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

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

A. Proof of the Decomposition of the Gradient of the Score Function

Proof. Firstly we have

$$\begin{aligned}\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})\right] =&\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\nabla_{\boldsymbol{v}}p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\right]=\int p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})\frac{\nabla_{\boldsymbol{v}}p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\mathrm{d}\boldsymbol{h}\\ =&\int\nabla_{\boldsymbol{v}}p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})\mathrm{d}\boldsymbol{h}=\nabla_{\boldsymbol{v}}\int p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})\mathrm{d}\boldsymbol{h}=\nabla_{\boldsymbol{v}}1=\mathbf{0},\end{aligned}\tag{1}$$

and similarly we have $\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})\right]=\mathbf{0}$. Thereby, we have

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

and similarly we have $\nabla_{\boldsymbol{\theta}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})=\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{\theta}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]$.
Taking derivatives to Eqn. (1) w.r.t. $\boldsymbol{\theta}$, we have

$$\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}{\partial\boldsymbol{\theta}}\right]+\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})\frac{\partial\log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}{\partial\boldsymbol{\theta}}\right]=\mathbf{0}.\tag{2}$$

The second term in the left side of Eqn. (2) can be written as

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

[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 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Thereby, we have

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

B. The Tractability of the Commonly Used Divergences between Posteriors

The tractability of the KL divergence or the Fisher divergence between the variational posterior and the true posterior in EBLVMs has been shown by Bao et al. (2020) and we restate the results for completeness. Besides, we also analyze the tractability of the reverse KL divergence, the total variation distance (Pollard, 2005), the maximum mean discrepancy (Li et al., 2017) and the Wasserstein distance (Arjovsky et al., 2017) between the two posteriors in EBLVMs.

The KL divergence is tractable. The gradient of the KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ w.r.t. $\boldsymbol{\phi}$ is

$$\begin{aligned}
\nabla_{\boldsymbol{\phi}} \mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})) =&\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} = \nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) p_{\boldsymbol{\theta}}(\boldsymbol{v}) \mathcal{Z}(\boldsymbol{\theta})}{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})} \\
=&\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})} + \cancel{\nabla_{\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) \mathcal{Z}(\boldsymbol{\theta})} = \nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}.
\end{aligned}$$

The last term doesn't depend on the partition function or the marginal distribution of an EBLVM and thereby is tractable.

The Fisher divergence is tractable. The Fisher divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ is

$$\begin{aligned}
\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})) =&\frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}||\nabla_{\boldsymbol{h}} \log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) - \nabla_{\boldsymbol{h}} \log p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})||_2^2 \\
=&\frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}||\nabla_{\boldsymbol{h}} \log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) - \nabla_{\boldsymbol{h}} \log \frac{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{p_{\boldsymbol{\theta}}(\boldsymbol{v})\mathcal{Z}(\boldsymbol{\theta})}||_2^2 \\
=&\frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}||\nabla_{\boldsymbol{h}} \log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) - \nabla_{\boldsymbol{h}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}) - \cancel{\nabla_{\boldsymbol{h}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})\mathcal{Z}(\boldsymbol{\theta})}||_2^2 \\
=&\frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}||\nabla_{\boldsymbol{h}} \log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) - \nabla_{\boldsymbol{h}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})||_2^2.
\end{aligned}$$

Again, the last term doesn't depend on the partition function or the marginal distribution of an EBLVM and thereby is tractable. Furthermore, its gradient w.r.t. $\boldsymbol{\phi}$ is tractable.

The reverse KL divergence is generally intractable. The gradient of the reverse KL divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ w.r.t. $\boldsymbol{\phi}$ is $\nabla_{\boldsymbol{\phi}} \mathcal{D}_{RKL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})) = \nabla_{\boldsymbol{\phi}} \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} = -\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. Since $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ is generally intractable, the reverse KL divergence is generally intractable.

The total variation distance is generally intractable. The gradient of total variation distance (Pollard, 2005) between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ w.r.t. $\boldsymbol{\phi}$ is $\nabla_{\boldsymbol{\phi}} V(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}), p_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})) = \nabla_{\boldsymbol{\phi}} \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}|1 - \frac{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}|$. Since $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ is generally intractable, the total variation distance is generally intractable.

Maximum mean discrepancy is generally intractable. Given a kernel k, the gradient of the square of maximum mean discrepancy (Li et al., 2017) between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ w.r.t. $\boldsymbol{\phi}$ is

$$\begin{aligned}
\nabla_{\boldsymbol{\phi}} M_k(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}), p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})) &= \nabla_{\boldsymbol{\phi}}\big(\mathbb{E}_{\boldsymbol{h},\boldsymbol{h}'\sim q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} k(\boldsymbol{h},\boldsymbol{h}') + \mathbb{E}_{\boldsymbol{h},\boldsymbol{h}'\sim p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} k(\boldsymbol{h},\boldsymbol{h}') - 2\mathbb{E}_{\boldsymbol{h}\sim q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}),\boldsymbol{h}'\sim p_{\boldsymbol{\theta}}(\boldsymbol{h}'|\boldsymbol{v})} k(\boldsymbol{h},\boldsymbol{h}')\big) \\
&= \nabla_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{h},\boldsymbol{h}'\sim q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} k(\boldsymbol{h},\boldsymbol{h}') - 2\mathbb{E}_{\boldsymbol{h}'\sim p_{\boldsymbol{\theta}}(\boldsymbol{h}'|\boldsymbol{v})} \nabla_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{h}\sim q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} k(\boldsymbol{h},\boldsymbol{h}').
\end{aligned}$$

Since $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ is generally intractable, the maximum mean discrepancy is generally intractable.

The Wasserstein distance is generally intractable. The Wasserstein distance (Arjovsky et al., 2017) between $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ is $W(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}), p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})) = \frac{1}{K} \sup_{f:||f||_{Lip}\leq K} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} f(\boldsymbol{h}) - \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} f(\boldsymbol{h})$. Generally $\{f : ||f||_{Lip} \leq K\}$ is approximated by a neural network $f_{\boldsymbol{\eta}}$ with weight clipping and the Wasserstein distance is optimized by a bi-level optimization (Arjovsky et al., 2017). The lower level problem requires samples from the generally intractable posterior $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$. Thereby, the Wasserstein distance is generally intractable.

C. Proof of Theorem 1 and Theorem 2

Lemma 1. *Suppose P, Q are two probability measures on Ω, and $\boldsymbol{f} : \Omega \to \mathbb{R}^m$, then we have $||\mathbb{E}_P \boldsymbol{f} - \mathbb{E}_Q \boldsymbol{f}||_2 \leq ||\boldsymbol{f}||_\infty \sqrt{2\mathcal{D}_{KL}(Q||P)}$, where $||\boldsymbol{f}||_\infty \triangleq \sup_{\omega\in\Omega} ||\boldsymbol{f}(\omega)||_2$.*

Proof. Let $S = (P+Q)/2$, then P, Q are absolutely continuous w.r.t. S, and we have

$$\begin{aligned}||\mathbb{E}_P \boldsymbol{f} - \mathbb{E}_Q \boldsymbol{f}||_2 =& || \int \boldsymbol{f} \mathrm{d}P - \int \boldsymbol{f} \mathrm{d}Q||_2 = || \int \boldsymbol{f} \frac{\mathrm{d}P}{\mathrm{d}S} \mathrm{d}S - \int \boldsymbol{f} \frac{\mathrm{d}Q}{\mathrm{d}S} \mathrm{d}S||_2 = || \int \boldsymbol{f} (\frac{\mathrm{d}P}{\mathrm{d}S} - \frac{\mathrm{d}Q}{\mathrm{d}S}) \mathrm{d}S||_2 \\ \leq& \int ||\boldsymbol{f}||_2 \, |\frac{\mathrm{d}P}{\mathrm{d}S} - \frac{\mathrm{d}Q}{\mathrm{d}S}| \mathrm{d}S \leq ||\boldsymbol{f}||_\infty \int |\frac{\mathrm{d}P}{\mathrm{d}S} - \frac{\mathrm{d}Q}{\mathrm{d}S}| \mathrm{d}S.\end{aligned}$$

According to Pinsker's inequality (Tsybakov, 2008), we have $\int |\frac{\mathrm{d}P}{\mathrm{d}S} - \frac{\mathrm{d}Q}{\mathrm{d}S}| \mathrm{d}S \leq \sqrt{2\mathcal{D}_{KL}(Q||P)}$. Thereby, $||\mathbb{E}_P \boldsymbol{f} - \mathbb{E}_Q \boldsymbol{f}||_2 \leq ||\boldsymbol{f}||_\infty \sqrt{2\mathcal{D}_{KL}(Q||P)}$. □

Theorem 1. *(VaES, KL divergence) 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.*

Proof. According to Lemma 1, we have

$$||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] - \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] \, ||_2 \leq \sup_{\boldsymbol{h}} ||\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})||_2 \sqrt{2\mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.$$

By the boundedness of $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})$, $\exists A < \infty, \forall \boldsymbol{v}, \forall \boldsymbol{h}, \forall \boldsymbol{\theta}, ||\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})||_2 \leq A$. Let $C = \sqrt{2}A$, then

$$\begin{aligned}&||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] - \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})||_2 = ||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] - \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] \, ||_2 \\ \leq& A \sqrt{2\mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))} = C \sqrt{\mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.\end{aligned}$$

□

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

Lemma 2. *Suppose $\boldsymbol{a}$, $\boldsymbol{b}$ are two vectors, then $||\boldsymbol{a}\boldsymbol{b}^\top||_2 = ||\boldsymbol{a}||_2 ||\boldsymbol{b}||_2$.*

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

Theorem 2. *(VaGES, KL divergence) 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.*

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

$$||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] - \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})||_2 \leq C_1 \sqrt{\mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.$$

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

$$||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})} [\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})] - \nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})||_2 \leq C_2 \sqrt{\mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))},$$

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

$$||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right] - \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial \nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\right]||_2 \leq C_3\sqrt{\mathcal{D}_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))}.$$

By the boundedness of $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$ and $\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})$, $\nabla_{\boldsymbol{v}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\frac{\partial \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ is also bounded. Thereby, $\exists C_4 < \infty$, s.t.

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

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 $||\nabla_{\boldsymbol{\theta}} \log \tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})||_2$. Then, by the triangle inequality and Lemma. 2, we have

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

□

D. Proof of Theorem 3 and Theorem 4

Definition 2. *Suppose p is a probability density on $\mathbb{R}^n$ and $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$, we define $\mathcal{S}_p\boldsymbol{g}(\boldsymbol{x})\triangleq\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^\top\boldsymbol{g}(\boldsymbol{x})+\mathrm{Tr}(\nabla_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x}))$.*

Lemma 3. *(Liu & Wang, 2016) Suppose p is a probability density on $\mathbb{R}^n$ and $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$ is a function satisfying $\lim_{||\boldsymbol{x}||\to\infty}p(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})=\boldsymbol{0}$, then $\mathbb{E}_{p(\boldsymbol{x})}\left[\mathcal{S}_p\boldsymbol{g}(\boldsymbol{x})\right]=0$.*

Proof.

$$\boldsymbol{0}=\int\nabla_{\boldsymbol{x}}(p(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}=\int p(\boldsymbol{x})\nabla_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})+p(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^\top\mathrm{d}\boldsymbol{x}=\mathbb{E}_{p(\boldsymbol{x})}\left[\nabla_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{g}(\boldsymbol{x})\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^\top\right].$$

Thereby,

$$0=\mathrm{Tr}(\mathbb{E}_{p(\boldsymbol{x})}\left[\nabla_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{g}(\boldsymbol{x})\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^\top\right])=\mathbb{E}_{p(\boldsymbol{x})}\left[\mathrm{Tr}(\nabla_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x}))+\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})^\top\boldsymbol{g}(\boldsymbol{x})\right]=\mathbb{E}_{p(\boldsymbol{x})}\left[\mathcal{S}_p\boldsymbol{g}(\boldsymbol{x})\right].\ \square$$

Lemma 4. *Suppose p,q are probability densities on $\mathbb{R}^n$ and $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$ satisfies $\lim_{||\boldsymbol{x}||\to\infty}q(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})=\boldsymbol{0}$, we have*

$$|\mathbb{E}_q\mathcal{S}_p\boldsymbol{g}|\leq\sqrt{\mathbb{E}_{q(\boldsymbol{x})}||\boldsymbol{g}(\boldsymbol{x})||^2}\sqrt{D_F(q||p)}$$

Proof. By Lemma 3, we have $\mathbb{E}_q\mathcal{S}_q\boldsymbol{g}=0$. Thereby,

$$\begin{aligned}&|\mathbb{E}_q\mathcal{S}_p\boldsymbol{g}|=|\mathbb{E}_q\mathcal{S}_p\boldsymbol{g}-\mathbb{E}_q\mathcal{S}_q\boldsymbol{g}|=|\mathbb{E}_{q(\boldsymbol{x})}\boldsymbol{g}(\boldsymbol{x})^\top(\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})-\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x}))|\\ \leq&\mathbb{E}_{q(\boldsymbol{x})}||\boldsymbol{g}(\boldsymbol{x})||\ ||\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})-\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x})||\leq\sqrt{\mathbb{E}_{q(\boldsymbol{x})}||\boldsymbol{g}(\boldsymbol{x})||^2\mathbb{E}_{q(\boldsymbol{x})}||\nabla_{\boldsymbol{x}}\log p(\boldsymbol{x})-\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x})||^2}\\ =&\sqrt{\mathbb{E}_{q(\boldsymbol{x})}||\boldsymbol{g}(\boldsymbol{x})||^2}\sqrt{D_F(q||p)}.\end{aligned}\ \square$$

Definition 3. (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 $\mathcal{S}_p\boldsymbol{g}=f-\mathbb{E}_pf$.*

Remark. *The solution of the Stein equation exists. For example, let $h=f-\mathbb{E}_pf$, then*

$$g_1(\boldsymbol{x})=\frac{1}{p(\boldsymbol{x})}\int_{-\infty}^{x_1}p(t,x_2,\cdots,x_n)h(t,x_2,\cdots,x_n)\mathrm{d}t,\quad g_2(\boldsymbol{x})=\cdots=g_n(\boldsymbol{x})=0$$

is a solution.

Definition 4. *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 $\forall i\in\mathbb{Z}\cap[1,m]$, $\lim_{||\boldsymbol{x}||\to\infty}q(\boldsymbol{x})\boldsymbol{g}_{f_i}^p(\boldsymbol{x})=0$.*

Definition 5. (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 $\kappa_{\boldsymbol{f}}^{p,q}\triangleq\sqrt{\mathbb{E}_{q(\boldsymbol{x})}\sum_{i=1}^m||\boldsymbol{g}_{f_i}^p(\boldsymbol{x})||_2^2}$, referred to as the Stein factor of $\boldsymbol{f}$ w.r.t. p,q.*

Lemma 5. *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, then we have $||\mathbb{E}_q\boldsymbol{f}-\mathbb{E}_p\boldsymbol{f}||_2\leq\kappa_{\boldsymbol{f}}^{p,q}\sqrt{\mathcal{D}_F(q||p)}$.*

Proof. By Lemma 4, we have $|\mathbb{E}_qf_i-\mathbb{E}_pf_i|=|\mathbb{E}_q(f_i-\mathbb{E}_pf_i)|=|\mathbb{E}_q\mathcal{S}_p\boldsymbol{g}_{f_i}^p|\leq\sqrt{\mathbb{E}_{q(\boldsymbol{x})}||\boldsymbol{g}_{f_i}^p(\boldsymbol{x})||_2^2}\sqrt{D_F(q||p)}$.

Thereby, $||\mathbb{E}_q\boldsymbol{f}-\mathbb{E}_p\boldsymbol{f}||=\sqrt{\sum_{i=1}^n|\mathbb{E}_qf_i-\mathbb{E}_pf_i|^2}\leq\sqrt{\sum_{i=1}^n\mathbb{E}_{q(\boldsymbol{x})}||\boldsymbol{g}_{f_i}^p(\boldsymbol{x})||_2^2\ D_F(q||p)}=\kappa_{\boldsymbol{f}}^{p,q}\sqrt{D_F(q||p)}$. $\square$

Theorem 3. *(continuous $\boldsymbol{h}$, VaES) 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.*

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

Theorem 4. *(continuous $\boldsymbol{h}$, VaGES) 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.*

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

$$||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]-\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})||_2\leq C_1\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))},$$

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

$$||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{\theta}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]-\nabla_{\boldsymbol{\theta}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})||_2\leq C_2\sqrt{\mathcal{D}_F(q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})||p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v}))},$$

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

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

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

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

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}&||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial\boldsymbol{\theta}}\right]-\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\mathbb{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&||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\left(\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial\boldsymbol{\theta}}\right]-\mathbb{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(\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]-\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]\right)\mathbb{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\\ =&||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]||_2\,||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\frac{\partial\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{\partial\boldsymbol{\theta}}\right]-\mathbb{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\\ &+||\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]-\mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})\right]||_2\,||\mathbb{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&(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,

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

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})) + \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] - \frac{\partial \nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})}{\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})) + \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}_{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})) - \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] ||_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})) - \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 \\
&+ ||\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] - \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] ||_2 \\
\leq&(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}
$$

□

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

Theorem 5. *Suppose* $\lim_{||\boldsymbol{v}||\to\infty} p_D(\boldsymbol{v})(\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_D(\boldsymbol{v})) = \boldsymbol{0}$ *and* $\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] = \boldsymbol{I}$*, then* $\mathcal{D}_F^m(p_D||p_{\boldsymbol{\theta}}) = \mathcal{D}_F(p_D||p_{\boldsymbol{\theta}})$*, where* $\mathcal{D}_F(p_D||p_{\boldsymbol{\theta}}) = \frac{1}{2}\mathbb{E}_{p_D(\boldsymbol{v})}||\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} \log p_D(\boldsymbol{v})||_2^2$ *is the Fisher divergence between* p_D *and* $p_{\boldsymbol{\theta}}$*.*

Proof. By the assumption $\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right] = \boldsymbol{I}$, we have $\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v})\boldsymbol{\epsilon}\right] = \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v}))$. Thereby, $\mathcal{D}_F^m(p_D||p_{\boldsymbol{\theta}}) = \max_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_{p_D(\boldsymbol{v})}\left[\nabla_{\boldsymbol{v}} \log p_{\boldsymbol{\theta}}(\boldsymbol{v})^\top \boldsymbol{f}(\boldsymbol{v}) + \mathrm{Tr}(\nabla_{\boldsymbol{v}} \boldsymbol{f}(\boldsymbol{v})) - \frac{1}{2}||\boldsymbol{f}(\boldsymbol{v})||_2^2\right]$.

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

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

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

F. Additional Experimental Details

F.1. Learning EBLVMs with KSD

Additional setting. We generate 60,000 samples for training and 10,000 samples for testing on checkerboard. The dimension of $\boldsymbol{h}$ is 4. We use the Adam optimizer and the learning rate is 10^{-3}. We train 100,000 iterations and the batch size is 100. The log-likelihood is estimated by annealed importance sampling (Salakhutdinov & Murray, 2008), where we use 2,000 samples and 2,000 middle states to estimate the log-partition function. We run 1,000 steps Gibbs sampling to sample from GRBMs. The variational parameter $\boldsymbol{\phi}$ is updated for $K = 5$ times on each minibatch.

F.2. Learning EBLVMs with Score Matching

F.2.1. Comparison in GRBMs

Following BiSM (Bao et al., 2020), we split 1,400 images for training, 300 images for validation and 265 images for testing on Frey face[1]; we generate 60,000 samples for training and 10,000 samples for testing on checkerboard; the dimension of $\boldsymbol{h}$ is 400 on Frey face and 4 on checkerboard; we use the Adam optimizers with learning rates 2×10^{-4} for training on Frey face and 10^{-3} for training on checkerboard; we train 20,000 iterations on Frey face and 100,000 iterations on checkerboard; the batch size is 100 for both datasets; we select the best model trained on Frey face according to the validation log-likelihood. The log-likelihood is estimated by annealed importance sampling (Salakhutdinov & Murray, 2008), where we use 2,000 samples and 2,000 middle states to estimate the log-partition function. We run 1,000 steps Gibbs sampling to sample from GRBMs. The time comparison on Frey face is conducted on 1 GeForce GTX 1080 Ti GPU with 2,000 training iterations. The number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is $L = 2$ and the variational parameter $\boldsymbol{\phi}$ is updated for $K = 5$ times on each minibatch by default.

F.2.2. Learning Deep EBLVMs

Additional setting. Following BiSM (Bao et al., 2020), we split 60,000 samples for training on MNIST, 50,000 samples for training on CIFAR10 and 182,637 samples for training on CelebA; the dimension of $\boldsymbol{h}$ is 50; we use the Adam optimizers with learning rates 10^{-4} for training on MNIST and 5×10^{-5} for training on CIFAR10 and CelebA; we train 100,000 iterations on MNIST and 300,000 iterations on CIFAR10 and CelebA; the batch size is 100 for all datasets. g_1 consists of a $6k$-layer ResNet, where $k = 2, 3, 3$ for MNIST, CIFAR10 and CelebA respectively and g_3 is an MLP containing one fully connected layer. The structure of the deep EBLVM is shown in Fig. 1. The number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is $L = 2$ and the variational parameter $\boldsymbol{\phi}$ is updated for $K = 5$ times on each minibatch. Following a similar protocol with Song & Ermon (2019); Li et al. (2019); Bao et al. (2020), we save one checkpoint every 5000 iterations and select the best CIFAR10 and CelebA models according to the FID score on 1000 samples. The FID score reported in Section 4.2 in the full paper is estimated on 50,000 samples using the official code[2].

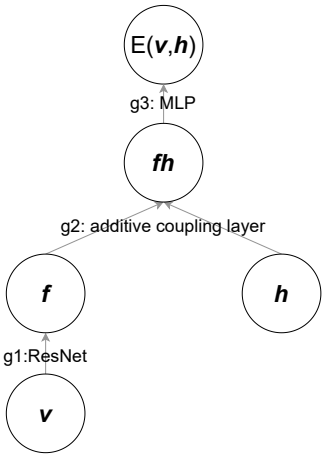


Figure 1. The structure of the deep EBLVM.

Hyperparamter selection. Since we compare with BiSM (Bao et al., 2020), we use the same hyperparameters as BiSM when they can be shared (e.g., the model types and structures, the divergence to learn $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, the dimensions of $\boldsymbol{h}$, the batch size, the optimizers and corresponding learning rates). As for L (the number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, we find it enough to set it to 2 (the minimal number of samples required in a sample covariance matrix) in our considered models. As for K, we set it to 5, so that it will ensure the convergence of training $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and meanwhile have an acceptable computation cost. As for the step size and the standard deviation of the noise in Langevin dynamics, we grid search the optimal one, as shown in Tab. 1. We find that the optimal step size is approximately proportional to the dimension of $\boldsymbol{h}$, perhaps because Langevin dynamics converges to its stationary distribution slower when $\boldsymbol{h}$ has a higher dimension. The standard deviation can work in range $[10^{-4}, 10^{-2}]$.

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

[2] https://github.com/bioinf-jku/TTUR

Table 1. Grid search of the step size and the standard deviation of the noise in Langevin dynamics under different dimensions of $\boldsymbol{h}$. We use the CIFAR10 dataset. The result is represented by the FID score on 1000 samples.

(a) dimension($\boldsymbol{h}$) = 20

	10^{-4}	10^{-3}	10^{-2}
1×10^{-3}	diverge	diverge	diverge
2×10^{-3}	56.42	54.87	54.46
$\mathbf{4 \times 10^{-3}}$	55.90	58.07	54.55
6×10^{-3}	56.48	55.09	57.03
10×10^{-3}	diverge	diverge	diverge

(b) dimension($\boldsymbol{h}$) = 50

	10^{-4}	10^{-3}	10^{-2}
2.5×10^{-3}	diverge	diverge	diverge
5×10^{-3}	56.74	57.55	58.19
$\mathbf{10 \times 10^{-3}}$	55.74	54.71	58.01
15×10^{-3}	58.03	60.52	56.18
25×10^{-3}	diverge	diverge	diverge

(c) dimension($\boldsymbol{h}$) = 100

	10^{-4}	10^{-3}	10^{-2}
10×10^{-3}	diverge	diverge	diverge
$\mathbf{20 \times 10^{-3}}$	56.57	60.04	56.31
30×10^{-3}	diverge	diverge	diverge

Sampling. Since we compare with BiSM (Bao et al., 2020), we use the same sampling algorithm as BiSM. For deep EBLVMs, we first randomly select a training data point and inference its approximate posterior mean; we then sample from $p(\boldsymbol{v}|\boldsymbol{h})$ with $\boldsymbol{h}$ equal to the approximate posterior mean using the annealed Langevin dynamics technique (Li et al., 2019). The temperature range is $[1, 100]$ and the step size is 0.02 in annealed Langevin dynamics.

Devices and training time. The time and memory consumption of VaGES with different batch sizes (BS) is displayed in Tab. 2. We also include that of BiSM. The time and memory consumption in deep EBLVMs of VaGES and BiSM is consistent with GRBMs (see Fig. 3 in the full paper). Besides, training an EBLVM takes about 2.8 times as long as training an EBM (the one trained by MDSM in Tab. 1 (a) in the full paper). The additional cost is reasonable, since (i) the EBLVM improves the expressive power (see Tab. 1 (a) in the full paper) based on a similar model structure and a comparable amount of parameters (244MB for the EBLVM and 238MB for the EBM), and (ii) the EBLVM enables manipulation in the latent space (see Fig. 4 in the full paper).

Table 2. Training time of 2k iterations/memory consumption on GeForce RTX 2080 Ti in deep EBLVMs. L=2,K=5.

Dataset	Setting	VaGES	BiSM (N=0)	BiSM (N=2)	BiSM (N=5)
CIFAR10	1GPU BS=64	24m/5.8GB	19m/6.8GB	25m/7.8GB	34m/9.5GB
	2GPUs BS=100	35m/8.2GB	28m/11.1GB	43m/13.2GB	66m/16.4GB
CelebA	1GPU BS=16	26m/8.0GB	21m/7.9GB	26m/9.2GB	35m/10.2GB
	6GPUs BS=100	90m/52.4GB	67m/51.9GB	100m/58.7GB	124m/60.1GB

F.3. Evaluating EBLVMs with Exact Fisher Divergence

Additional setting. The GRBM is initialized as a standard Gaussian distribution by letting $\boldsymbol{b} = \mathbf{0}, \boldsymbol{c} = \mathbf{0}, W = \mathbf{0}, \sigma = 1$, 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. We use the Adam optimizer and the learning rate is 2×10^{-4}. We train 20,000 iterations and the batch size is 100. The number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is $L = 1$ and the variational parameter $\boldsymbol{\phi}$ is updated for $K = 5$ times on each minibatch. $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ 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. $\mathcal{D}$ is the KL divergence to learn $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$. $\boldsymbol{f_\eta}$ is a multilayer perceptron (MLP) with 2 hidden layers and each layer has the same width.

F.4. Numerical Validation of Theorems

In the two posteriors $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$, we fix $\boldsymbol{v}$, $\boldsymbol{\theta}$ and only vary $\boldsymbol{\phi}$ to plot the relationship between the biases and the divergences. As for $\boldsymbol{v}$, we randomly select a sample from the Frey face training dataset and fix $\boldsymbol{v}$ as the sample. As for

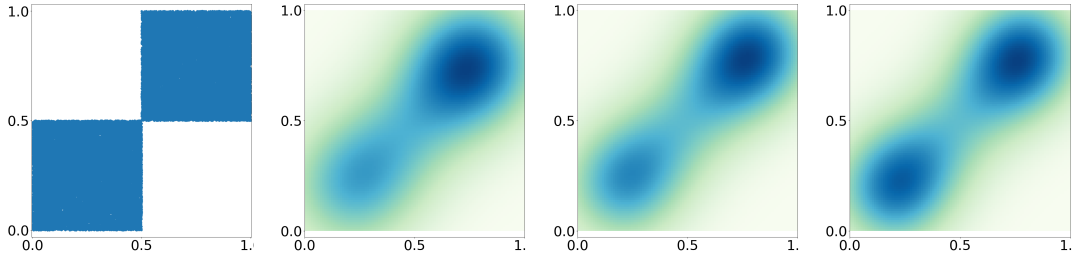
$\boldsymbol{\theta}$, we randomly initialize it with the uniform noise and don't change it anymore. As for $\boldsymbol{\phi}$, we initialize it such that the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is equal to the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$. After initialization, we perturb $\boldsymbol{\phi}$ with an increasing Gaussian noise level and record the corresponding biases and divergences.

The dimension of $\boldsymbol{h}$ is 400. As for the GRBM, $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a Bernoulli distribution parametermized by a fully connected layer with the sigmoid activation. As for the Gaussian model, $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is a Gaussian distribution parametermized by a fully connected layer.

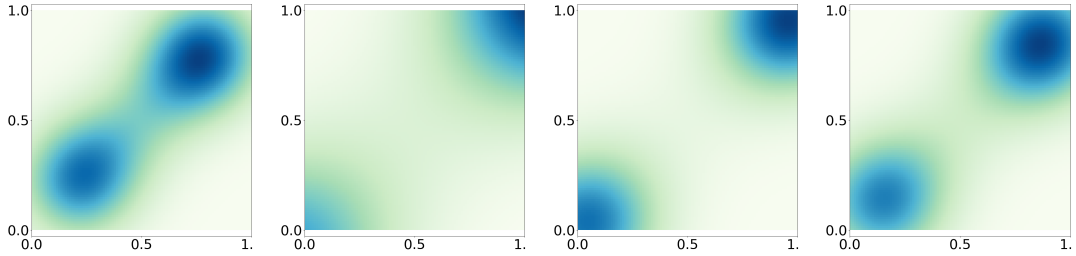
G. Additional Results

G.1. Learning EBLVMs with KSD

The density of the checkerboard dataset is shown in Fig. 2 (a). The densities of GRBMs learned by KSD, VaGES-KSD and IS-KSD are shown in Fig. 2 (b-h). Our VaGES-KSD is comparable to the KSD baseline and is better than the IS-KSD baseline. The result is consistent with the test log-likelihood results in Figure 2 in the full paper.



(a) Data density (b) VaGES-KSD (L=2) (c) VaGES-KSD (L=5) (d) VaGES-KSD (L=10)



(e) KSD (f) IS-KSD (L=2) (g) IS-KSD (L=5) (h) IS-KSD (L=10)

Figure 2. Density plots of GRBMs trained by KSD, VaGES-KSD and IS-KSD. L is the number of samples from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$.

G.2. Learning EBLVMs with Score Matching

G.2.1. Comparison in GRBMs

We compare with DSM (Vincent, 2011), BiSM (Bao et al., 2020), CD-based methods (Hinton, 2002; Tieleman, 2008) and noise contrastive estimation (NCE)-based methods (Gutmann & Hyvärinen, 2010; Rhodes & Gutmann, 2019) on the checkerboard dataset. In Fig. 3, we plot the test log-likelihood of different methods under the same setting. The result of VaGES-DSM is similar to CD, DSM and BiDSM and slightly better than PCD and NCE-based methods after convergence. The convergence speed of VaGES-DSM is faster than BiDSM. Besides, we show the densities of GRBMs learned by these methods in Fig. 4. The performance of VaGES-DSM is similar to CD, DSM and BiDSM and better than PCD, NCE and VNCE, which agrees with the test log-likelihood results after convergence in Fig. 3.

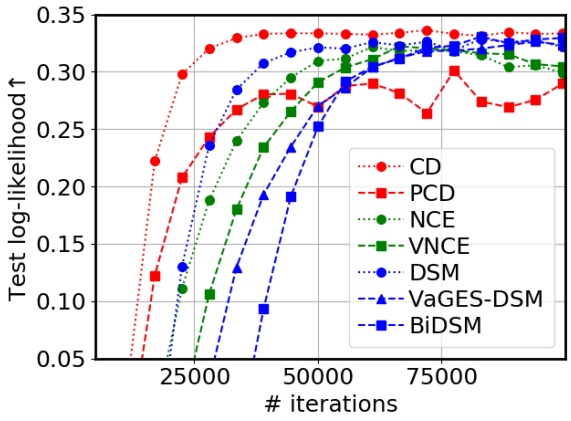


Figure 3. Comparison of different methods on checkerboard. The test log-likelihood is averaged over 10 runs.

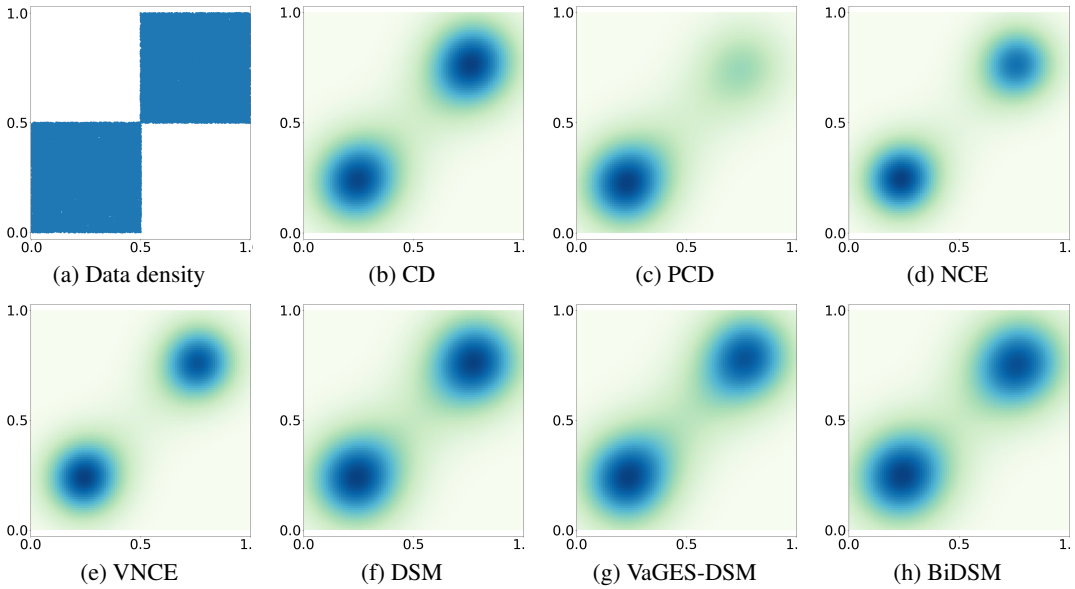
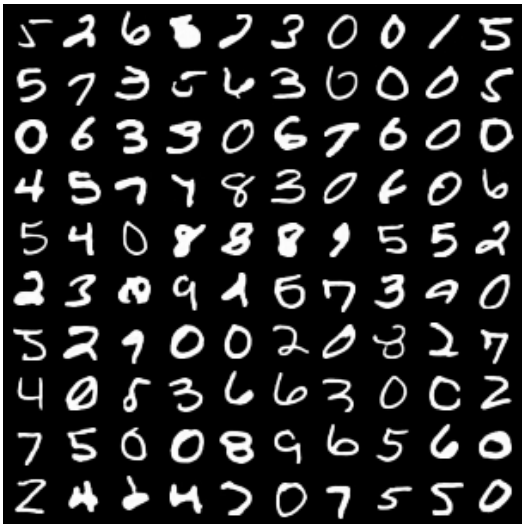


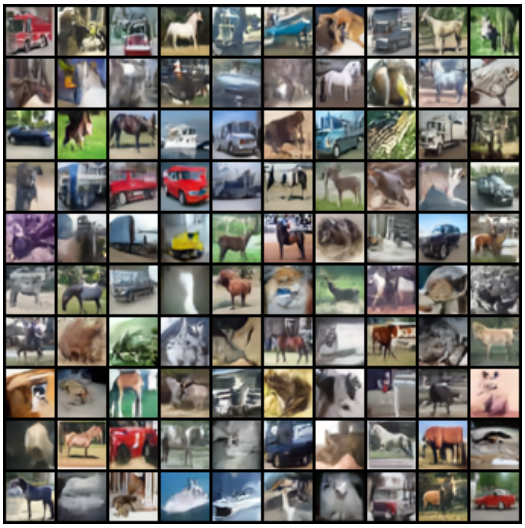
Figure 4. Density plots of GRBMs trained by different methods on checkerboard.

G.2.2. Learning Deep EBLVMs

Sample quality. We show samples from EBLVMs learned on MNIST, CIFAR10 and CelebA in Fig. 5. We also evaluate the Inception Score on CIFAR10 and VaGES gets 7.53, which is better than baselines such as VAE-EBLVM (Han et al., 2020) (7.17) and EBM (Du & Mordatch, 2019) (6.78).



(a) MNIST

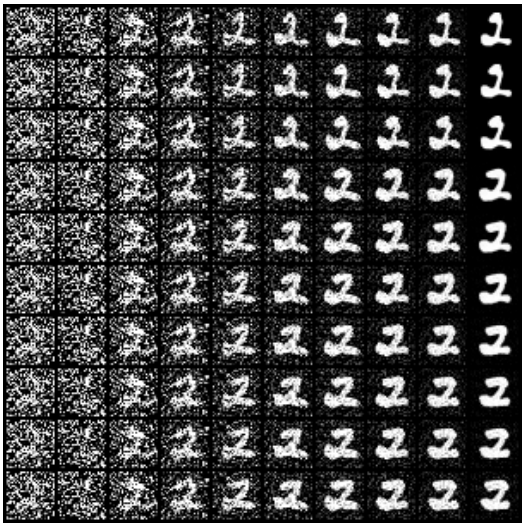


(b) CIFAR10

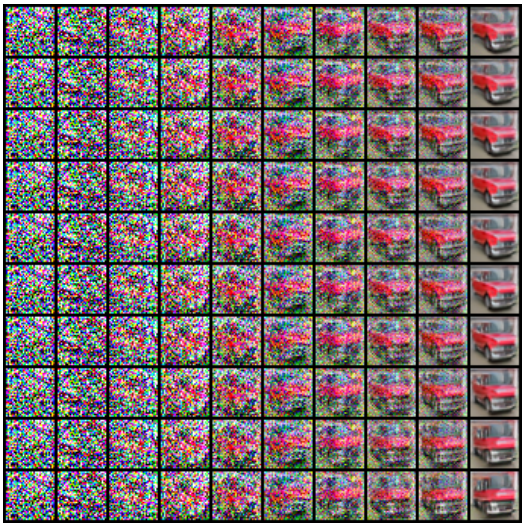
(c) CelebA

Figure 5. Samples from EBLVMs.

Interpolation in the latent space. We show more interpolation results in Fig. 6.



(a) MNIST



(b) CIFAR10

(c) CelebA

Figure 6. Interpolation of annealed Langevin dynamics trajectories in the latent space in EBLVMs.

Sensitivity analysis. We study how hyperparameters influence the performances of VaGES-SM in deep EBLVMs. The result is shown in Tab. 3. Increasing the number of convolutional layers will improve the performance, while the dimension of $\boldsymbol{h}$, the number of $\boldsymbol{h}$ sampled from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ and the noise level in Langevin dynamics don't affect the result very much. Setting both the number of times updating $\boldsymbol{\phi}$ and the number of Langevin dynamics steps to 5 is enough for a stable training. Besides, we also try using the KL divergence to learn the variational posterior and get a FID of 29.13, which doesn't affect the result very much.

Table 3. Sensitivity analysis on different hyperparameters (evaluated by FID ↓ on CIFAR10). Div means the training diverges.

(a) Dimensions of $\boldsymbol{h}$

	20	50	100
FID	26.55	28.09	27.78

(b) # convolutional layers

	12	18	24
FID	36.17	28.09	25.98

(c) # $\boldsymbol{h}$ sampled from $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$

	2	5	10
FID	28.09	31.21	29.26

(d) Noise level in Langevin dynamics

	10^{-4}	10^{-3}	10^{-2}
FID	28.09	25.83	30.47

(e) # times updating $\boldsymbol{\phi}$ (K)
Langevin dynamics steps (C)

K \ C	0	5	10	15
0	FID	Div	Div	Div
5	Div	28.09	Div	Div
10	Div	Div	29.52	27.88
15	Div	Div	Div	Div

H. Additional Attempts on Improving Estimates

We can directly apply the control variate technique (Owen, 2013) to VaES. By noticing that

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\nabla_{\boldsymbol{v}}\log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}) = \int q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})\frac{\nabla_{\boldsymbol{v}}q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}{q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})}\mathrm{d}\boldsymbol{h} = \nabla_{\boldsymbol{v}}\int q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})\mathrm{d}\boldsymbol{h} = \boldsymbol{0}, \tag{3}$$

we can subtract $\nabla_{\boldsymbol{v}}\log q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ from VaES without changing the value of the expectation, and the resulting variational estimate is

$$\text{VaES-CV}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{1}{L}\sum_{i=1}^{L}\nabla_{\boldsymbol{v}}\log\frac{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h}_i)}{q_{\boldsymbol{\phi}}(\boldsymbol{h}_i|\boldsymbol{v})}, \quad \boldsymbol{h}_i \overset{\text{i.i.d}}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v}). \tag{4}$$

When the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{h}|\boldsymbol{v})$ is equal to the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})$, $\text{VaES-CV}(\boldsymbol{v};\boldsymbol{\theta},\boldsymbol{\phi}) = \nabla_{\boldsymbol{v}}\log\frac{\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}{p_{\boldsymbol{\theta}}(\boldsymbol{h}|\boldsymbol{v})} = \nabla_{\boldsymbol{v}}\log\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{v})$ is exactly equal to the score function and has zero bias and variance. Empirically, we study how the control variate influences the performance over different objectives on the checkerboard dataset in GRBMs. As shown in Fig. 7, the control variate only marginally improves the performance of VaGES-KSD and makes no difference to VaGES-DSM. As a result, we don't make the control variate a default technique in VaES, since it will introduce some extra computation, while the improvement is marginal.

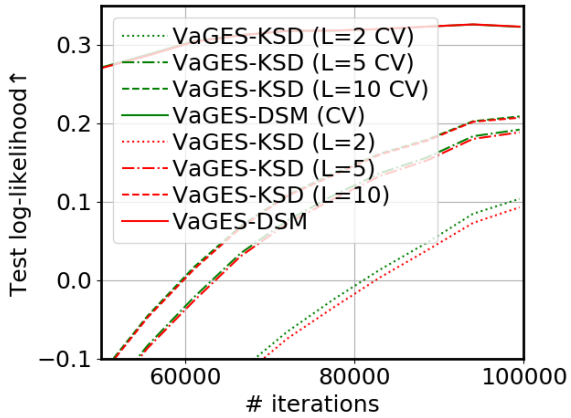


Figure 7. How the control variate (CV) influences the performance over different objectives on the checkerboard dataset in GRBMs. The test log-likelihood is averaged over 10 runs.

I. An Introduction to BiSM

BiSM (Bao et al., 2020) approximates the score function via variational inference first:

$$\nabla_{\boldsymbol{v}} \log p(\boldsymbol{v};\boldsymbol{\theta}) = \nabla_{\boldsymbol{v}} \log \frac{\tilde{p}(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})} - \nabla_{\boldsymbol{v}} \log \mathcal{Z}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{v}} \log \frac{\tilde{p}(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})},$$

and then gets the gradient of a certain objective via solving a complicated bi-level optimization problem:

$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{J}_{Bi}(\boldsymbol{\theta},\boldsymbol{\phi}^*(\boldsymbol{\theta})), \quad \mathcal{J}_{Bi}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{v},\boldsymbol{\epsilon})}\mathbb{E}_{q(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\phi})}\mathcal{F}\left(\nabla_{\boldsymbol{v}} \log \frac{\tilde{p}(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}{q(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\phi})}, \boldsymbol{\epsilon}, \boldsymbol{v}\right),$$

where Θ is the hypothesis space of the model, $\mathcal{F}$ depends on the certain objective, $q(\boldsymbol{v},\boldsymbol{\epsilon})$ is the joint distribution of the data and additional noise and $\boldsymbol{\phi}^*(\boldsymbol{\theta})$ is defined as follows:

$$\boldsymbol{\phi}^*(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\phi}\in\Phi} \mathcal{G}(\boldsymbol{\theta},\boldsymbol{\phi}), \text{ with } \mathcal{G}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{v},\boldsymbol{\epsilon})}\mathcal{D}\left(q(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\phi})||p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta})\right).$$

BiSM uses gradient unrolling to solve the problem, where the lower level problem $\boldsymbol{\phi}^*(\boldsymbol{\theta})$ is approximated by the output of N steps gradient descent on $\mathcal{G}(\boldsymbol{\theta},\boldsymbol{\phi})$ w.r.t. $\boldsymbol{\phi}$, which is denoted by $\boldsymbol{\phi}^N(\boldsymbol{\theta})$. Finally, the model is updated with the approximate gradient $\nabla_{\boldsymbol{\theta}}\mathcal{J}_{Bi}(\boldsymbol{\theta},\boldsymbol{\phi}^N(\boldsymbol{\theta}))$, whose bias converges to zero in a linear rate in terms of N when G is strongly convex. The gradient unrolling requires an $O(N)$ time and memory.

Gradient unrolling of small steps is of large bias and that of large steps is time and memory consuming. Thus, BiSM with 0 gradient unrolling suffers from an additional bias besides the variational approximation. Instead, VaGES directly approximates the gradient of score function and its bias is controllable as presented in Sec. 2.2 in the full paper.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017.

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.

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.

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.

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

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

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

Li, C., Chang, W., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30*, pp. 2203–2213, 2017.

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. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems 29*, pp. 2370–2378, 2016.

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

Pollard, D. *Inequalities for probability and statistics*. 2005.

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. and Murray, I. On the quantitative analysis of deep belief networks. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, volume 307 of *ACM International Conference Proceeding Series*, pp. 872–879. ACM, 2008.

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.

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.

Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

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