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

Fan Bao [1,2] Kun Xu [1] Chongxuan Li [1] Lanqing Hong [2] Jun Zhu [1] Bo Zhang [1]

Abstract

This paper presents new estimates of the *score function* and its gradient with respect to the model parameters in a general energy-based latent variable model (EBLVM). The score function and its gradient can be expressed as combinations of expectation and covariance terms over the (generally intractable) posterior of the latent variables. New estimates are obtained by introducing a variational posterior to approximate the true posterior in these terms. The variational posterior is trained to minimize a certain divergence (e.g., the KL divergence) between itself and the true posterior. Theoretically, the divergence characterizes upper bounds of the bias of the estimates. In principle, our estimates can be applied to a wide range of objectives, including *kernelized Stein discrepancy* (KSD), *score matching* (SM)-based methods and *exact Fisher divergence* with a minimal model assumption. In particular, these estimates applied to SM-based methods outperform existing methods in learning EBLVMs on several image datasets.

1. Introduction

Energy-based models (EBMs) (LeCun et al., 2006) associate an energy to each configuration of visible variables, which naturally induces a probability distribution by normalizing the exponential negative energy. Such models have found applications in a wide range of areas, such as image synthesis (Du & Mordatch, 2019; Nijkamp et al., 2020), out-of-distribution detection (Grathwohl et al., 2020a; Liu et al., 2020) and controllable generation (Nijkamp et al., 2019). Based on EBMs, energy-based latent variable models (EBLVMs) (Swersky et al., 2011; Vértes et al., 2016) incorporate latent variables to further improve the expressive power (Salakhutdinov & Hinton, 2009; Bao et al., 2020) and to enable representation learning (Welling et al., 2004; Salakhutdinov & Hinton, 2009; Srivastava & Salakhutdinov, 2012) and conditional sampling (see results in Sec. 4.2).

Notably, the *score function* (SF) of an EBM is defined as the gradient of its log-density with respect to the visible variables, which is independent of the partition function and thereby tractable. Due to its tractability, the SF-based methods are appealing in both learning (Hyvärinen, 2005; Liu et al., 2016) and evaluating (Grathwohl et al., 2020b) EBMs, compared to many other approaches (Hinton, 2002; Tieleman, 2008; Gutmann & Hyvärinen, 2010; Salakhutdinov, 2008) (see a comprehensive discussion in Sec. 6). In particular, *score matching* (SM) (Hyvärinen, 2005), which minimizes the expected squared distance between the score function of the data distribution and that of the model distribution, and its variants (Kingma & LeCun, 2010; Vincent, 2011; Saremi et al., 2018; Song et al., 2019; Li et al., 2019; Pang et al., 2020) have shown promise in learning EBMs.

Unfortunately, the score function and its gradient with respect to the model parameters in EBLVMs are generally intractable without a strong structural assumption (e.g., the joint distribution of visible and latent variables is in the exponential family (Vértes et al., 2016)) and thereby SF-based methods are not directly applicable to general EBLVMs. Recently, bi-level score matching (BiSM) (Bao et al., 2020) applies SM to general EBLVMs by reformulating a SM-based objective as a bilevel optimization problem, which learns a deep EBLVM on natural images and outperforms an EBM of the same size. However, BiSM is not problemless — practically BiSM is optimized by gradient unrolling (Metz et al., 2017) of the lower level optimization, which is time and memory consuming (see a comparison in Sec. 4.1).

In this paper, we present variational estimates of the score function and its gradient with respect to the model parameters in general EBLVMs, referred to as *VaES* and *VaGES* respectively. The score function and its gradient can be expressed as combinations of expectation and covariance terms over the posterior of the latent variables. VaES and VaGES are obtained by introducing a variational posterior, which approximates the true posterior in these terms by minimizing a certain divergence between itself and the true

[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. Correspondence to: Jun Zhu <dcszj@tsinghua.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

posterior. Theoretically, 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 (Johnson, 2004) between the variational posterior and the true one (see Theorem 1,2,3,4).

VaES and VaGES are generally applicable in a wide range of SF-based methods, including *kernelized Stein discrepancy* (KSD) (Liu et al., 2016) (see Sec. 3) and *score matching* (SM)-based methods (Vincent, 2011; Li et al., 2019) (see Sec. 4) for learning EBLVMs and estimating the *exact Fisher divergence* (Hu et al., 2018; Grathwohl et al., 2020b) (see Sec. 5) between the data distribution and the model distribution for evaluating EBLVMs. In particular, VaES and VaGES applied to SM-based methods are superior in time and memory consumption compared with the complex gradient unrolling in the strongest baseline BiSM (see Sec. 4.1) and applicable to learn deep EBLVMs on the MNIST (LeCun et al., 2010), CIFAR10 (Krizhevsky et al., 2009) and CelebA (Liu et al., 2015) datasets (see Sec. 4.2). We also present latent space interpolation results of deep EBLVMs (see Sec. 4.2), which haven't been investigated in previous EBLVMs to our knowledge.

2. Method

As mentioned in Sec. 1, the *score function* has been used in a wide range of methods (Hyvärinen, 2005; Liu et al., 2016; Grathwohl et al., 2020b), while it is generally intractable in EBLVMs without a strong structural assumption. In this paper, we present estimates of the score function and its gradient w.r.t. the model parameters in a general EBLVM. Formally, an EBLVM defines a joint probability distribution over the visible variables $\boldsymbol{v}$ and latent variables $\boldsymbol{h}$ as follows

$$p_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) = \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})/\mathcal{Z}(\boldsymbol{\theta}) = e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}/\mathcal{Z}(\boldsymbol{\theta}), \quad (1)$$

where $\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ is the energy function parameterized by $\boldsymbol{\theta}$, $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) = e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}$ is the unnormalized distribution 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. As shown by Vértes et al. (2016), the score function of an EBLVM can be written as an expectation over the posterior:

$$\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right], \quad (2)$$

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

$$\frac{\partial \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial \boldsymbol{\theta}} = \underbrace{\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]}_{\text{denoted as } e(\boldsymbol{v};\boldsymbol{\theta})} + \underbrace{\mathrm{Cov}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}))}_{\text{denoted as } c(\boldsymbol{v};\boldsymbol{\theta})}, \quad (3)$$

where the first 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 $\boldsymbol{h} \sim p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ and the second term is the covariance between two random vectors $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ and $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ with $\boldsymbol{h} \sim p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$.

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

Eq. (2) and (3) naturally suggest Monte Carlo estimates, while both of them need samples from the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$, which is generally intractable. As for approximate inference, we present an amortized variational approach (Kingma & Welling, 2014) that considers

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right], \quad (4)$$

and

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] + \mathrm{Cov}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}(\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}), \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})), \quad (5)$$

respectively, where $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a variational posterior parameterized by $\boldsymbol{\phi}$. Note that the covariance term in Eq. (5) is a variational estimate of that in Eq. (3), not derived from taking gradient to Eq. (4). We optimize some divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$

$$\min_{\boldsymbol{\phi}} \mathbb{E}_{p_D(\boldsymbol{v})}\left[\mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))\right], \quad (6)$$

where $p_D(\boldsymbol{v})$ denotes the data distribution. Specifically, if we set $\mathcal{D}$ in Eq. (6) as the KL divergence or the Fisher divergence, then Eq. (6) is tractable. For clarity, we refer the readers to Appendix B for a derivation and a general analysis of the tractability of other divergences.

Below, we consider Monte Carlo estimates based on the variational approximation. According to Eq. (4), the variational estimate of the score function (VaES) is

$$\mathrm{VaES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{1}{L}\sum_{i=1}^{L} \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i), \quad (7)$$

where $\boldsymbol{h}_i \overset{\text{i.i.d}}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and L is the number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. Similarly, according to Eq. (5), the variational estimate of the expectation term $e(\boldsymbol{v};\boldsymbol{\theta})$ is

$$\hat{e}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{1}{L}\sum_{i=1}^{L} \frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{\partial \boldsymbol{\theta}}, \quad (8)$$

where $\boldsymbol{h}_i \overset{\text{i.i.d}}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. As for the covariance term $c(\boldsymbol{v};\boldsymbol{\theta})$, we estimate it with the sample covariance matrix (Fan et al., 2016). According to Eq. (5), the variational estimate is

$$\hat{c}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{1}{L-1}\sum_{i=1}^{L}\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)\frac{\partial\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{\partial\boldsymbol{\theta}} - \frac{1}{(L-1)L}\sum_{i=1}^{L}\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)\sum_{i=1}^{L}\frac{\partial\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{\partial\boldsymbol{\theta}}, \tag{9}$$

where $\boldsymbol{h}_i \overset{\text{i.i.d}}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, $\nabla_{\boldsymbol{v}}\cdot$ outputs column vectors and $\frac{\partial\cdot}{\partial\boldsymbol{\theta}}$ outputs row vectors. With Eq. (8) and (9), the variational gradient estimate of the score function (VaGES) is

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

Remark: Although Eq. (8) includes second derivatives, in practice we only need to calculate the product of them and vectors. Similar to the Hessian-vector products (Song et al., 2019), only two backpropagations are required in the calculation (since $\boldsymbol{z}^\top\frac{\partial\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{\partial\boldsymbol{\theta}} = \frac{\partial\boldsymbol{z}^\top\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{\partial\boldsymbol{\theta}}$).

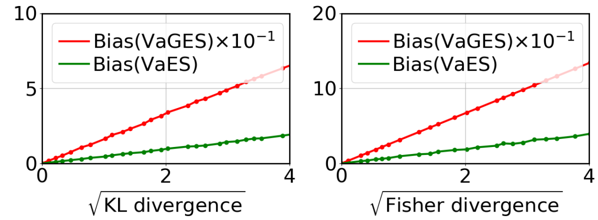
2.2. Bounding the Bias

Notice that $\mathrm{VaES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ and $\mathrm{VaGES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ are actually biased estimates due to the difference between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$. Firstly, we show that the bias of $\mathrm{VaES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ and $\mathrm{VaGES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ can be bounded by the square root of the KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ under some assumptions on boundedness, as characterized in Theorem 1 and Theorem 2.

Theorem 1. *(VaES, KL, proof in Appendix C) Suppose $\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ is bounded w.r.t. $\boldsymbol{v},\boldsymbol{h}$ and $\boldsymbol{\theta}$, then the bias of $\mathrm{VaES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ can be bounded by the square root of the KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ up to multiplying a constant.*

Theorem 2. *(VaGES, KL, proof in Appendix C) Suppose $\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{\theta}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ and $\frac{\partial\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial\boldsymbol{\theta}}$ are bounded w.r.t. $\boldsymbol{v},\boldsymbol{h}$ and $\boldsymbol{\theta}$, then the bias of $\mathrm{VaGES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ can be bounded by the square root of the KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ up to multiplying a constant.*

Remark: The boundedness assumptions in Theorem 1,2 are not difficult to satisfy when $\boldsymbol{h}$ is discrete. For example, if $\boldsymbol{v}$ comes from a compact set (which is naturally satisfied on image data), $\boldsymbol{\theta}$ comes from a compact set (which can be satisfied by adding weight decay to $\boldsymbol{\theta}$), $\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ is second continuously differentiable w.r.t. $\boldsymbol{v}$ and $\boldsymbol{\theta}$ (which is satisfied in neural networks composed of smooth functions) and $\boldsymbol{h}$ comes from a finite set (which is satisfied in most commonly used discrete distributions, e.g., the Bernoulli distribution), then $\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{\theta}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ and $\frac{\partial\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial\boldsymbol{\theta}}$ are bounded w.r.t. $\boldsymbol{v},\boldsymbol{h}$ and $\boldsymbol{\theta}$ by the extreme value theorem of continuous functions on compact sets.



(a) GRBM, discrete $\boldsymbol{h}$ (b) GM, continuous $\boldsymbol{h}$

Figure 1. The biases of VaES and VaGES v.s. the square root of (a) the KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ in a GRBM and (b) the Fisher divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ in a Gaussian model (GM). See Appendix F.4 for experimental details.

In Fig. 1 (a), we numerically validate Theorem 1,2 in Gaussian restricted Boltzmann machines (GRBMs) (Hinton & Salakhutdinov, 2006). The biases of VaES and VaGES in such models are linearly proportional to the square root of the KL divergence. Theorem 1,2 strongly motivate us to set $\mathcal{D}$ as the KL divergence in Eq. (6) when $\boldsymbol{h}$ is discrete.

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 $\mathrm{VaES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ and $\mathrm{VaGES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ can be bounded by the square root of the Fisher divergence (Johnson, 2004) between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$, as characterized in Theorem 3 and Theorem 4. We first introduce some definitions, which will be used in the theorems.

Definition 1. (Ley et al., 2013) *Suppose p is a probability density defined on $\mathbb{R}^n$ and $f:\mathbb{R}^n\to\mathbb{R}$ is a function, we define $\boldsymbol{g}_f^p$ as a solution of the Stein equation $\mathcal{S}_p\boldsymbol{g} = f - \mathbb{E}_p f$, where $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$ and $\mathcal{S}_p\boldsymbol{g}(\boldsymbol{x}) \triangleq \nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^\top\boldsymbol{g}(\boldsymbol{x}) + \mathrm{Tr}(\nabla_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x}))$. (See Appendix D for the existence of the solution.)*

Definition 2. *Suppose p,q are probability densities defined on $\mathbb{R}^n$ and $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^m$ is a function, we say $\boldsymbol{f}$ satisfies the Stein regular condition w.r.t. p,q iff $\forall i\in\mathbb{Z}\cap[1,m]$, $\lim\limits_{||\boldsymbol{x}||\to\infty} q(\boldsymbol{x})\boldsymbol{g}_{f_i}^p(\boldsymbol{x}) = 0$.*

Definition 3. (Ley et al., 2013) *Suppose p,q are probability densities defined on $\mathbb{R}^n$ and $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^m$ is a function satisfying the Stein regular condition w.r.t. p,q, we define $\kappa_{\boldsymbol{f}}^{p,q} \triangleq \sqrt{\mathbb{E}_{q(\boldsymbol{x})}\sum\limits_{i=1}^{m}||\boldsymbol{g}_{f_i}^p(\boldsymbol{x})||_2^2}$, referred to as the Stein factor of $\boldsymbol{f}$ w.r.t. p,q.*

Theorem 3. *(VaES, Fisher, proof in Appendix D) Suppose (1) $\forall(\boldsymbol{v},\boldsymbol{\theta},\boldsymbol{\phi})$, $\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ as a function of $\boldsymbol{h}$ satisfies the Stein regular condition w.r.t. $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and (2) the Stein factor of $\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ as a function of $\boldsymbol{h}$ w.r.t. $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}), q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is bounded w.r.t. $\boldsymbol{v},\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, then the bias of $\mathrm{VaES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ can be bounded by the square root of the Fisher divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ up to multiplying a constant.*

Theorem 4. *(VaGES, Fisher, proof in Appendix D) Suppose* (1) $\forall(\boldsymbol{v},\boldsymbol{\theta},\boldsymbol{\phi})$, $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ *and* $\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ *as functions of* $\boldsymbol{h}$ *satisfy the Stein regular condition w.r.t.* $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ *and* $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ *and* (2) *the Stein factors of* $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ *and* $\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ *as functions of* $\boldsymbol{h}$ *w.r.t.* $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}), q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ *are bounded w.r.t.* $\boldsymbol{v},\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$, (3) $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ *and* $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ *are bounded w.r.t.* $\boldsymbol{v},\boldsymbol{h}$ *and* $\boldsymbol{\theta}$, *then the bias of* $\mathrm{VaGES}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi})$ *can be bounded by the square root of the Fisher divergence between* $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ *and* $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ *up to multiplying a constant.*

Although the boundedness of the Stein factors has only been verified under some simple cases (Ley et al., 2013), and hasn't been extended to more complex cases, e.g., when $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ is parameterized by a neural network, we find it work in practice when choosing $\mathcal{D}$ in Eq. (6) as the Fisher divergence (see Sec. 4.2) to learn $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. In Fig. 1 (b), we numerically validate Theorem 3,4 in a Gaussian model (GM), whose energy is $\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{2\sigma^2}||\boldsymbol{v}-\boldsymbol{b}||^2 + \frac{1}{2}||\boldsymbol{h}-\boldsymbol{c}||^2 - \boldsymbol{v}^\top W \boldsymbol{h}$ with $\boldsymbol{\theta} = (\sigma, W, \boldsymbol{b}, \boldsymbol{c})$. The biases of VaES and VaGES in such models are linearly proportional to the square root of the Fisher divergence.

2.3. Langevin Dynamics Corrector

Sometimes $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ can be complex for $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ to approximate, especially in deep models. To improve the inference performance, we can run a few Langevin dynamics steps (Welling & Teh, 2011) to correct samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, which has been successfully applied to correct the solution of a numerical SDE solver (Song et al., 2020). Specifically, a Langevin dynamics step updates $\boldsymbol{h}$ by

$$\boldsymbol{h} \leftarrow \boldsymbol{h} + \frac{\alpha}{2} \underbrace{\nabla_{\boldsymbol{h}} \log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}_{\substack{\shortparallel \\ \nabla_{\boldsymbol{h}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}} + \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \alpha). \tag{11}$$

In practice, the standard deviation of $\boldsymbol{\epsilon}$ is often smaller than $\sqrt{\alpha}$ to allow a faster convergence (Du & Mordatch, 2019; Nijkamp et al., 2020; 2019; Grathwohl et al., 2020a). Since Langevin dynamics requires $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ to be differentiable w.r.t. $\boldsymbol{h}$, we only apply the corrector when $\boldsymbol{h}$ is continuous.

3. Learning EBLVMs with KSD

In this section, we show that VaES and VaGES can extend kernelized Stein discrepancy (KSD) (Liu et al., 2016) to learn general EBLVMs. The KSD between the data distribution $p_D(\boldsymbol{v})$ and the model distribution $p_{\boldsymbol{\theta}}(\boldsymbol{v})$ with the kernel

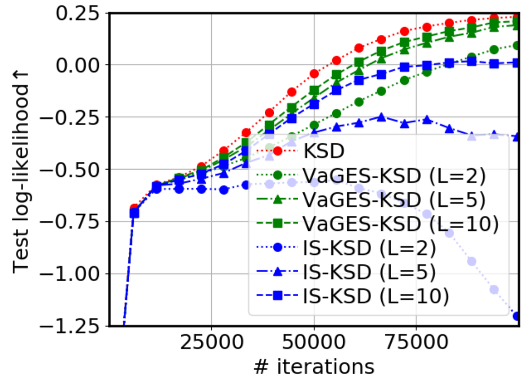


Figure 2. Comparison of KSD, VaGES-KSD and IS-KSD on checkerboard. The test log-likelihood is averaged over 10 runs.

$k(\boldsymbol{v},\boldsymbol{v}')$ is defined as

$$\begin{aligned}
&\mathrm{KSD}(p_D, p_{\boldsymbol{\theta}}) \\
\triangleq\ &\mathbb{E}_{\boldsymbol{v},\boldsymbol{v}'\sim p_D}[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top k(\boldsymbol{v},\boldsymbol{v}')\nabla_{\boldsymbol{v}'} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}') \\
&\qquad + \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \nabla_{\boldsymbol{v}'} k(\boldsymbol{v},\boldsymbol{v}') \\
&\qquad + \nabla_{\boldsymbol{v}} k(\boldsymbol{v},\boldsymbol{v}')^\top \nabla_{\boldsymbol{v}'} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}') \\
&\qquad + \mathrm{Tr}(\nabla_{\boldsymbol{v}}\nabla_{\boldsymbol{v}'} k(\boldsymbol{v},\boldsymbol{v}'))],
\end{aligned} \tag{12}$$

which properly measures the difference between $p_D(\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{v})$ under some mild assumptions (Liu et al., 2016). To learn an EBLVM, we use the gradient-based optimization to minimize $\mathrm{KSD}(p_D, p_{\boldsymbol{\theta}})$, where the gradient w.r.t. $\boldsymbol{\theta}$ is

$$\begin{aligned}
\frac{\partial \mathrm{KSD}(p_D, p_{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} &= 2\mathbb{E}_{\boldsymbol{v},\boldsymbol{v}'\sim p_D}\Big[(k(\boldsymbol{v},\boldsymbol{v}')\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) \\
&\qquad + \nabla_{\boldsymbol{v}} k(\boldsymbol{v},\boldsymbol{v}'))^\top \frac{\partial \nabla_{\boldsymbol{v}'} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}')}{\partial \boldsymbol{\theta}}\Big].
\end{aligned} \tag{13}$$

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

$$\begin{aligned}
&\frac{2}{M}\sum_{i=1}^{M}[k(\boldsymbol{v}_i,\boldsymbol{v}_i')\mathrm{VaES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi}) \\
&\qquad + \nabla_{\boldsymbol{v}} k(\boldsymbol{v}_i,\boldsymbol{v}_i')]^\top \mathrm{VaGES}(\boldsymbol{v}_i';\boldsymbol{\theta},\boldsymbol{\phi}),
\end{aligned} \tag{14}$$

where the union of $\boldsymbol{v}_{1:M}$ and $\boldsymbol{v}'_{1:M}$ is a mini-batch from the data distribution, and $\mathrm{VaES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi})$ and $\mathrm{VaGES}(\boldsymbol{v}_i';\boldsymbol{\theta},\boldsymbol{\phi})$ are independent.

3.1. Experiments

Setting. Since KSD itself suffers from the curse of dimensionality issue (Gong et al., 2020), we illustrate the validity of VaGES-KSD by learning Gaussian restricted Boltzmann machines (GRBMs) (Welling et al., 2004; Hinton & Salakhutdinov, 2006) on the 2-D checkerboard dataset, whose distribution is shown in Appendix G.1. The energy function of a GRBM is

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{2\sigma^2}||\boldsymbol{v}-\boldsymbol{b}||^2 - \boldsymbol{c}^\top \boldsymbol{h} - \boldsymbol{v}^\top W \boldsymbol{h}, \qquad (15)$$

with learnable parameters $\boldsymbol{\theta} = (\sigma, W, \boldsymbol{b}, \boldsymbol{c})$. Since GRBM has a tractable posterior, we can directly learn it using KSD (see Eq. (12)). We also compare another baseline where $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})$ is estimated by importance sampling (Owen, 2013) (i.e., $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}) \approx \frac{1}{L}\sum_{i=1}^{L} \frac{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{\mathrm{unif}(\boldsymbol{h}_i)}, \boldsymbol{h}_i \overset{\text{i.i.d}}{\sim} \mathrm{unif}(\boldsymbol{h})$, where unif means the uniform distribution) and we call it IS-KSD. We use the RBF kernel $k(\boldsymbol{v},\boldsymbol{v}') = \exp(-\frac{||\boldsymbol{v}-\boldsymbol{v}'||_2^2}{2\sigma^2})$ with $\sigma = 0.1$. $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a Bernoulli distribution parametermized by a fully connected layer with the sigmoid activation and we use the Gumbel-Softmax trick (Jang et al., 2017) for reparameterization of $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ with 0.1 as the temperature. $\mathcal{D}$ in Eq. (6) is the KL divergence. See Appendix F.1 for more experimental details.

Result. As shown in Fig. 2, VaGES-KSD outperforms IS-KSD on all L (the number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$) and is comparable to the KSD (the test log-likelihood curves of VaGES-KSD are very close to KSD when $L = 5, 10$).

4. Learning EBLVMs with Score Matching

In this section, we show that VaES and VaGES can extend two score matching (SM)-based methods (Vincent, 2011; Li et al., 2019) to learn general EBLVMs. The first one is the denoising score matching (DSM) (Vincent, 2011), which minimizes the Fisher divergence $\mathcal{D}_F$ between a perturbed data distribution $p_{\sigma_0}(\boldsymbol{v})$ and the model distribution $p_{\boldsymbol{\theta}}(\boldsymbol{v})$:

$$\begin{aligned}&\mathcal{J}_{DSM}(\boldsymbol{\theta}) \triangleq \mathcal{D}_F(p_{\sigma_0}||p_{\boldsymbol{\theta}}) \qquad (16)\\ \equiv&\frac{1}{2}\mathbb{E}_{p_D(\boldsymbol{w})p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})}||\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}}\log p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})||_2^2,\end{aligned}$$

where $p_D(\boldsymbol{w})$ is the data distribution, $p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{v}|\boldsymbol{w},\sigma_0^2 I)$ is the Gaussian perturbation with a fixed noise level σ_0, $p_{\sigma_0}(\boldsymbol{v}) = \int p_D(\boldsymbol{w})p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})\mathrm{d}\boldsymbol{w}$ is the perturbed data distribution and $\equiv$ means equivalence in parameter optimization. The second one is the multiscale denoising score matching (MDSM) (Li et al., 2019), which uses different noise levels to scale up DSM to high-dimensional data:

$$\begin{aligned}&\mathcal{J}_{MDSM}(\boldsymbol{\theta}) \qquad (17)\\ \triangleq&\frac{1}{2}\mathbb{E}_{p_D(\boldsymbol{w})p(\sigma)p_\sigma(\boldsymbol{v}|\boldsymbol{w})}||\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}}\log p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})||_2^2,\end{aligned}$$

where $p(\sigma)$ is a prior distribution over the flexible noise level σ and σ_0 is a fixed noise level. We extend DSM and MDSM to learn EBLVMs. Firstly we write Eq. (16) and (17) in a general form for the convenience of discussion

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{p_D(\boldsymbol{w},\boldsymbol{v})}||\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}}\log p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})||_2^2, \qquad (18)$$

where $p_D(\boldsymbol{w},\boldsymbol{v})$ is the joint distribution of $\boldsymbol{w}$ and $\boldsymbol{v}$ (specifically, $p_D(\boldsymbol{w},\boldsymbol{v}) = p_D(\boldsymbol{w})p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})$ for Eq. (16) and $p_D(\boldsymbol{w},\boldsymbol{v}) = \int_\sigma p_D(\boldsymbol{w})p(\sigma)p_\sigma(\boldsymbol{v}|\boldsymbol{w})\mathrm{d}\sigma$ for Eq. (17)). We use gradient-based optimization to minimize $\mathcal{J}(\boldsymbol{\theta})$ and its gradient w.r.t. $\boldsymbol{\theta}$ is

$$\begin{aligned}\frac{\mathcal{J}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} =&\mathbb{E}_{p_D(\boldsymbol{w},\boldsymbol{v})}\Big\{\big[\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v})\\ &- \nabla_{\boldsymbol{v}}\log p_{\sigma_0}(\boldsymbol{v}|\boldsymbol{w})\big]\frac{\partial\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial\boldsymbol{\theta}}\Big\}. \qquad (19)\end{aligned}$$

Estimating $\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v})$ and $\frac{\partial\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial\boldsymbol{\theta}}$ with VaES and VaGES respectively, the variational stochastic gradient estimate of the SM-based methods (VaGES-SM) is

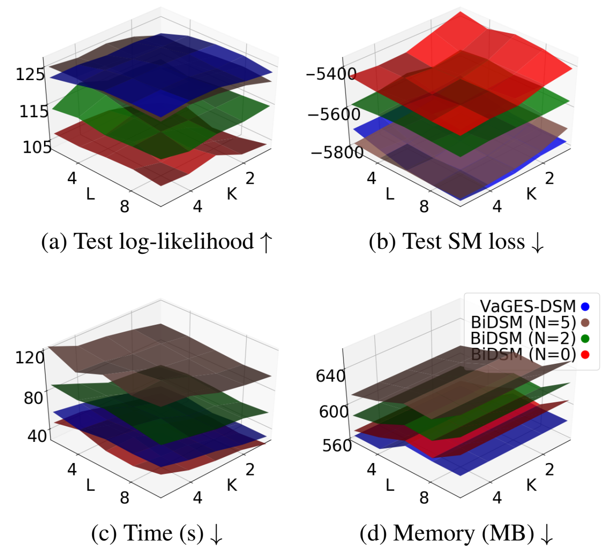
$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}&\big[\mathrm{VaES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi})\\ &- \nabla_{\boldsymbol{v}}\log p_{\sigma_0}(\boldsymbol{v}_i|\boldsymbol{w}_i)\big]\mathrm{VaGES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi}), \qquad (20)\end{aligned}$$

where $(\boldsymbol{w}_{1:M},\boldsymbol{v}_{1:M})$ is a mini-batch from $p_D(\boldsymbol{w},\boldsymbol{v})$, and $\mathrm{VaES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi})$ and $\mathrm{VaGES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi})$ are independent. We explicitly denote our methods as VaGES-DSM or VaGES-MDSM according to which objective is used.

4.1. Comparison in GRBMs

Setting. We firstly consider the GRBM (see Eq. (15)), which is a good benchmark to compare existing methods. We consider the Frey face dataset, which consists of gray-scaled face images of size 20×28, and the checkerboard dataset (mentioned in Sec. 3.1). $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a Bernoulli distribution parametermized by a fully connected layer with the sigmoid activation and we use the Gumbel-Softmax trick (Jang et al., 2017) for reparameterization of $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ with 0.1 as the temperature. $\mathcal{D}$ in Eq. (6) is the KL divergence. See Appendix F.2.1 for more experimental details.

Comparison with BiSM. As mentioned in Sec. 1, bi-level score matching (BiSM) (Bao et al., 2020) learns general EBLVMs by reformulating Eq. (16) and (17) as bi-level optimization problems, referred to as BiDSM and BiMDSM respectively, which serves as the most direct baseline (see Appendix I for an introduction to BiSM). We make a comprehensive comparison between VaGES-SM and BiSM on the Frey face dataset, which explores all hyperparameters involved in these two methods. Specifically, both methods sample from a variational posterior and optimize the variational posterior, so they share two hyperparameters L (the number of $\boldsymbol{h}$ sampled from the variational posterior) and K (the number of times updating the variational posterior on each mini-batch). Besides, BiSM has an extra hyperparameter N, which specifies the number of gradient unrolling steps. In Fig. 3, we plot the test log-likelihood, the test score matching (SM) loss (Hyvärinen, 2005), which is the Fisher divergence up to an additive constant, the time



(a) Test log-likelihood ↑ (b) Test SM loss ↓

(c) Time (s) ↓ (d) Memory (MB) ↓

Figure 3. Comparing VaGES-DSM and BiDSM on Frey face. The test SM loss is the Fisher divergence up to a additive constant.

and memory consumption by grid search across L, K and N. VaGES-DSM takes about 40% time and 90% memory on average to achieve a similar performance with BiDSM (N=5). VaGES-DSM outperforms BiDSM (N=0,2), and meanwhile it takes less time than BiDSM (N=2) and the least memory.

Comparison with other methods. For completeness, we also compare with DSM (Vincent, 2011), CD-based methods (Hinton, 2002; Tieleman, 2008) and noise contrastive estimation (NCE)-based methods (Gutmann & Hyvärinen, 2010; Rhodes & Gutmann, 2019) on the checkerboard dataset. VaGES-DSM is competitive to the baselines (e.g., CD and DSM) that leverage the tractability of the posterior in GRBM and outperforms the variational NCE (Rhodes & Gutmann, 2019). See results in Appendix G.2.1.

4.2. Learning Deep EBLVMs

Setting. Then we show that VaGES-SM can scale up to learn general deep EBLVMs and generate natural images. Following BiSM (Bao et al., 2020), we consider the deep EBLVM with the following energy function

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) = g_3(g_2(g_1(\boldsymbol{v};\boldsymbol{\theta}_1),\boldsymbol{h});\boldsymbol{\theta}_2), \tag{21}$$

where $\boldsymbol{h}$ is continuous, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, g_1 a vector-valued neural network that outputs a feature sharing the same dimension with $\boldsymbol{h}$, g_2 is an additive coupling layer (Dinh et al., 2015) and g_3 is a scalar-valued neural network. We consider the MNIST dataset (LeCun et al., 2010), which consists of gray-scaled hand-written digits of size 28×28, the CIFAR10 dataset (Krizhevsky et al., 2009), which consists of color natural images of size 32×32 and the CelebA dataset (Liu et al., 2015), which consists of color face images. We resize CelebA to 64×64. $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a Gaussian distribution parameterized by a 3-layer convolutional neural network. $\mathcal{D}$ in Eq. (6) is the Fisher divergence. We use the Langevin dynamics corrector (see Sec. 2.3) to correct samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. As for the step size α in Langevin dynamics, we empirically find that the optimal one is approximately proportional to the dimension of $\boldsymbol{h}$ (see Appendix F.2.2), perhaps because Langevin dynamics converges to its stationary distribution slower when $\boldsymbol{h}$ has a higher dimension and thereby requires a larger step size, so we fix the ratio of the step size α in Langevin dynamics to the dimension of $\boldsymbol{h}$ on one dataset and the ratio is 2×10^{-4} on MNIST and CIFAR10 and 10^{-5} on CelebA. The standard deviation of the noise $\boldsymbol{\epsilon}$ in Langevin dynamics is 10^{-4}. See Appendix F.2.2 for more experimental details, e.g., how hyperparameters are selected and time consumption.

Table 1. FID on CIFAR10 and CelebA (64×64). † Averaged by 5 runs. ‡ Since BiSM doesn't report a FID on CelebA, the value is evaluated in our reproduction.

(a) CIFAR10

Methods	FID ↓
Flow-CE (Gao et al., 2020)	37.30
VAE-EBLVM (Han et al., 2020)	30.1
CoopNets (Xie et al., 2018)	33.61
EBM (Du & Mordatch, 2019)	38.2
MDSM (Li et al., 2019)	31.7
BiMDSM (Bao et al., 2020)	$29.43 \pm 2.76^{\dagger}$
VaGES-MDSM (ours)	$\mathbf{28.93} \pm 1.91^{\dagger}$

(b) CelebA

Methods	FID ↓
BiMDSM (Bao et al., 2020)	$32.43^{\ddagger}$
VaGES-MDSM (ours)	**31.53**

Sample quality. Since the EBLVM defined by Eq. (21) has an intractable posterior, BiSM is the only applicable baseline mentioned in Section 4.1. We quantitatively evaluate the sample quality with the FID score (Heusel et al., 2017) in Tab. 1 and VaGES-MDSM achieves comparable performance to BiMDSM. Since there are relatively few baselines of learning deep EBLVMs, we also compare with other baselines involving learning EBMs in Tab. 1. We mention that VAE-EBLVM (Han et al., 2020) and CoopNets (Xie et al., 2018) report FID results on a subset of CelebA and on a different resolution of CelebA respectively. For fairness, we don't include these results on CelebA. We provide image samples and Inception Score results in Appendix G.2.2.

Interpolation in the latent space. As argued in prior work (Bao et al., 2020), the conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{h})$

(a) MNIST (b) CIFAR10 (c) CelebA

Figure 4. Interpolation of annealed Langevin dynamics trajectories in the latent space. Each row displays a trajectory of $p_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{h})$ for a fixed latent variable $\boldsymbol{h}$ and $\boldsymbol{h}$ changes smoothly between rows. Different trajectories share the same initial point.

of a deep EBLVM can be multimodal, so we want to explore how a local modal of $p_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{h})$ evolves as $\boldsymbol{h}$ changes. Since $\boldsymbol{v}$ is of high dimension, we are not able to plot the landscape of $p_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{h})$. Alternatively, we study how a fixed particle moves in different landscapes according to the annealed Langevin dynamics (Li et al., 2019) by interpolating $\boldsymbol{h}$. Specifically, we select two points in the latent space and interpolate linearly between them to get a class of conditional distributions $\{p_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{h}_i)\}_{i=1}^T$. Then we fix an initial point $\boldsymbol{v}_0$ and run annealed Langevin dynamics on these conditional distributions with a shared noise. These annealed Langevin dynamics trajectories are shown in Fig. 4. Starting from the same initial noise, the trajectory evolves smoothly as $\boldsymbol{h}$ changes smoothly. More interpolation results can be found in Appendix G.2.2.

Sensitivity analysis. We also study how hyperparameters influence the performances of VaGES-SM in deep EBLVMs, including the dimension of $\boldsymbol{h}$, the number of convolutional layers, the number of $\boldsymbol{h}$ sampled from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, the number of times of updating $\boldsymbol{\phi}$, the number of Langevin dynamics steps, the noise level in Langevin dynamics and which divergence to use in Eq. (6). These results can be found in Appendix G.2.2. In brief, increasing the number of convolutional layers will improve the performance, while the dimension of $\boldsymbol{h}$, the number of $\boldsymbol{h}$ sampled from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, the noise level in Langevin dynamics and using the KL divergence instead of the Fisher divergence in Eq. (6) don't affect the result very much. Besides, setting both the number of times updating $\boldsymbol{\phi}$ and the number of Langevin dynamics steps to 5 is enough for a stable training process.

5. Evaluating EBLVMs with Exact Fisher Divergence

In this setting we are given an EBLVM and a set of samples $\{\boldsymbol{v}_i\}_{i=1}^n$ from the data distribution $p_D(\boldsymbol{v})$. We want to measure how well the model distribution $p_{\boldsymbol{\theta}}(\boldsymbol{v})$ approximates $p_D(\boldsymbol{v})$, which needs an absolute value representing the difference between them. This task is more difficult than model comparison, which only needs relative values (e.g., log-likelihood) to compare between different models. Specifically, we consider the Fisher divergence of the maximum form (Hu et al., 2018; Grathwohl et al., 2020b) to measure how well $p_{\boldsymbol{\theta}}(\boldsymbol{v})$ approximates $p_D(\boldsymbol{v})$:

$$\mathcal{D}_F^m(p_D||p_{\boldsymbol{\theta}}) \triangleq \max_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_{p_D(\boldsymbol{v})}\mathbb{E}_{p(\boldsymbol{\epsilon})}\big[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v})\boldsymbol{\epsilon} - \frac{1}{2}||\boldsymbol{f}(\boldsymbol{v})||_2^2\big], \tag{22}$$

where $\mathcal{F}$ is the set of functions $\boldsymbol{f}:\mathbb{R}^d \to \mathbb{R}^d$ satisfying $\lim_{||\boldsymbol{v}||\to\infty} p_D(\boldsymbol{v})\boldsymbol{f}(\boldsymbol{v}) = \boldsymbol{0}$ and $p(\boldsymbol{\epsilon})$ is a noise distribution (e.g., Gaussian distribution) introduced for computation efficiency (Grathwohl et al., 2020b). Such a form can also be understood as the Stein discrepancy (Gorham & Mackey, 2017) with $\mathcal{L}_2$ constraint on the function space $\mathcal{F}$ and has been applied to evaluate GRBMs (Grathwohl et al., 2020b). Under some mild assumptions (see Appendix E), Eq. (22) is equal to the exact Fisher divergence. In practice, $\mathcal{F}$ is approximated by a neural network $\{\boldsymbol{f}_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in H\}$, where $\boldsymbol{\eta}$ is the parameter and H is the parameter space. We optimize the right hand side of Eq. (22) and the gradient w.r.t. $\boldsymbol{\eta}$ is

$$\mathbb{E}_{p_D(\boldsymbol{v})}\mathbb{E}_{p(\boldsymbol{\epsilon})}\big[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \frac{\partial \boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{v})}{\partial \boldsymbol{\eta}} + \frac{\partial \boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{v}} \boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{v})\boldsymbol{\epsilon}}{\partial \boldsymbol{\eta}} - \frac{1}{2}\frac{\partial ||\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{v})||_2^2}{\partial \boldsymbol{\eta}}\big]. \tag{23}$$

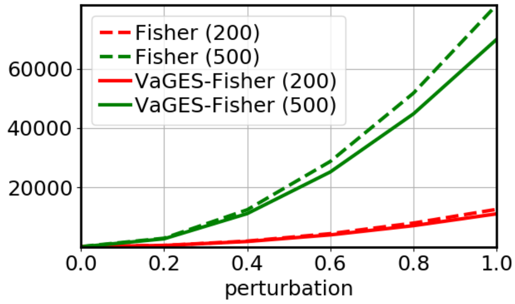


Figure 5. The Fisher divergence estimated by VaGES-Fisher v.s. the accurate Fisher divergence (Fisher).

Estimating $\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})$ with VaES, the variational stochastic gradient estimate is

$$\frac{1}{M}\sum_{i=1}^{M}\Big[\mathrm{VaES}(\boldsymbol{v}_i;\boldsymbol{\theta},\boldsymbol{\phi})^\top \frac{\partial \boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{v}_i)}{\partial \boldsymbol{\eta}} + \frac{\partial \boldsymbol{\epsilon}_i^\top \nabla_{\boldsymbol{v}} \boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{v}_i)\boldsymbol{\epsilon}_i}{\partial \boldsymbol{\eta}} - \frac{1}{2}\frac{\partial ||\boldsymbol{f}_{\boldsymbol{\eta}}(\boldsymbol{v}_i)||_2^2}{\partial \boldsymbol{\eta}}\Big], \quad (24)$$

where $\boldsymbol{v}_{1:M}$ is a mini-batch from $p_D(\boldsymbol{v})$ and $\boldsymbol{\epsilon}_{1:M}$ is a mini-batch from $p(\boldsymbol{\epsilon})$. We refer to this method as VaGES-Fisher.

5.1. Experiments

We validate the effectiveness of VaGES-Fisher in GRBMs. We initialize a GRBM as $p_D(\boldsymbol{v})$, perturb its weight with increasing noise and estimate the Fisher divergence between the initial GRBM and the perturbed one by VaGES-Fisher. The dimensions of $\boldsymbol{v}$ and $\boldsymbol{h}$ are the same and we experiment on dimensions of 200 and 500. See Appendix F.3 for more experimental details. The result is shown in Fig. 5. We compare the Fisher divergence estimated by VaGES-Fisher with the accurate Fisher divergence. Under both dimensions, our estimated Fisher divergence is close to the accurate one.

6. Related Work

Learning EBLVMs with MLE. A popular class of learning methods is the maximum likelihood estimate (MLE). Such methods have to deal with both the model posterior and the partition function. Some methods impose structural assumptions on the model. Among them, contrastive divergence (CD) (Hinton, 2002) and its variants (Tieleman, 2008; Qiu et al., 2020) estimate the gradient of the partition function with respect to the model parameters using Gibbs sampling (Geman & Geman, 1984) when the posterior can be sampled efficiently. Furthermore, when the model is fully visible, scalable methods (Du et al., 2018; Du & Mordatch, 2019; Nijkamp et al., 2020; 2019; Grathwohl et al., 2020a) are applicable by estimating such gradient with Langevin dynamics. In deep models with multiple layers of latent variables such as DBNs (Hinton et al., 2006; Lee et al., 2009) and DBMs (Salakhutdinov & Hinton, 2009), the greedy layer-wise learning algorithm (Hinton et al., 2006; Bengio et al., 2006) is applicable by exploring their hierarchical structures. Some recent methods (Kuleshov & Ermon, 2017; Li et al., 2020) attempt to learn general EBLVMs by variational inference. However, such methods suffer from high variance (Kuleshov & Ermon, 2017) or high bias (Li et al., 2020) of the variational bounds for the partition function in high-dimensional space.

Learning EBLVMs without partition functions. As mentioned in Sec. 1, the score matching (SM) is a popular class of learning methods, which eliminate the partition function by considering the score function, and the recent bi-level score matching (BiSM) is a scalable extension of SM to learn general EBLVMs. Prior to BiSM, extensions of SM (Swersky et al., 2011; Vértes et al., 2016) impose structural assumptions on the model, e.g., a tractable posterior (Swersky et al., 2011) or a jointly exponential family model (Vértes et al., 2016). Recently, Song & Ermon (2019; 2020) use a score network to directly model the score function of the data distribution. Incorporating latent variables to such a model is currently an open problem and we leave it as future work. The noise contrastive estimation (NCE) (Gutmann & Hyvärinen, 2010) is also a learning method which eliminates the partition function and variational noise contrastive estimation (VNCE) (Rhodes & Gutmann, 2019) is an extension of NCE to general EBLVMs based on variational methods. However, such methods have to manually design a noise distribution, which in principle should be close to the data distribution and is challenging in high-dimensional space. There are alternative approaches to handling the partition function, e.g., Doubly Dual Embedding (Dai et al., 2019a), Adversarial Dynamics Embedding (Dai et al., 2019b) and Fenchel Mini-Max Learning (Tao et al., 2019), and extending theses methods for learning EBLVMs would be interesting future work.

Evaluating EBLVMs. The log-likelihood is commonly used to compare fully visible models on the same dataset and the annealed importance sampling (Neal, 2001; Salakhutdinov, 2008) is often used to estimate the partition function in the log-likelihood. For a general EBLVM, a lower bound of the log-likelihood (Kingma & Welling, 2014; Burda et al., 2016) can be estimated by introducing a variational posterior together with estimating the partition function. The log-likelihood is a relative value and the corresponding absolute value (i.e., the KL divergence) is generally intractable since the entropy of the data distribution is unknown.

7. Conclusion

We propose new variational estimates of the score function and its gradient with respect to the model parameters in general energy-based latent variable models (EBLVMs). We rewrite the score function and its gradient as combinations of expectation and covariance terms over the model posterior. We approximate the model posterior with a variational posterior and analyze its bias. With samples drawn from the variational posterior, the expectation terms are estimated by Monte Carlo and the covariance terms are estimated by sample covariance matrix. We apply our estimates to kernelized Stein discrepancy (KSD), score matching (SM)-based methods and exact Fisher divergence. In particular, these estimates applied to SM-based methods are more time and memory efficient compared with the complex gradient unrolling in the strongest baseline bi-level score matching (BiSM) and meanwhile can scale up to natural images in deep EBLVMs. Besides, we also present interpolation results of annealed Langevin dynamics trajectories in the latent space in deep EBLVMs.

Software and Data

Codes in https://github.com/baofff/VaGES.

Acknowledgements

This work was supported by NSFC Projects (Nos. 61620106010, 62061136001, 61621136008, U1811461, 62076145), Beijing NSF Project (No. JQ19016), Beijing Academy of Artificial Intelligence (BAAI), Tsinghua-Huawei Joint Research Program, a grant from Tsinghua Institute for Guo Qiang, Tiangong Institute for Intelligent Computing, and the NVIDIA NVAIL Program with GPU/DGX Acceleration. C. Li was supported by the fellowship of China postdoctoral Science Foundation (2020M680572), and the fellowship of China national postdoctoral program for innovative talents (BX20190172) and Shuimu Tsinghua Scholar.

References

Bao, F., Li, C., Xu, T., Su, H., Zhu, J., and Zhang, B. Bi-level score matching for learning energy-based latent variable models. In *Advances in Neural Information Processing Systems 33*, 2020.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19*, pp. 153–160. MIT Press, 2006.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance weighted autoencoders. In *4th International Conference on Learning Representations*, 2016.

Dai, B., Dai, H., Gretton, A., Song, L., Schuurmans, D., and He, N. Kernel exponential family estimation via doubly dual embedding. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2321–2330. PMLR, 2019a.

Dai, B., Liu, Z., Dai, H., He, N., Gretton, A., Song, L., and Schuurmans, D. Exponential family estimation via adversarial dynamics embedding. In *Advances in Neural Information Processing Systems 32*, pp. 10977–10988, 2019b.

Dinh, L., Krueger, D., and Bengio, Y. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations*, 2015.

Du, C., Xu, K., Li, C., Zhu, J., and Zhang, B. Learning implicit generative models by teaching explicit ones. *arXiv preprint arXiv:1807.03870*, 2018.

Du, Y. and Mordatch, I. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems 32*, pp. 3603–3613, 2019.

Fan, J., Liao, Y., and Liu, H. An overview of the estimation of large covariance and precision matrices, 2016.

Gao, R., Nijkamp, E., Kingma, D. P., Xu, Z., Dai, A. M., and Wu, Y. N. Flow contrastive estimation of energy-based models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 7515–7525. IEEE, 2020.

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

Gong, W., Li, Y., and Hernández-Lobato, J. M. Sliced kernelized stein discrepancy. *arXiv preprint arXiv:2006.16531*, 2020.

Gorham, J. and Mackey, L. W. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1292–1301. PMLR, 2017.

Grathwohl, W., Wang, K., Jacobsen, J., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. In *8th International Conference on Learning Representations*. OpenReview.net, 2020a.

Grathwohl, W., Wang, K., Jacobsen, J., Duvenaud, D., and Zemel, R. S. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3732–3747. PMLR, 2020b.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Han, T., Nijkamp, E., Zhou, L., Pang, B., Zhu, S., and Wu, Y. N. Joint training of variational auto-encoder and latent energy-based model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 7975–7984. IEEE, 2020.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pp. 6626–6637, 2017.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation*, 18 (7):1527–1554, 2006.

Hu, T., Chen, Z., Sun, H., Bai, J., Ye, M., and Cheng, G. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

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

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.

Johnson, O. T. *Information theory and the central limit theorem*. World Scientific, 2004.

Kingma, D. P. and LeCun, Y. Regularized estimation of image statistics by score matching. In *Advances in Neural Information Processing Systems 23*, pp. 1126–1134. Curran Associates, Inc., 2010.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuleshov, V. and Ermon, S. Neural variational inference and learning in undirected graphical models. In *Advances in Neural Information Processing Systems 30*, pp. 6734–6743, 2017.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Lee, H., Grosse, R. B., Ranganath, R., and Ng, A. Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, volume 382 of *ACM International Conference Proceeding Series*, pp. 609–616. ACM, 2009.

Ley, C., Swan, Y., et al. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18, 2013.

Li, C., Du, C., Xu, K., Welling, M., Zhu, J., and Zhang, B. To relieve your headache of training an mrf, take advil. In *8th International Conference on Learning Representations*. OpenReview.net, 2020.

Li, Z., Chen, Y., and Sommer, F. T. Annealed denoising score matching: Learning energy-based models in high-dimensional spaces. *arXiv preprint arXiv:1910.07762*, 2019.

Liu, Q., Lee, J. D., and Jordan, M. I. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 276–284. JMLR.org, 2016.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems 33*, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 3730–3738. IEEE Computer Society, 2015.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. In *5th International Conference on Learning Representations*. OpenReview.net, 2017.

Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Nijkamp, E., Hill, M., Zhu, S., and Wu, Y. N. Learning non-convergent non-persistent short-run MCMC toward energy-based model. In *Advances in Neural Information Processing Systems 32*, pp. 5233–5243, 2019.

Nijkamp, E., Hill, M., Han, T., Zhu, S., and Wu, Y. N. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 5272–5280. AAAI Press, 2020.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Pang, T., Xu, K., Li, C., Song, Y., Ermon, S., and Zhu, J. Efficient learning of generative models via finite-difference score matching. In *Advances in Neural Information Processing Systems 33*, 2020.

Qiu, Y., Zhang, L., and Wang, X. Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *8th International Conference on Learning Representations*. OpenReview.net, 2020.

Rhodes, B. and Gutmann, M. U. Variational noise-contrastive estimation. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2741–2750. PMLR, 2019.

Salakhutdinov, R. Learning and evaluating boltzmann machines. *Utml Tr*, 2:21, 2008.

Salakhutdinov, R. and Hinton, G. Deep Boltzmann machines. In *Proceedings of the twelfth international conference on artificial intelligence and statistics*, 2009.

Saremi, S., Mehrjou, A., Schölkopf, B., and Hyvärinen, A. Deep energy estimator networks. *arXiv preprint arXiv:1805.08306*, 2018.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32*, pp. 11895–11907, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems 33*, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 574–584. AUAI Press, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Srivastava, N. and Salakhutdinov, R. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems 25*, pp. 2231–2239, 2012.

Swersky, K., Ranzato, M., Buchman, D., Marlin, B. M., and de Freitas, N. On autoencoders and score matching for energy based models. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1201–1208. Omnipress, 2011.

Tao, C., Chen, L., Dai, S., Chen, J., Bai, K., Wang, D., Feng, J., Lu, W., Bobashev, G. V., and Carin, L. On fenchel mini-max learning. In *Advances in Neural Information Processing Systems 32*, pp. 10427–10439, 2019.

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, volume 307 of *ACM International Conference Proceeding Series*, pp. 1064–1071. ACM, 2008.

Vértes, E., Gatsby Unit, U., and Sahani, M. Learning doubly intractable latent variable models via score matching. 2016.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688. Omnipress, 2011.

Welling, M., Rosen-Zvi, M., and Hinton, G. E. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*, pp. 1481–1488, 2004.

Xie, J., Lu, Y., Gao, R., Zhu, S.-C., and Wu, Y. N. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):27–45, 2018.