathbb{R}^d \to \mathbb{R}^d$, s.t. $\lim_{\|v\|\to\infty} p_t(v)f(v) = 0$ and $p(\epsilon)$ is a noise distribution (e.g., Gaussian distribution) introduced for computation efficiency [6]. Under some mild assumptions (see Appendix D), we have $D_F^m(p_t\|p_\theta) = D_F(p_t\|p_\theta)$. In practice, $\mathcal{F}$ is approximated by a neural network $\{f : \phi \in \Phi\}$, where ϕ is the parameter and Φ is the parameter space. We optimize the right hand side of Eqn. (21) and the gradient w.r.t. θ is

$$\mathbb{E}_{p_t(v)}\mathbb{E}_{p(\epsilon)}\left[2\nabla_v \log p_\theta(v)^\top \frac{\partial f_\phi(v)}{\partial \theta} + 2\frac{\partial^\top \nabla_v f(v)}{\partial \theta} - \frac{1}{2}\frac{\partial \|f_\phi(v)\|_2^2}{\partial \theta}\right].$$

Estimating $\nabla_v \log p_\theta(v)$ with VaES, the variational stochastic gradient estimate is

$$\frac{1}{M}\sum_{i=1}^{M} \mathrm{VaES}(v_i; \epsilon, \phi)^\top \frac{\partial f_\phi(v_i)}{\partial \theta} + 2\frac{\partial^\top \nabla_{v_i} f(v_i)\,\epsilon_i}{\partial \theta} - \frac{1}{2}\frac{\partial \|f_\phi(v_i)\|_2^2}{\partial \theta},$$

where $v_{1:M}$ is a minibatch from $p_t(v)$ and $\epsilon_{1:M}$ is a minibatch of noise from $p(\epsilon)$. We refer to this method as VaGES-Fisher. In practice, Eqn. (21) is optimized on the training data, validated on the validation data (if there is one) and tested on the testing data.

Result. We validate the effectiveness of VaGES-Fisher in GRBMs. We initialize a GRBM as $p_t(v)$, perturb its weight with increasing noise and provide the Fisher divergence between the initial GRBM and the perturbed one. The dimensions of v and h are same and we experiment on dimensions of 200 and 500. Further details of experimental settings can be found in Appendix E. The result is shown in Fig. 4. We compare our estimated Fisher divergence with the accurate Fisher divergence. Under both dimensions, our estimated one is close to the accurate one.

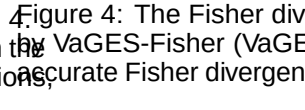

Figure 4: The Fisher divergence estimated by VaGES-Fisher (VaGES-Fisher) v.s. the accurate Fisher divergence (Fisher).

5 Challenges in High-Dimensional Spaces and Deep EBLVMs

Although our methods achieve promising results when the data is simple (e.g., the checkerboard dataset) or when the model is simple (e.g., GRBMs), there are still challenges in high-dimensional spaces or in deep EBLVMs. Consider the VaGES estimator, it's a matrix of size $\dim(v) \times \dim(\epsilon)$. So either more complex data or models will increase the size of the matrix, making it harder to estimate. In our estimate, the bias might matter more than the variance, since increasing L doesn't increase the visual quality of samples, as shown in Fig. 3. It is an initial hypothesis and requires further experiments for validation. Besides, VaGES is based on the sample covariance matrix, which is relatively simple and there are might be advanced substitutes with better properties.

References

[1] Fan Bao, Chongxuan Li, Kun Xu, Hang Su, Jun Zhu, and Bo Zhang. Bi-level score matching for learning energy-based latent variable models. arXiv preprint arXiv:2010.07856, 2020.

[2] Chao Du, Kun Xu, Chongxuan Li, Jun Zhu, and Bo Zhang. Learning implicit generative models by teaching explicit ones. arXiv preprint arXiv:1807.03870, 2018.

[3] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. arXiv preprint arXiv:1903.08689, 2019.

[4] Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices, 2016.

[5] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7518–7528, 2020.

[6] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling.

[7] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classi er is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263, 2019.

[8] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Arti cial Intelligence and Statistics, pages 297–304, 2010.

[9] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8):1771–1800, 2002.

[10] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. science, 313(5786):504–507, 2006.

[11] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. Neural computation, 18(7):1527–1554, 2006.

[12] Tianyang Hu, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng. Stein neural sampler. arXiv preprint arXiv:1810.03545, 2018.

[13] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6(Apr):695–709, 2005.

[14] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016.

[15] Oliver Thomas Johnson. Information theory and the central limit theorem. World Scienti c, 2004.

[16] Volodymyr Kuleshov and Stefano Ermon. Neural variational inference and learning in undirected graphical models. In Advances in Neural Information Processing Systems, pages 6734–6743, 2017.

[17] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006.

[18] Christophe Ley, Yvik Swan, et al. Stein's density approach and information inequalities. Electronic Communications in Probability, 18, 2013.

[19] Chongxuan Li, Chao Du, Kun Xu, Max Welling, Jun Zhu, and Bo Zhang. To relieve your headache of training an mrf, take advil. arXiv preprint arXiv:1901.08400, 2019.

[20] Zengyi Li, Yubei Chen, and Friedrich T Sommer. Annealed denoising score matching: Learning energy-based models in high-dimensional spaces. arXiv preprint arXiv:1910.07762, 2019.

[21] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In Advances in neural information processing systems, pages 2378–2386, 2016.

[22] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of- t tests. In International conference on machine learning, pages 276–284, 2016.

[23] Radford M. Neal. Probabilistic inference using markov chain monte carlo methods, 1993.

[24] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. arXiv preprint arXiv:1903.12370, 2019.

[25] Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. On learning non-convergent short-run mcmc toward energy-based model. arXiv preprint arXiv:1904.09770, 2019.

[26] Benjamin Rhodes and Michael U. Gutmann. Variational noise-contrastive estimation. In Kamalika Chaudhuri and Masashi Sugiyama, editors, Proceedings of Machine Learning Research, volume 89 of Proceedings of Machine Learning Research, pages 2741–2750. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/rhodes19a.html .

[27] Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann machines. In Proceedings of the twelfth international conference on arti cial intelligence and statistics, 2009.

[28] Ruslan Salakhutdinov and Hugo Larochelle. Ef cient learning of deep Boltzmann machines. In Proceedings of the thirteenth international conference on arti cial intelligence and statistics, pages 693–700, 2010.

[29] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. arXiv preprint arXiv:1905.07088, 2019.

[30] Kevin Swersky, Marc'Aurelio Ranzato, David Buchman, Nando D Freitas, and Benjamin M Marlin. On autoencoders and score matching for energy based models. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 1201–1208, 2011.

[31] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In Proceedings of the 25th international conference on Machine learning, pages 1064–1071. ACM, 2008.

[32] Alexandre B Tsybakov. Introduction to nonparametric estimation. Springer Science & Business Media, 2008.

[33] Eszter Vértes, UCL Gatsby Unit, and Maneesh Sahani. Learning doubly intractable latent variable models via score matching. http://approximateinference.org/2016/accepted/VertesSahani2016.pdf, 2016.

[34] Pascal Vincent. A connection between score matching and denoising autoencoders. Neural computation, 23(7):1661–1674, 2011.

[35] Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. In Advances in neural information processing systems, pages 1481–1488, 2005.

[36] Qiu Yixuan, Zhang Lingsong, and Wang Xiao. Unbiased contrastive divergence algorithm for training energy-based latent variable models. In International Conference on Learning Representations, 2020.

A Proof of Equation (3)

Proof. Firstly we have

$$E_{p(h|v;\theta)}[\nabla_\theta \log p(h|v;\theta)] = E_{p(h|v;\theta)}\left[\frac{\nabla_\theta p(h|v;\theta)}{p(h|v;\theta)}\right] = \int_h p(h|v;\theta)\frac{\nabla_\theta p(h|v;\theta)}{p(h|v;\theta)}dh$$
$$= \nabla_\theta \int_h p(h|v;\theta)dh = \nabla_\theta 1 = 0; \quad (22)$$

and similarly we have $E_{p(h|v;\theta)}[\nabla_\theta \log p(h|v;\theta)] = 0$. Thereby, we have

$$\nabla_\theta \log p(v;\theta) = \nabla_\theta \log p(v;\theta) + E_{p(h|v;\theta)}[\nabla_\theta \log p(h|v;\theta)] = E_{p(h|v;\theta)}[\nabla_\theta \log p(v,h;\theta)];$$

and similarly we have $\nabla_\theta \log p(v;\theta) = E_{p(h|v;\theta)}[\nabla_\theta \log p(v,h;\theta)]$.

Taking derivatives to Eqn. (22) w.r.t., we have

$$E_{p(h|v;\theta)}\left[\nabla_\theta \log p(h|v;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right] + E_{p(h|v;\theta)}\left[\frac{\partial \nabla_\theta \log p(h|v;\theta)}{\partial}\right] = 0. \quad (23)$$

The leftmost term of Eqn. (23) can be written as

$$E_{p(h|v;\theta)}\left[\nabla_\theta \log p(h|v;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right]$$
$$= E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v,h;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right] - E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right]$$
$$= E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v,h;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right]$$
$$= E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v,h;\theta)\frac{\partial \log p(v,h;\theta)}{\partial}\right] - E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v,h;\theta)\frac{\partial \log p(v;\theta)}{\partial}\right]$$
$$= E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v,h;\theta)\frac{\partial \log p(v,h;\theta)}{\partial}\right] - E_{p(h|v;\theta)}[\nabla_\theta \log p(v,h;\theta)]\,E_{p(h|v;\theta)}\left[\frac{\partial \log p(v;\theta)}{\partial}\right]$$
$$= E_{p(h|v;\theta)}\left[\nabla_\theta \log p(v,h;\theta)\frac{\partial \log p(v,h;\theta)}{\partial}\right]$$
$$E_{p(h|v;\theta)}[\nabla_\theta \log p(v,h;\theta)]\,E_{p(h|v;\theta)}\left[\frac{\partial \log p(v,h;\theta)}{\partial}\right]$$
$$= \mathrm{Cov}_{p(h|v;\theta)}(\nabla_\theta \log p(v,h;\theta), \nabla_\theta \log p(v,h;\theta)).$$

Thereby, we have

$$\frac{\partial \nabla_\theta \log p(v;\theta)}{\partial} = \frac{\partial \nabla_\theta \log p(v;\theta)}{\partial} + E_{p(h|v;\theta)}\left[\nabla_\theta \log p(h|v;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right]$$
$$+ E_{p(h|v;\theta)}\left[\frac{\partial \nabla_\theta \log p(h|v;\theta)}{\partial}\right]$$
$$= E_{p(h|v;\theta)}\left[\nabla_\theta \log p(h|v;\theta)\frac{\partial \log p(h|v;\theta)}{\partial}\right]$$
$$+ E_{p(h|v;\theta)}\left[\frac{\partial \nabla_\theta \log p(v,h;\theta)}{\partial}\right]$$
$$= \mathrm{Cov}_{p(h|v;\theta)}(\nabla_\theta \log p(v,h;\theta), \nabla_\theta \log p(v,h;\theta))$$
$$+ E_{p(h|v;\theta)}\left[\frac{\partial \nabla_\theta \log p(v,h;\theta)}{\partial}\right];$$

□

B Proof of Theorem 1 and Theorem 2

Lemma 1. Suppose p, q are probability mass functions defined on Ω and $f : \Omega \to \mathbb{R}^m$, then the approximation error of estimating $\mathbb{E}_p f$ using $\mathbb{E}_q f$ can be bounded as

$$\|\mathbb{E}_p f - \mathbb{E}_q f\|_2 \le \|f\|_1 \sqrt{2 D_{KL}(q \| p)},$$

where $\|f\|_1 := \sup_{\omega \in \Omega} \|f(\omega)\|_2$.

Proof. Firstly, we have

$$\|\mathbb{E}_p f - \mathbb{E}_q f\|_2 = \Big\| \sum_{\omega \in \Omega} (p(\omega) - q(\omega)) f(\omega) \Big\|_2 \le \sum_{\omega \in \Omega} |p(\omega) - q(\omega)| \, \|f(\omega)\|_2 \le \sum_{\omega \in \Omega} |p(\omega) - q(\omega)| \, \|f\|_1 .$$

According to Pinsker's inequality (3.2.2), we have $\sum_{\omega \in \Omega} |p(\omega) - q(\omega)| \le \sqrt{2 D_{KL}(p \| q)}$. Thereby, $\|\mathbb{E}_p f - \mathbb{E}_q f\|_2 \le \sqrt{2 D_{KL}(p \| q)} \|f\|_1$. □

Theorem 1. (h is discrete, VaES) Suppose $r_v \log p_\theta(v; h)$ is bounded w.r.t. $v; h$ and θ, then the bias of $\mathrm{VaES}(v; \theta; \cdot)$ can be bounded by the square root of the KL divergence between $q_\theta(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Proof. According to Lemma 1, we have

$$\|\mathbb{E}_{q_\theta(h|v)}[r_v \log p_\theta(v; h)] - \mathbb{E}_{p_\theta(h|v)}[r_v \log p_\theta(v; h)]\|_2 \le \sup_h \|r_v \log p_\theta(v; h)\|_2 \sqrt{2 D_{KL}(q_\theta(h|v) \| p_\theta(h|v))}.$$

By the boundedness of $r_v \log p_\theta(v; h)$, $\exists A < \infty, \forall v, \forall h, \forall \theta, \|r_v \log p_\theta(v; h)\|_2 \le A$. Let $C = \sqrt{2} A$, then

$$\begin{aligned} &\|\mathbb{E}_{q_\theta(h|v)}[r_v \log p_\theta(v; h)] - r_v \log p_\theta(v)\|_2 \\ =\ &\|\mathbb{E}_{q_\theta(h|v)}[r_v \log p_\theta(v; h)] - \mathbb{E}_{p_\theta(h|v)}[r_v \log p_\theta(v; h)]\|_2 \le A \sqrt{2 D_{KL}(q_\theta(h|v) \| p_\theta(h|v))} = C \sqrt{D_{KL}(q_\theta(h|v) \| p_\theta(h|v))}. \end{aligned}$$

□

Definition 1. Suppose A is a matrix, we define $\|A\|_2 := \sqrt{\sum_{i,j} A_{i,j}^2}$.

Lemma 2. Suppose a, b are two vectors, then $\|ab^\top\|_2 = \|a\|_2 \|b\|_2$.

Proof. $\|ab^\top\|_2 = \sqrt{\sum_{i,j} a_i^2 b_j^2} = \|a\|_2 \|b\|_2$. □

Theorem 2. (h is discrete, VaGES) Suppose $r_v \log p_\theta(v; h)$, $r_\theta \log p_\theta(v; h)$ and $\frac{\partial r_v \log p_\theta(v; h)}{\partial \theta}$ are bounded w.r.t. $v; h$ and θ, then the bias of $\mathrm{VaGES}(v; \theta; \cdot)$ can be bounded by the square root of the KL divergence between $q_\theta(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Proof. According to Thm. 1, $\exists C_1 < \infty$, s.t.

$$\|\mathbb{E}_{q_\theta(h|v)}[r_v \log p_\theta(v; h)] - r_\theta \log p_\theta(v)\|_2 \le C_1 \sqrt{D_{KL}(q_\theta(h|v) \| p_\theta(h|v))}.$$

Similarly, $\exists C_2 < \infty$, s.t.

$$\|\mathbb{E}_{q_\theta(h|v)}[r_\theta \log p_\theta(v; h)] - r_\theta \log p_\theta(v)\|_2 \le C_2 \sqrt{D_{KL}(q_\theta(h|v) \| p_\theta(h|v))},$$

and $\exists C_3 < 1$, s.t.

$$\|E_{q(h|v)}\frac{\partial_{\theta_v}\log p_\theta(v;h)}{\partial\theta} - E_{p(h|v)}\frac{\partial_{\theta_v}\log p_\theta(v;h)}{\partial\theta}\|_2 \le C_3\sqrt{D_{KL}(q(h|v)\|p(h|v))}.$$

By the boundedness of $\partial_{\theta_v}\log p_\theta(v;h)$ and $\partial_{\theta_h}\log p_\theta(v;h)$, $\partial_{\theta_v}\log p_\theta(v;h)\frac{\partial\log p_\theta(v;h)}{\partial\theta}$ is also bounded. Thereby, $\exists C_4 < 1$, s.t.

$$\|E_{q(h|v)}\partial_{\theta_v}\log p_\theta(v;h)\frac{\partial\log p_\theta(v;h)}{\partial\theta} - E_{p(h|v)}\partial_{\theta_v}\log p_\theta(v;h)\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \le C_4\sqrt{D_{KL}(q(h|v)\|p(h|v))}.$$

By the boundedness of $\partial_{\theta_v}\log p_\theta(v;h)$ and $\partial_{\theta_h}\log p_\theta(v;h)$, we can assume $C < 1$ is a constant that bounds $\|\partial_{\theta_v}\log p_\theta(v;h)\|_2$ and $\|\partial_{\theta_h}\log p_\theta(v;h)\|_2$. Then, by the triangle inequality and Lemma. 2, we have

$$\begin{aligned}
&\|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]E_{q(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta} \\
&\quad - E_{p(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\le \|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]E_{q(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta} - E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\quad + \|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)] - E_{p(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]\|E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&= \|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]\|_2\,\|E_{q(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta} - E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\quad + \|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)] - E_{p(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]\|_2\,\|E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\le C\|E_{q(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta} - E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\quad + C\|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)] - E_{p(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]\|_2 \\
&\le CC_2\sqrt{D_{KL}(q(h|v)\|p(h|v))} + CC_1\sqrt{D_{KL}(q(h|v)\|p(h|v))} \\
&= (CC_2 + CC_1)\sqrt{D_{KL}(q(h|v)\|p(h|v))}.
\end{aligned}$$

Thereby,

$$\begin{aligned}
&\|Cov_{q(h|v)}(\partial_{\theta_v}\log p_\theta(v;h);\partial_{\theta_h}\log p_\theta(v;h)) \\
&\quad - Cov_{p(h|v)}(\partial_{\theta_v}\log p_\theta(v;h);\partial_{\theta_h}\log p_\theta(v;h))\|_2 \\
&\le \|E_{q(h|v)}\partial_{\theta_v}\log p_\theta(v;h)\frac{\partial\log p_\theta(v;h)}{\partial\theta} - E_{p(h|v)}\partial_{\theta_v}\log p_\theta(v;h)\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\quad + \|E_{q(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]E_{q(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta} \\
&\qquad - E_{p(h|v)}[\partial_{\theta_v}\log p_\theta(v;h)]E_{p(h|v)}\frac{\partial\log p_\theta(v;h)}{\partial\theta}\|_2 \\
&\le C_4\sqrt{D_{KL}(q(h|v)\|p(h|v))} + (CC_2 + CC_1)\sqrt{D_{KL}(q(h|v)\|p(h|v))} \\
&= (C_4 + CC_2 + CC_1)\sqrt{D_{KL}(q(h|v)\|p(h|v))}.
\end{aligned}$$

As a result,

$$\begin{aligned}
&\|\mathrm{Cov}_{q(h|v)}(r_v \log p_\theta(v;h), r_\theta \log p_\theta(v;h)) + \mathbb{E}_{q(h|v)}\left[\frac{\partial_v \log p_\theta(v;h)}{\partial_\theta \log p_\theta(v;h)}\right] \\
&\quad \frac{\partial_v \log p_\theta(v;\cdot)}{\partial_\theta}\|_2 \\
=&\|\mathrm{Cov}_{q(h|v)}(r_v \log p_\theta(v;h), r_\theta \log p_\theta(v;h)) + \mathbb{E}_{q(h|v)}\left[\frac{\partial_v \log p_\theta(v;h)}{\partial_\theta \log p_\theta(v;h)}\right] \\
&\quad \mathrm{Cov}_{p(h|v)}(r_v \log p_\theta(v;h), r_\theta \log p_\theta(v;h)) \quad \mathbb{E}_{p(h|v)}\left[\frac{\partial_v \log p_\theta(v;h)}{\partial_\theta \log p_\theta(v;h)}\right]\|_2 \\
\le&\|\mathrm{Cov}_{q(h|v)}(r_v \log p_\theta(v;h), r_\theta \log p_\theta(v;h)) \\
&\quad \mathrm{Cov}_{p(h|v)}(r_v \log p_\theta(v;h), r_\theta \log p_\theta(v;h))\|_2 \\
&+ \|\mathbb{E}_{q(h|v)}\left[\frac{\partial_v \log p_\theta(v;h)}{\partial_\theta}\right] \quad \mathbb{E}_{p(h|v)}\left[\frac{\partial_v \log p_\theta(v;h)}{\partial_\theta}\right]\|_2 \\
\le&(C_4 + CC_2 + CC_1) \sqrt{D_{KL}(q(h|v)\|p(h|v))} + C_3 \sqrt{D_{KL}(q(h|v)\|p(h|v))} \\
=&(C_4 + CC_2 + CC_1 + C_3) \sqrt{D_{KL}(q(h|v)\|p(h|v))}.
\end{aligned}$$

□

C Proof of Theorem 3 and Theorem 4

De nition 2. Suppose p is a probability density on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n$, we de ne $S_p f(v) := r_v \log p(v)^\top f(v) + \mathrm{Tr}(r_v f(v))$.

Lemma 3. [21] Suppose p is a probability density on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a function satisfying $\lim_{\|v\|\to\infty} p(v)f(v) = 0$, then $\mathbb{E}_{p(v)}[S_p f(v)] = 0$.

Proof.

$$\begin{aligned}
0 = \int r_v(p(v)f(v))dv &= \int p(v) r_v f(v) + p(v) f(v) r_v \log p(v)^\top dv \\
&= \mathbb{E}_{p(v)}\left[r_v f(v) + f(v) r_v \log p(v)^\top\right].
\end{aligned}$$

Thereby,

$$\begin{aligned}
0 = \mathrm{Tr}(\mathbb{E}_{p(v)}\left[r_v f(v) + f(v) r_v \log p(v)^\top\right]) &= \mathbb{E}_{p(v)}\left[\mathrm{Tr}(r_v f(v)) + r_v \log p(v)^\top f(v)\right] \\
&= \mathbb{E}_{p(v)}[S_p f(v)].
\end{aligned}$$

□

Lemma 4. Suppose p, q are probability densities on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ satis es $\lim_{\|v\|\to\infty} q(v)f(v) = 0$, we have

$$|\mathbb{E}_q S_p f| \le \sqrt{\mathbb{E}_{q(v)}\|f(v)\|^2} \sqrt{D_F(q\|p)}$$

Proof. By Lemma 3, we have $\mathbb{E}_q S_q f = 0$. Thereby,

$$\begin{aligned}
|\mathbb{E}_q S_p f| = |\mathbb{E}_q S_p f - \mathbb{E}_q S_q f| &= |\mathbb{E}_{q(v)} f(v)^\top (r_v \log p(v) - r_v \log q(v))| \\
&\le \mathbb{E}_{q(v)}\|f(v)\| \|r_v \log p(v) - r_v \log q(v)\| \\
&\le \sqrt{\mathbb{E}_{q(v)}\|f(v)\|^2} \sqrt{\mathbb{E}_{q(v)}\|r_v \log p(v) - r_v \log q(v)\|^2} \\
&= \sqrt{\mathbb{E}_{q(v)}\|f(v)\|^2} \sqrt{D_F(q\|p)}.
\end{aligned}$$

□

De nition 3. [18] Suppose p is a probability density de ned on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a function, we de ne g_f^p as a solution of the Stein equation $S_p g = f - E_p f$.

Remark. The solution of the Stein equation exists. For example, let $h = f - E_p f$, then

$$g_1(v) = \frac{1}{p(v)} \int_{-\infty}^{v_1} p(t, v_2, \cdots, v_n) h(t, v_2, \cdots, v_n) dt, \quad g_2(v) = \cdots = g_n(v) = 0$$

is a solution.

De nition 4. Suppose p, q are probability densities de ned on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ is a function, we say f satis es the Stein regular condition w.r.t. p, q iff $\forall i \in \mathbb{Z} \cap [1, m], \lim_{\|v\| \to \infty} q(v) g_{f_i}^p(v) = 0$.

De nition 5. [18] Suppose p, q are probability densities de ned on $\mathbb{R}^n$ and $f_S : \mathbb{R}^n \to \mathbb{R}^m$ is a function satisfying satis es the Stein regular condition w.r.t. p, q, we de ne $\Gamma_f^{p,q} := \sqrt{E_{q(x)} \sum_{i=1}^m \|g_{f_i}^p(x)\|_2^2}$, referred to as the Stein factor of f w.r.t. p, q.

Lemma 5. Suppose p, q are probability densities de ned on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ is a function satisfying the Stein regular condition w.r.t. p, q, then the approximation error of estimating $E_p f$ using $E_q f$ can be bounded as

$$\|E_q f - E_p f\|_2 \leq \Gamma_f^{p,q} \sqrt{D_F(q \| p)}.$$

Proof. By Lemma 4, we have

$$|E_q f_i - E_p f_i| = |E_q(f_i - E_p f_i)| = |E_q S_p g_{f_i}^p| \leq \sqrt{E_{q(v)} \|g_{f_i}^p(v)\|_2^2} \sqrt{D_F(q \| p)}.$$

Thereby, we have

$$\|E_q f - E_p f\|_2 = \sqrt{\sum_{i=1}^m |E_q f_i - E_p f_i|^2} \leq \sqrt{\sum_{i=1}^m E_{q(v)} \|g_{f_i}^p(v)\|_2^2} \sqrt{D_F(q \| p)} = \Gamma_f^{p,q} \sqrt{D_F(q \| p)},$$

where $\Gamma_f^{p,q} := \sqrt{E_{q(v)} \sum_{i=1}^m \|g_{f_i}^p(v)\|_2^2}$. □

Theorem 3. (h is continuous, VaES) Suppose (1) $\forall (v, \cdots, \cdots)$, $r_v \log p_\theta(v; h)$ as a function of h satis es the Stein regular condition w.r.t. $p(h|v, \cdots, \cdots)$ and $q(h|v, \cdots, \cdots)$ and (2) the Stein factor of $r_v \log p_\theta(v; h)$ as a function of h w.r.t. $p(h|v, \cdots, \cdots), q(h|v, \cdots, \cdots)$ is bounded w.r.t. $v, \cdots$ and $\cdots$, then the bias of $\mathrm{VaES}(v, \cdots, \cdots)$ can be bounded by the square root of the Fisher divergence between $q(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Proof. It can be directly derived from Lemma 5. □

Theorem 4. (h is continuous, VaGES) Suppose (1) $\forall (v, \cdots, \cdots)$, $r_v \log p_\theta(v; h)$, $r_\cdots \log p_\theta(v; h)$, $r_v \log p_\theta(v; h) \frac{\partial \log p_\theta(v;h)}{\partial \cdots}$ and $\frac{\partial r_v \log p_\theta(v;h)}{\partial \cdots}$ as functions of h satisfy the Stein regular condition w.r.t. $p(h|v, \cdots, \cdots)$ and $q(h|v, \cdots, \cdots)$ and (2) the Stein factors of $r_v \log p_\theta(v; h)$, $r_\cdots \log p_\theta(v; h)$, $r_v \log p_\theta(v; h) \frac{\partial \log p_\theta(v;h)}{\partial \cdots}$ and $\frac{\partial r_v \log p_\theta(v;h)}{\partial \cdots}$ as functions of h w.r.t. $p(h|v, \cdots, \cdots), q(h|v, \cdots, \cdots)$ are bounded w.r.t. $v, \cdots$ and $\cdots$, (3) $r_v \log p_\theta(v; h)$ and $r_\cdots \log p_\theta(v; h)$ are bounded w.r.t. v, h and $\cdots$, then the bias of $\mathrm{VaGES}(v, \cdots, \cdots)$ can be bounded by the square root of the Fisher divergence between $q(h|v)$ and $p_\theta(h|v)$ up to multiplying a constant.

Proof. According to Lemma 5, $\exists C_1 < \infty$, s.t.

$$\|E_{q(h|v)}[r_v \log p_\theta(v; h)] - r_v \log p_\theta(v)\|_2 \leq C_1 \sqrt{D_F(q(h|v) \| p(h|v))},$$

$\exists C_2 < \infty$, s.t.

$$\|E_{q(h|v)}[r_\cdots \log p_\theta(v; h)] - r_\cdots \log p_\theta(v)\|_2 \leq C_2 \sqrt{D_F(q(h|v) \| p(h|v))},$$

$\exists C_3 < \infty$, s.t.

$$
\begin{aligned}
&||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
\leq & C_3\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))},
\end{aligned}
$$

and $\exists C_4 < \infty$, s.t.

$$
\begin{aligned}
&||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
\leq & C_4\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.
\end{aligned}
$$

By the boundedness of $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ and $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, we can assume $C < \infty$ is a constant that bounds $||\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})||_2$ and $||\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})||_2$. After establishing the above bounds w.r.t. the Fisher divergence, the rest proof is exactly the same as Theorem 2. For completeness , we restate the proof as follows. By the triangle inequality and Lemma. 2, we have

$$
\begin{aligned}
&||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] \\
&- \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
\leq & ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\left(\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]\right)||_2 \\
&+ ||\left(\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\right)\mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
= & ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]||_2\, ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
&+ ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]||_2\, ||\mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
\leq & C||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
&+ C||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]||_2 \\
\leq & CC_2\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} + CC_1\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} \\
= & (CC_2 + CC_1)\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.
\end{aligned}
$$

Thereby,

$$
\begin{aligned}
&||\mathrm{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})) \\
&- \mathrm{Cov}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}))||_2 \\
\leq & ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
&+ ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] \\
&\quad - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \\
\leq & C_4\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} + (CC_2 + CC_1)\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} \\
= & (C_4 + CC_2 + CC_1)\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&||\mathrm{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})) + \mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] \\
&\quad - \frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}(\boldsymbol{v};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}||_2 \\
=&||\mathrm{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})) + \mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] \\
&\quad - \mathrm{Cov}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})) - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] ||_2 \\
\leq&||\mathrm{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})) \\
&\quad - \mathrm{Cov}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}))||_2 \\
&\quad + ||\mathsf{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathsf{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] ||_2 \\
\leq&(C_4 + CC_2 + CC_1)\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} + C_3\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} \\
=&(C_4 + CC_2 + CC_1 + C_3)\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.
\end{aligned}
$$

□

D Consistency between $\mathcal{D}_F^m$ and $\mathcal{D}_F$

Theorem 5. *Suppose* $\lim_{||\boldsymbol{v}||\to\infty} p_t(\boldsymbol{v})(\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_t(\boldsymbol{v})) = \mathtt{0}$ *and* $\mathsf{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] = \boldsymbol{I}$, *then* $\mathcal{D}_F^m(p_t||p_{\boldsymbol{\theta}}) = \mathcal{D}_F(p_t||p_{\boldsymbol{\theta}})$.

Proof. Notice that $\mathsf{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v}) \boldsymbol{\epsilon}\right] = \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v}))$ by the assumption $\mathsf{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] = \boldsymbol{I}$ and $\mathcal{D}_F^m(p_t||p_{\boldsymbol{\theta}})$ can be simplified as

$$
\mathcal{D}_F^m(p_t||p_{\boldsymbol{\theta}}) = \max_{\boldsymbol{f}\in\mathcal{F}} \mathsf{E}_{p_t(\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v})) - \frac{1}{2}||\boldsymbol{f}(\boldsymbol{v})||_2^2\right].
$$

Suppose $\boldsymbol{f} \in \mathcal{F}$, i.e., $\boldsymbol{f}$ is a function from R^d to R^d and $\lim_{||\boldsymbol{v}||\to\infty} p_t(\boldsymbol{v})\boldsymbol{f}(\boldsymbol{v}) = \mathtt{0}$, by the Stein's identity, we have $\mathsf{E}_{p_t(\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log p_t(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v}))\right] = \mathtt{0}$. Thereby, we have

$$
\begin{aligned}
&\mathsf{E}_{p_t(\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v})) - \frac{1}{2}||\boldsymbol{f}(\boldsymbol{v})||_2^2\right] \\
=&\mathsf{E}_{p_t(\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v})) - \frac{1}{2}||\boldsymbol{f}(\boldsymbol{v})||_2^2\right] - \\
&\mathsf{E}_{p_t(\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log p_t(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v}))\right] \\
=&\mathsf{E}_{p_t(\boldsymbol{v})}\left[(\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_t(\boldsymbol{v}))^\top \boldsymbol{f}(\boldsymbol{v}) - \frac{1}{2}||\boldsymbol{f}(\boldsymbol{v})||_2^2\right] \\
\leq&\frac{1}{2}\mathsf{E}_{p_t(\boldsymbol{v})}\left[||\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_t(\boldsymbol{v}))||_2^2\right] = \mathcal{D}_F(p_t||p_{\boldsymbol{\theta}}).
\end{aligned}
$$

The equality is achieved when $\boldsymbol{f}(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_t(\boldsymbol{v})$, which is a function in $\mathcal{F}$ by assumption. As a result, $\mathcal{D}_F^m(p_t||p_{\boldsymbol{\theta}}) = \mathcal{D}_F(p_t||p_{\boldsymbol{\theta}})$.

□

E Experimental Settings

For experiments on GRBMs, $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is always a Bernoulli distribution parametermized by a fully connected layer with the sigmoid activation and we use the Gumbel-Softmax trick [14] for reparameterization of $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ with 0.1 as the temperature. By default, the variational parameter $\boldsymbol{\phi}$ is updated for $K = 5$ times for each minibatch. The batch size is 100 for all experiments. We use the Adam optimizer for all experiments.

E.1 Learning EBLVMs with Kernelized Stein Discrepancy

The density of the checkerboard dataset is shown in Fig. 5 (a). The dimension of $\boldsymbol{h}$ is 4. The learning rate is 0.001.

E.2 Learning EBLVMs with Score Matching

Experiments on GRBMs: The Frey face dataset[2] consists of gray-scaled face images of size 20 × 28. Following BiSM, we split 1,400 images for training, 300 images for validation and 265 images for testing; the dimension of $\boldsymbol{h}$ is 400; the learning rate is 0.0002.

Experiments on deep EBLVMs: The MNIST dataset consists of gray-scaled hand-written digits of size 28 × 28. The dimension of $\boldsymbol{h}$ is 20. Following BiSM, we split 60,000 samples for training and 10,000 samples for testing on MNIST; g_1 consists of a 12-layer ResNet and a linear layer which aligns the output dimension of g_1 with the dimension of $\boldsymbol{h}$; g_3 is a fully connected layer; $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a Gaussian distribution parameterized by a 3-layer convolutional neural network (CNN). For BiMDSM, the times of updating parameters in the lower level problem is $K = 5$ and the number of steps for gradient unrolling in the higher level problem is $N = 0$.

E.3 Evaluating EBLVMs

$\boldsymbol{f}_{\boldsymbol{\eta}}$ is a multilayer perceptron (MLP) with 2 hidden layers and each layer has the same width. The GRBM is initialized as a standard Gaussian distribution by letting $\boldsymbol{b} = \mathtt{0}, \boldsymbol{c} = \mathtt{0}, W = \mathtt{0}, \sigma = 1$, so we can get accurate samples from it. We get 20,000 samples from the initial GRBM, and split 16,000 samples for training, 2,000 samples for validation and 2,000 samples for testing. The learning rate is 0.0002. The number of $\boldsymbol{h}$ sampled for each $\boldsymbol{v}$ is $L = 1$.

[2] http://www.cs.nyu.edu/~roweis/data.html