Fisher Auto-Encoders

Khalil Elkhalil[1] **Ali Hasan**[1] **Jie Ding**[2] **Sina Farsiu**[1] **Vahid Tarokh**[1]
[1]Duke University, USA [2]University of Minnesota, USA

Abstract

It has been conjectured that the Fisher divergence is more robust to model uncertainty than the conventional Kullback-Leibler (KL) divergence. This motivates the design of a new class of robust generative auto-encoders (AE) referred to as Fisher auto-encoders. Our approach is to design Fisher AEs by minimizing the Fisher divergence between the intractable joint distribution of observed data and latent variables, with that of the postulated/modeled joint distribution. In contrast to KL-based variational AEs (VAEs), the Fisher AE can exactly quantify the distance between the true and the model-based posterior distributions. Qualitative and quantitative results are provided on both MNIST and celebA datasets demonstrating the competitive performance of Fisher AEs in terms of robustness compared to other AEs such as VAEs and Wasserstein AEs.

1 Introduction

In recent years, generative modeling became a very active research area with impressive achievements. The most popular generative schemes are often given by variational auto-encoders (VAEs) (Kingma and Welling, 2014), generative adversarial networks (GANs) (Goodfellow et al., 2014) and their variants. VAEs rely on the maximum likelihood principle to learn the underlying data generating distribution by considering a parametric model. Due to the intractability of the parametric model, VAEs employ approximate inference by considering an approximate posterior to get a variational bound on the log-likelihood of the model distribution. Despite its elegance, this approach has the drawback of generating low-quality samples due to the fact that the approximate posterior could be quite different from the true one. On the other hand, GANs have proven to be more impressive when it comes to the visual quality of the generated samples, while the training often involves nontrivial fine-tuning and is unstable. In addition to difficult training, GANs also suffer from "mode collapse" where the generated samples are not diverse enough to capture the diversity and variability in the true data distribution (Goodfellow et al., 2014).

In this work, we propose a new class of robust auto-encoders that also serve as a generative model. The main idea is to develop a 'score' function (Hyvärinen, 2005; Parry et al., 2012) of the observed data and postulated model, so that its minimization problem is equivalent to minimizing the Fisher divergence (Ding et al., 2019) between the underlying data generating distribution and the postulated/modeled distribution. By doing this, we are able to leverage the potential advantages of Fisher divergence in terms of computation and robustness. In the context of parameter estimation, minimizing the Fisher divergence has led to the Hyvärinen score (Hyvärinen, 2005), which serves as a potential surrogate for the logarithmic score. The main advantage of the Hyvärinen score over logarithmic sore is its significant computational advantage for estimating probability distributions that are known only up to a multiplicative constant, e.g. those in mixture models and complex time series models (Hyvärinen, 2005; Parry et al., 2012; Ding et al., 2019; Shao et al., 2019). Our work will extend the use of Fisher divergence and Hyvärinen score in the context of variational auto-encoders.

Similar to the logarithmic score, the Hyvärinen score is also intractable to compute due to the intractable integration over the latent variables. One way to mitigate this difficulty is to bound the Hyvärinen score and obtain a variational bound to optimize instead. However, unlike the logarithmic score, this strategy seems to be very complicated and a variational bound seems to be out of reach. Alternatively, it turns out that the variational bound in VAEs can be recovered by minimizing the KL divergence between the joint distribution over the data and latent variable and the modeled joint distribution which can be easily calculated as the product of the prior and the decoder distribution (Kingma and Welling, 2019). Following the same principle, we propose to minimize the Fisher divergence between the two joint distributions over the model parameters. This minimization results in a loss function that shares similar properties as regular VAEs but more powerful from an inference point of view.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

It turns out that our developed loss function is the sum of three terms: the first one is the tractable Fisher divergence between the approximate and the model posteriors, the second is similar to the reconstruction loss in VAEs obtained by evaluating the Hyvärinen score on the decoder distribution, and the last term can be seen as a stability measure that promotes the invariance property in feature extraction in the encoder. Therefore, the new loss function is different from the regular variational bound in regular VAEs in the following aspects: 1) it considers the minimization of the distance between the approximate and the model posteriors which turns out to be difficult when considering the KL divergence due to the intractable normalization constant in the model posterior, 2) it allows to produce robust features by considering a stability measure of the approximate posterior. Experimental results on MNIST (LeCun and Cortes, 2010) and CelebA (Liu et al., 2015) datasets validate these aspects and demonstrate the potential of the proposed Fisher AE as compared to some existing schemes such as VAEs and Wasserstein AEs. Moreover, thanks to the stability measure in the Fisher loss function, the encoder is proved to have more stable and robust reconstruction when the data is perturbed by noise as compared to other schemes playing a similar role as denoising auto-encoders (Vincent, 2011).

Related works. Previous works on learning variational auto-encoders initiated by the work of (Kingma and Welling, 2014) are fundamentally maximum likelihood methods that learn the underlying data distribution by the proxy of an evidence lower bound (ELBO) on the log-likelihood. The accuracy of such bound is mainly related to the KL divergence between the true and the postulated posteriors. This has been the focus of many works trying to minimize the inference gap resulting from the postulated posterior. For instance, normalizing flows (Rezende and Mohamed, 2015) employs rich posterior approximations using tractable and flexible transformations on initial densities. In the same category, the work in (Pu et al., 2017) provides an efficient way of directly sampling from the true posterior using the Stein Variational Gradient Descent (SVGD) method. On the other hand, Wasserstein auto-encoders (WAEs) proposed in (Ilya et al., 2018) follow a different path by looking at the Wasserstein distance between the true and the model distributions. Relying on the Monge-Kantorovich formulation, the Wasserstein distance naturally emerges as an optimization over an encoder-decoder structure with a reasonable geometry over the latent manifold.

Main contributions. First, we develop a new type of AEs that is based on minimizing the Fisher divergence between the underlying data/latent joint distribution and the postulated model joint distribution. Our derived loss function may be decomposed as *divergence between posteriors* + *reconstruction loss* + *stability measure*. Second, our derived method is conceptually appealing as it is reminiscent of the classical evidence lower bound (ELBO) derived from Kullback-Leibler (KL) divergence. Third, we affirmatively address the conjecture made in some earlier work that Fisher divergence can be more robust than KL divergence in modeling complex nonlinear models (Ding et al., 2019; Lyu, 2009) in the context of VAEs. Our results indicate that Fisher divergence may serve as a competitive learning machinery for challenging deep learning tasks.

Outline. In Section 2, we provide a brief overview on VAEs and some theoretical concepts related to the Fisher divergence and the Hyvärinen score. In Section 3, we provide the technical details related to the proposed Fisher auto-encoder. Then, in Section 4 we give both qualitative and quantitative results regarding the performance of the proposed Fisher AE. Finally, we provide some concluding remarks in Section 5.

2 Background on VAEs and Fisher divergence

2.1 Variational auto-encoders

By considering a probabilistic model of the data observations $\mathbf{x} \in \mathbb{R}^D$ given by $p_\theta(\mathbf{x})$, the goal of variational inference is to optimize the model parameters θ to match the true unknown data distribution $p_\star(\mathbf{x})$ in some sense. One way to match the true data distribution is to minimize the Kullback-Leibler (KL) divergence as follows:

$$\begin{aligned}\theta^\star &= \arg\min_\theta \mathbb{D}_{\text{KL}}\left[p_\star || p_\theta\right] \\ &= \arg\min_\theta \mathbb{E}_{p_\star(\mathbf{x})} - \log p_\theta(\mathbf{x}) \\ &= \arg\min_\theta \mathbb{E}_{p_\star(\mathbf{x})} - \log \int p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z},\end{aligned} \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^d$ are latent variables with prior distribution $p(\mathbf{z})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ is a likelihood function corresponding to the decoder modeled by the parameters θ using a neural network. Unfortunately, the intergration over the latent variables $\mathbf{z}$ in (1) is usually intractable and an upper bound on the negative marginal log-likelihood is often optimized instead. By introducing an alternative posterior over the latent variables given by $q_\phi(\mathbf{z}|\mathbf{x})$ and by direct application of the Jensen's inequality, we have

$$\begin{aligned}-\log p_\theta(\mathbf{x}) &= -\log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z} \\ &\leq \mathbb{D}_{\text{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) \\ &= \mathcal{L}_{\text{VAE}}\left(\mathbf{x}; \phi, \theta\right),\end{aligned} \tag{2}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is an approximate posterior corresponding to the encoder parameterized by ϕ. The bound in (2) is often called the evidence lower bound (ELBO) (w.r.t the log-likelihood) and it is optimized w.r.t both model parameters ϕ and θ:

$$\phi^*, \theta^* = \arg\min_{\phi,\theta} \mathbb{E}_{p_\star(\mathbf{x})} \mathcal{L}_{\text{VAE}}\left(\mathbf{x}; \phi, \theta\right). \tag{3}$$

The common practice is to consider a Gaussian model for the posterior $q_\phi(\mathbf{z}|\mathbf{x})$, i.e., $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\mu(\mathbf{x}), \sigma(\mathbf{x})^2\right)$ where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})^2$ are the output of a neural network taking as input the data sample $\mathbf{x}$ and parameterized by ϕ. This allows to reparametrize $\mathbf{z}$ as $\mathbf{z} = \mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \epsilon$, where $\odot$ denotes the point-wise multiplication and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ which permits to efficiently solve (3) using stochastic gradient variational Bayes (SGVB) as in (Kingma and Welling, 2014).

2.2 Fisher divergence and the Hyvärinen score

A standard procedure in data fitting and density estimation is to select from a parameter space Θ, the probability distribution p_θ, $\theta \in \Theta$ that minimizes a certain divergence $\mathbb{D}[.||.]$ with respect to the unknown true data distribution $p_\star$. For a certain class of divergences, expanding the divergence w.r.t the true probability distribution yields: $\mathbb{D}[p_\star||p_\theta] = c_\star + \mathbb{E}_{p_\star(\mathbf{x})} s[p_\theta(\mathbf{x})]$, where $c_\star$ is a constant that depends only on the data and $s[.] : \mathbb{R}^+ \to \mathbb{R}$ is a score function associated to $\mathbb{D}[.||.]$. Clearly, the smaller the score $s[p_\theta(\mathbf{x})]$, the better the data point $\mathbf{x} \sim p_\star$ fits the model p_θ. In practice, given a set of observations $\{\mathbf{x}_i\}_{i=1,\cdots,N} \sim_{i.i.d} p_\star$, one would minimize the sample average $N^{-1}\sum_{i=1}^N s[p_\theta(\mathbf{x}_i)]$ over $\theta \in \Theta$. The most popular example of these scoring functions (Parry et al., 2012) is the logarithmic score given by $-\log p_\theta(\mathbf{x})$ which is obtained by minimizing the Kullback-Leibler (KL) divergence, i.e. $\mathbb{D} = \mathbb{D}_{\text{KL}}$. In this case, the procedure of minimizing the score function is widely known as maximum likelihood (ML) estimation and has been extensively applied in statistics and machine learning. Popular instances of ML estimation include logistic regression when minimizing the cross-entropy loss w.r.t a Bernoulli model of the data and regression when minimizing the squared loss in the presence of a Gaussian model of the data (Bishop, 2006). In the context of variational inference, the logarithmic score is fundamental in the construction of variational autoencoders (Kingma and Welling, 2014) as we showed in the previous section.

Recently, the Hyvärinen score (Hyvärinen, 2005; Ding et al., 2019; Liu et al., 2016; Lyu, 2009) that we denote by $s_\nabla[.]$ has been proposed as an alternative to the logarithmic score. It turns out that the Hyvärinen score can be obtained by minimizing the Fisher divergence defined as

$$\mathbb{D}_\nabla[p_\star||p_\theta] = \mathbb{E}_{p_\star(\mathbf{x})}\frac{1}{2}\left\|\nabla_{\mathbf{x}}\log p_\star(\mathbf{x}) - \nabla_{\mathbf{x}}\log p_\theta(\mathbf{x})\right\|^2, \tag{4}$$

where $\nabla_{\mathbf{x}}$ denotes the gradient w.r.t $\mathbf{x}$. Assuming the same regularity conditions as in Proposition 1 (Ding et al., 2019), we have

$$\mathbb{D}_\nabla[p_\star||p_\theta] = \mathbb{E}_{p_\star(\mathbf{x})}\frac{1}{2}\left\|\nabla_{\mathbf{x}}\log p_\star(\mathbf{x})\right\|^2 + s_\nabla[p_\theta(\mathbf{x})], \tag{5}$$

with

$$s_\nabla[p(\mathbf{x})] = \frac{1}{2}\left\|\nabla_{\mathbf{x}}\log p(\mathbf{x})\right\|^2 + \Delta_{\mathbf{x}}\log p(\mathbf{x}), \tag{6}$$

for some probability density function $p(\mathbf{x})$ and $\Delta_{\mathbf{x}} = \sum_{j=1}^D \frac{\partial^2}{\partial x_j^2} f(\mathbf{x})$ denotes the Laplacian of some function f w.r.t $\mathbf{x}$. The potential of both the Fisher divergence and the Hyvärinen score is their ability to deal with probability distributions that are known up to some multiplicative constant. This interesting property allows to consider larger class of unormalized distributions and therefore better fits the data. In the next section, we provide a detailed description of how we can extend the use of Fisher divergence and Hyvärinen score in the context of variational auto-encoders.

3 Proposed Fisher Auto-Encoder

Recall from (2) that instead of minimizing the logarithmic score $-\log p_\theta(\mathbf{x})$, we instead upper bound the score and minimize $\mathcal{L}_{\text{VAE}}(\mathbf{x}; \phi, \theta)$. Similarly, one would look for an upper bound to the Hyvärinen score $s_\nabla[p_\theta(\mathbf{x})]$ and minimize it w.r.t model parameters ϕ and θ. However, this is quite non-trivial as opposed to the logarithmic score in (2). Fortunately, the upper bound in (2) can be recovered by minimizing the KL divergence between the following two joint distributions: $q_{\star,\phi}(\mathbf{x}, \mathbf{z}) = p_\star(\mathbf{x})q_\phi(\mathbf{z}|\mathbf{x})$ and $p_{\eta,\theta}(\mathbf{x}, \mathbf{z}) = p_\eta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ where $q_\phi(\mathbf{z}|\mathbf{x})$, $p_\eta(\mathbf{z})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ are respectively the variational posterior, the prior and the decoder with parameters ϕ, η and θ.

$$\begin{aligned}
&\phi^\star_{\text{VAE}}, \eta^\star_{\text{VAE}}, \theta^\star_{\text{VAE}} \\
&= \arg\min_{\phi,\eta,\theta} \mathbb{D}_{\text{KL}}\left[q_{\star,\phi}(\mathbf{x}, \mathbf{z})||p_{\eta,\theta}(\mathbf{x}, \mathbf{z})\right] \\
&= \arg\min_{\phi,\eta,\theta} \mathbb{E}_{p_\star(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\star(\mathbf{x}) + \log\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\eta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})}\right] \\
&= \arg\min_{\phi,\eta,\theta} \mathbb{E}_{p_\star(\mathbf{x})}\{\mathbb{D}_{\text{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\eta(\mathbf{z})\right] \\
&\qquad\qquad - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z})\} \\
&= \arg\min_{\phi,\eta,\theta} \mathbb{E}_{p_\star(\mathbf{x})}\mathcal{L}_{\text{VAE}}(\mathbf{x}; \phi, \eta, \theta).
\end{aligned} \tag{7}$$

Following the same line of thought, we propose to minimize the Fisher divergence between $q_{\star,\phi}(\mathbf{x}, \mathbf{z})$ and $p_{\eta,\theta}(\mathbf{x}, \mathbf{z})$ as follows:

$$\begin{aligned}
&\phi^\star, \eta^\star, \theta^\star \\
&= \arg\min_{\phi,\eta,\theta} \mathbb{D}_\nabla\left[q_{\star,\phi}(\mathbf{x}, \mathbf{z})||p_{\eta,\theta}(\mathbf{x}, \mathbf{z})\right] \\
&= \arg\min_{\phi,\eta,\theta} \\
&\mathbb{E}_{q_{\star,\phi}(\mathbf{x},\mathbf{z})}\frac{1}{2}\left\|\nabla_{\mathbf{x},\mathbf{z}}\log q_{\star,\phi}(\mathbf{x}, \mathbf{z}) - \nabla_{\mathbf{x},\mathbf{z}}\log p_{\eta,\theta}(\mathbf{x}, \mathbf{z})\right\|^2,
\end{aligned} \tag{8}$$

where $\nabla_{\mathbf{x},\mathbf{z}}$ denotes the gradient w.r.t the augmented variable $\{\mathbf{x}, \mathbf{z}\}$. The following theorem provides a simplified

expression of the Fisher AE loss by expanding and simplifying the Fisher divergence in (8).

Theorem 1. *The minimization in* (8) *is equivalent to the following minimization problem:*

$$\begin{aligned}\phi^\star, \eta^\star, \theta^\star &= \arg\min_{\phi,\eta,\theta} \mathbb{D}_\nabla \left[q_{\star,\phi}(\mathbf{x},\mathbf{z})||p_{\eta,\theta}(\mathbf{x},\mathbf{z})\right] \\ &= \arg\min_{\phi,\eta,\theta} \mathbb{E}_{p_\star(\mathbf{x})} \mathcal{L}_{\textit{F-AE}}\left(\mathbf{x};\phi,\eta,\theta\right),\end{aligned} \tag{9}$$

where

$$\begin{aligned}&\mathcal{L}_{\textit{F-AE}}\left(\mathbf{x};\phi,\eta,\theta\right) \\ &= \underbrace{\mathbb{D}_\nabla \left[q_\phi(\mathbf{z}|\mathbf{x})||p_{\eta,\theta}(\mathbf{z}|\mathbf{x})\right]}_{①} \\ &+ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\underbrace{s_\nabla\left[p_\theta(\mathbf{x}|\mathbf{z})\right]}_{②} + \underbrace{\frac{1}{2}\left\|\nabla_\mathbf{x}\log q_\phi(\mathbf{z}|\mathbf{x})\right\|^2}_{③}\Big].\end{aligned} \tag{10}$$

Proof. A proof can be found in the supplementary material. □

The Fisher AE loss denoted by $\mathcal{L}_{\text{F-AE}}\left(\mathbf{x};\phi,\eta,\theta\right)$ in (10) is the sum of the following three terms: ① the Fisher divergence between the two posteriors $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_{\eta,\theta}(\mathbf{z}|\mathbf{x})$. In traditional VAEs, the KL divergence between these two posteriors is generally intractable since $p_{\eta,\theta}(\mathbf{z}|\mathbf{x}) = \frac{p_\eta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})}{p_{\eta,\theta}(\mathbf{x})}$ and $p_{\eta,\theta}(\mathbf{x})$ is hard to compute because $p_{\eta,\theta}(\mathbf{x}) = \int p_\eta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})d\mathbf{z}$. Interestingly, with the Fisher divergence this limitation is alleviated since $p_{\eta,\theta}(\mathbf{z}|\mathbf{x}) \propto p_\eta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ and we only need $\nabla_\mathbf{z}\log p_{\eta,\theta}(\mathbf{z}|\mathbf{x}) = \nabla_\mathbf{z}\log p_\eta(\mathbf{z}) + \nabla_\mathbf{z}\log p_\theta(\mathbf{x}|\mathbf{z})$ for computation. The second term given by ② is the Hyvärinen score of $p_\theta(\mathbf{x}|\mathbf{z})$ which is nothing but a reconstruction loss similar to $-\log p_\theta(\mathbf{x}|\mathbf{z})$ in regular VAEs. When $p_\theta(\mathbf{x}|\mathbf{z}) \propto e^{-\frac{1}{2}\|\mathbf{x}-f_\theta(\mathbf{z})\|^2}$, the reconstruction loss is given by the squared loss[1]: $\frac{1}{2}\left\|\mathbf{x}-f_\theta(\mathbf{z})\right\|^2$ which is the same as in regular VAEs under the same model, $f_\theta(.):\mathbb{R}^d\to\mathbb{R}^D$ is the decoder parametrized by θ. The last term ③ is a stability term that permits to produce robust features in the sense that the posterior distribution is robust against small perturbations in the input data. This is similar to contractive auto-encoders which promote the invariance property in feature extraction (Rifai et al., 2011).

Remark 1. *When $q_\phi(\mathbf{z}|\mathbf{x}) = p_{\eta,\theta}(\mathbf{z}|\mathbf{x})$, the Fisher AE loss becomes exactly the Hyvärinen score of the model distribution $p_{\eta,\theta}(\mathbf{x})$, i.e. $\mathcal{L}_{\textit{F-AE}}\left(\mathbf{x};\phi,\eta,\theta\right) = s_\nabla\left[p_{\eta,\theta}(\mathbf{x})\right]$. This is similar to traditional VAEs since we also have $\mathcal{L}_{\textit{VAE}}\left(\mathbf{x};\phi,\eta,\theta\right) = -\log p_{\eta,\theta}(\mathbf{x})$ in this case.*

Proof. When $q_\phi(\mathbf{z}|\mathbf{x}) = p_{\eta,\theta}(\mathbf{z}|\mathbf{x})$, $\mathbb{D}_\nabla\left[q_{\star,\phi}(\mathbf{x},\mathbf{z})||p_{\eta,\theta}(\mathbf{x},\mathbf{z})\right] = \mathbb{D}_\nabla\left[p_\star(\mathbf{x})||p_{\eta,\theta}(\mathbf{x})\right]$. The proof is concluded by relying on (5). □

[1]We omit the constant term coming from the Laplacian $\Delta_\mathbf{x}\log p_\theta(\mathbf{x}|\mathbf{z})$ since it is irrelevant to the minimization problem in (9).

Given a data point $\mathbf{x}$, the Fisher AE loss can be estimated using Monte Carlo with L samples from $q_\phi(\mathbf{z}|\mathbf{x})$ as follows:

$$\begin{aligned}&\mathcal{L}_{\text{F-AE}}\left(\mathbf{x};\phi,\eta,\theta\right) \\ &\simeq \mathcal{L}^{(L)}_{\text{F-AE}}\left(\mathbf{x};\phi,\eta,\theta\right) \\ &= \frac{1}{2L}\sum_{l=1}^{L}\Bigg[\left\|\nabla_\mathbf{z}\log q_\phi(\mathbf{z}^{(l)}|\mathbf{x}) - \nabla_\mathbf{z}\log p_\eta(\mathbf{z}^{(l)})\right. \\ &\qquad\qquad \left. - \nabla_\mathbf{z}\log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})\right\|^2 \\ &\qquad + \left\|\mathbf{x} - f_\theta(\mathbf{z}^{(l)})\right\|^2 + \left\|\nabla_\mathbf{x}\log q_\phi(\mathbf{z}^{(l)}|\mathbf{x})\right\|^2\Bigg],\end{aligned} \tag{11}$$

where $\mathbf{z}^{(l)} = \mu(\mathbf{x}) + \sigma(\mathbf{x})\odot\epsilon^{(l)}$, $\epsilon^{(l)}\sim\mathcal{N}(0,\mathbf{I})$. Moreover, $\nabla_\mathbf{z}\log q_\phi(\mathbf{z}^{(l)}|\mathbf{x}) = -\frac{\epsilon^{(l)}}{\sigma(\mathbf{x})}$ and both $\nabla_\mathbf{z}\log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})$ and $\nabla_\mathbf{x}\log q_\phi(\mathbf{z}^{(l)}|\mathbf{x})$ can be computed using automatic differentiation tools like Autograd in PyTorch. To solve the minimization in (9), we use stochastic gradient descent (SGD) with minibatch data of size N as in (Kingma and Welling, 2014). Details of the optimization are given by Algorithm 1.

Algorithm 1 Training the Fisher AE with SGD

1: **Initialize** ϕ, η and θ
2: **Repeat**:
3: Randomly sample a minibatch of training data $\{\mathbf{x}_i\}_{i=1}^N$
4: Compute gradient $\nabla_{\phi,\eta,\theta}\frac{1}{N}\sum_{i=1}^N\mathcal{L}^{(L)}_{\text{F-AE}}\left(\mathbf{x}_i;\phi,\eta,\theta\right)$
5: Update ϕ, η and θ with Adam Kingma and Ba (2014)
6: **Until** convergence
7: **Output:** ϕ_*, η_* and θ_*

3.1 Fisher AE with exponential family priors

As discussed earlier, employing the Fisher divergence has the advantage of dealing with probability distributions that are known up to some multiplicative constant. This powerful property allows to consider a rich family of distributions to model the prior $p(\mathbf{z})$. In this paper, we consider the use of exponential family whose general form is given by:

$$p_\eta(\mathbf{z}) \propto \exp\left(\eta^\top T(\mathbf{z}) + h(\mathbf{z})\right), \tag{12}$$

where η denotes the natural parameters, $h(\mathbf{z})$ is the carrier measure and $T(\mathbf{z})$ is referred to as a sufficient statistic (Wainwright and Jordan, 2008). Popular examples of the exponential family include the Bernoulli, Poisson and Gaussian distributions to name a few (Wainwright and Jordan, 2008). Note that the form given by the right hand side of (12) is not a valid PDF since it does not sum to 1, but it is sufficient to compute the gradient of the log-density w.r.t $\mathbf{z}$ which is given by $\nabla_\mathbf{z}\log p_\eta(\mathbf{z}) = \nabla_\mathbf{z}\left(\eta^\top T(\mathbf{z}) + h(\mathbf{z})\right)$.

Therefore, the term ① in (10) can be written as:

$$\mathbb{D}_{\nabla}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right] = \frac{1}{2}\int q_\phi(\mathbf{z}|\mathbf{x})\|\nabla_{\mathbf{z}}\log q_\phi(\mathbf{z}|\mathbf{x}) - \nabla_{\mathbf{z}}\left(\eta^\top T(\mathbf{z}) + h(\mathbf{z})\right) - \nabla_{\mathbf{z}}\log p_\theta(\mathbf{x}|\mathbf{z})\|^2 d\mathbf{z}.$$

which can be approximated using samples $\mathbf{z}^{(l)} \sim q_\phi(\mathbf{z}|\mathbf{x})$, $l = 1, \cdots, L$ as follows:

$$\mathbb{D}_{\nabla}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right] \simeq \frac{1}{2L}\sum_{l=1}^{L}\|\nabla_{\mathbf{z}}\log q_\phi(\mathbf{z}^{(l)}|\mathbf{x}) - \nabla_{\mathbf{z}}\left(\eta^\top T(\mathbf{z}^{(l)}) + h(\mathbf{z}^{(l)})\right) - \nabla_{\mathbf{z}}\log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})\|^2.$$

A popular class of distributions that belongs to the exponential family is given by the factorable polynomial exponential family (FPE) (Cobb et al., 1983) in which $p_\eta(\mathbf{z})$ is given by

$$p_\eta(\mathbf{z}) = p_\eta(z_1, \cdots, z_d) \propto \exp\left(\sum_{j=1}^{d}\sum_{k=1}^{K}\eta_{jk} z_j^k\right), \quad (13)$$

where K denotes the order of FPE family and $\{\eta_{jk}\}_{1\leq j\leq d, 1\leq k\leq K}$ is a set of parameters. The natural parameters, the sufficient statistic and the carrier measure in this case are given by:

$$\eta = [\eta_{11}, \eta_{12}, \cdots, \eta_{1K}, \cdots, \eta_{d1}, \eta_{d2}, \cdots, \eta_{dK}]^\top$$
$$T(\mathbf{z}) = \left[z_1, z_1^2, \cdots, z_1^K, \cdots, z_d, z_d^2, \cdots, z_d^K\right]^\top$$
$$h(\mathbf{z}) = 0.$$

With the model in (13), the gradient of $\log p_\eta(\mathbf{z})$ w.r.t $\mathbf{z}$ can be easily derived as

$$\frac{\partial}{\partial z_j}\log p_\eta(\mathbf{z}) = \sum_{k=1}^{K} k\eta_{jk} z_j^{k-1}, \quad j = 1, \cdots, d.$$

4 Experiments

In this section, we provide both qualitative and quantitative results that demonstrate the ability of our proposed Fisher AE model to produce high quality samples on real-world image datasets such as MNIST and CelebA. We compare results with both regular VAEs (Kingma and Welling, 2014) and Wasserstein Auto-Encoders with GAN penalty (WAE-GAN) (Ilya et al., 2018). In the supplementary material, we provide full details for the encoder/decoder architectures used by the different schemes for both MNIST and celebA datasets.

Setup

For optimization, we use Adam (Kingma and Ba, 2014) with a learning rate $\texttt{lr} = 2.10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, a mini-batch size of 128 and trained various models for 100 epochs. For all experiments, we pick $d = 8$ for MNIST and $d = 64$ for celebA and use exponential family priors for the Fisher AE as in (13). We notice that $K = 5$ (order of FPE family) seems to work better in all experiments whereas Gaussian priors are used for VAE and WAE-GAN. We use Gaussian posteriors for both Fisher AE and VAE such that $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})^2\right)$ where $\mu_\phi(.)$ and $\sigma_\phi(.)$ are determined by the encoder architecture for which details are postponed to the supplementary material.

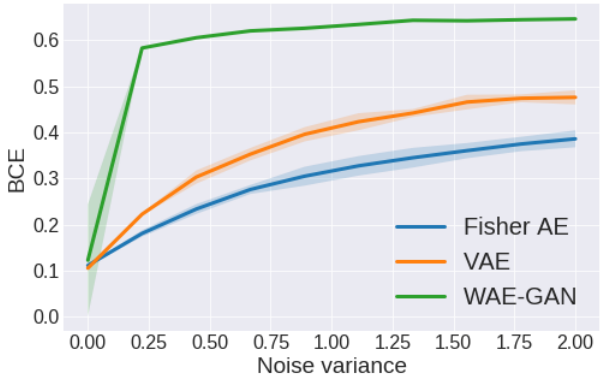

Figure 1: BCE vs. noise variance σ^2 for MNIST.

Sampling with SVGD

To sample from the exponential family prior after training, we use Stein Variational Gradient Descent (SVGD) (Liu and Wang, 2016) . Let M be the number of samples that we would like to sample from $p_{\eta_*}(\mathbf{z})$ denoted by $\{\mathbf{z}_i^*\}_{i=1}^M$. We start with $\{\mathbf{z}_i\}_{i=1}^M \sim_{i.i.d} \mathcal{N}(0, \mathbf{I})$ and we keep evolving these samples with a step-size 10^{-3} for $15,000$ iterations. These parameters (step-size and number of iterations) seem to work reasonably well across all experiments.

MNIST

Figure 2 exhibits a comparison between the three auto-encoders in terms of robustness, test reconstruction, and random sampling. In order to compare the robustness, we plot the reconstructed samples of the different schemes when the test data is corrupted by an isotropic Gaussian noise with a covariance matrix $0.2 \times \mathbf{I}_D$. The results of this experiment are given by the first row of Figure 2. Clearly, WAE-GAN completely fail to reconstruct the test data and Fisher AE seems to be more robust to noise. This result is confirmed quantitatively in Figure 1 where we plot the normalized binary cross-entropy (BCE) w.r.t the noise variance added to the test data, i.e. we feed the different trained models with $\texttt{data} = \texttt{test data} + \mathcal{N}(0, \sigma^2\mathbf{I}_D)$ and compute the BCE reconstruction loss w.r.t the true test data. In the second and

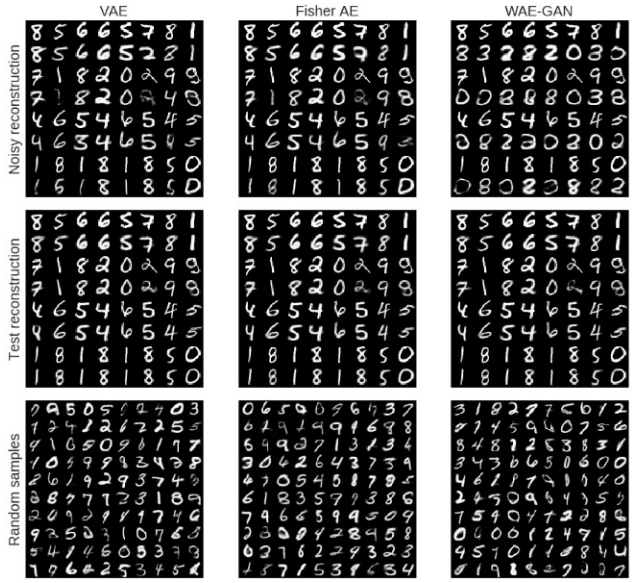

Figure 2: Performance of the Fisher AE trained on MNIST dataset in comparison with VAE and WAE-GAN. True test data are given by the odd rows in both reconstruction tasks (rows 1 and 2).

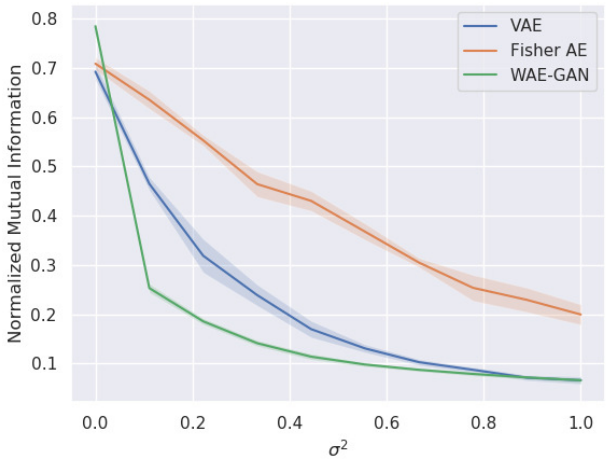

Figure 3: Robustness of latent space clustering in terms of the normalized mutual information on MNIST test set.

Algorithm	FID score
VAE	89.1 ± 1.1
Fisher AE (Gaussian prior)	89.1 ± 0.9
Fisher AE (Exp. prior)	$\mathbf{84.7 \pm 0.8}$
WAE-GAN	75.2 ± 1.0

Table 1: FID scores of the different generative models trained on CelebA (smaller is better).

third rows of Figure 2, we show both the reconstruction and generative performance of the different auto-encoders. For both test reconstruction and random sampling, the proposed Fisher AE exhibits a comparable performance to WAE-GAN which achieves the best generative performance thanks to the GAN penalty in the loss function (Ilya et al., 2018).

We further examine the robustness of the different models w.r.t latent representation when the data is corrupted by additive Gaussian noise. In Figure 4 using non-linear dimensionality reduction techniques such as t-SNE, we visualize the 2D latent structure of the different models. As shown in Figure 4, even with corrupted data, the latent structure of the Fisher AE is still preserved and the clusters associated to different classes are relatively distinguishable. This is not the case for VAE and WAE-GAN where the clusters in the latent space are somewhat mixed up when the data is perturbed by noise. This behavior is quantitatively confirmed in Figure 3 where test data is perturbed with Gaussian noise with variance σ^2, then encoded with each model encoder and projected with t-SNE and finally clustered using k-means. The quality of clustering is measured using the normalized mutual information: $\text{NMI} = \frac{2\mathcal{I}(\Omega;\mathbb{C})}{\mathcal{H}(\Omega)+\mathcal{H}(\mathbb{C})}$, where Ω is the model clusters for a given noise level, $\mathbb{C}$ is the true class labels, $\mathcal{I}(.;.)$ denotes the mutual information and $\mathcal{H}(.)$ denotes the entropy. Clearly, Fisher AE exhibits a better behavior in terms of clustering robustness where the decay in performance is nearly linear whereas for both VAE and WAE-GAN, the performance decays faster.

CelebA

For the CelebA dataset, it is clear from the first row (the noisy reconstructions) of Figure 6 that the proposed Fisher AE is more robust than both VAE and WAE-GAN when the test data is corrupted with an isotropic Gaussian noise with covariance matrix $2\mathbf{I}_D$. We further validate this property with different noise levels as depicted in Figure 5 where the Fisher AE outperforms VAE and WAE-GAN in the reconstruction MSE. Moreover, as shown in Figure 6, the Fisher AE generates better samples than VAE and has comparable quality to WAE. The visual quality of the samples is confirmed by the quantitative results summarized in Table 1 where the proposed Fisher AE with exponential family priors outperforms VAE in terms of the Fréchet Inception Distance (FID) and has relatively worse performance than WAE. Furthermore, sampling using the exponential prior provides additional challenges due to the difficulty of convergence of the algorithm. This may be alleviated with alternative sampling algorithms, but that remains beyond the scope of this paper.

5 Conclusion

In this paper, we introduced a new type of auto-encoders constructed based on the minimization of the Fisher diver-

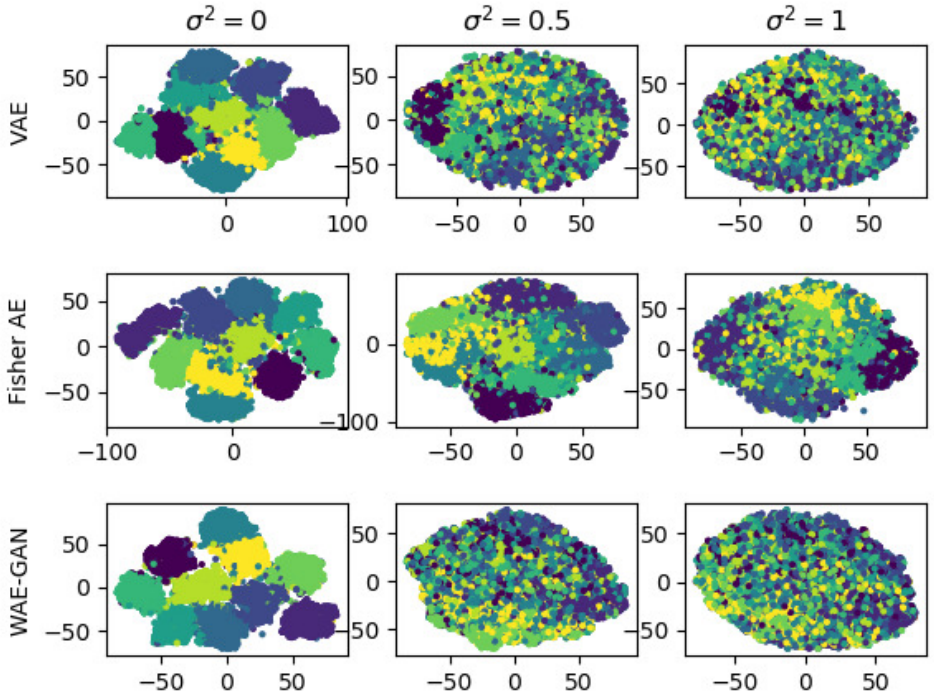

Figure 4: Visualisation of models latent representation on MNIST test set using t-SNE for different noise levels.

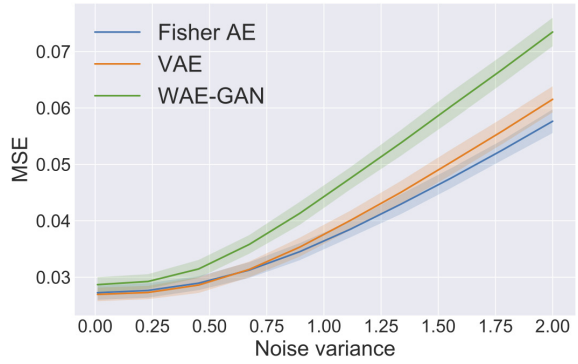

Figure 5: MSE vs. noise variance σ^2 for celebA. Errors are computed from variances in batches in the test set.

gence between the joint distribution over the data and latent variables and the model joint distribution. The resulting loss function has two interesting aspects: 1) it allows to directly minimize the tractable Fisher divergence between the approximate and the true posteriors and 2) considers a stability measure of the encoder that allows to produce robust features. Experimental results were provided to demonstrate the competitive performance of the proposed Fisher autoencoders as compared to some existing schemes like VAEs and Wasserstein AEs and their superiority in terms of robustness. An interesting but non trivial extension of the present work is to consider the modeling of the posterior distribution using exponential family priors.

Acknowledgements

This work was supported by the Army Research Office grant No. W911NF-15-1-0479.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Cobb, L., Koppstein, P., and Chen, N. H. (1983). Estimation and moment recursion relations for multimodal distributions of the exponential family. In *Journal of the American Statistical Association*, page 124–130.

Ding, J., Calderbank, R., and Tarokh, V. (2019). Gradient information for representation and modeling. *Advances in Neural Information Processing Systems 32*, pages 2396–2405.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

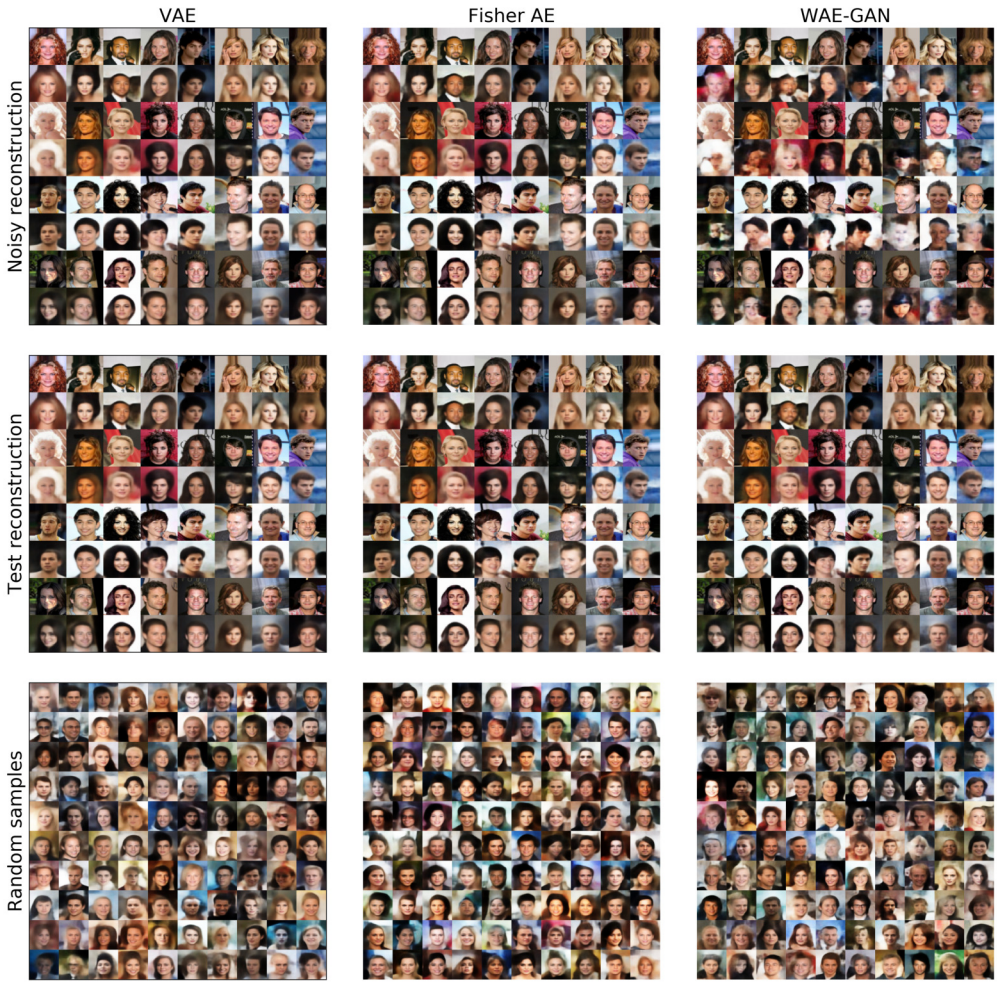

Figure 6: Performance of the Fisher AE trained on celebA dataset in comparison with VAE and WAE-GAN. True test data are given by the odd rows in the reconstruction tasks (rows 1 and 2).

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. In *J. Mach. Learn. Res.*, volume 6, page 695–709.

Ilya, T., Olivier, B., Sylvain, G., and Scholkopf, B. (2018). Wasserstein auto-encoders. In *ICLR*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *http://arxiv.org/abs/1412.6980*.

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

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database. *http://yann.lecun.com/exdb/mnist/*.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 276–284.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in Neural Information Processing Systems 29*, pages 2378–2386.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lyu, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 359–366, Arlington, Virginia, USA. AUAI Press.

Parry, M., Dawid, A. P., and Lauritzen, S. (2012). Proper local scoring rules. In *Ann. Stat*, page 561–592.

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. (2017). Vae learning via stein variational gradient descent. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4236–4245. Curran Associates, Inc.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1530–1538.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, pages 833–840.

Shao, S., Jacob, P. E., Ding, J., and Tarokh, V. (2019). Bayesian model comparison with the hyvärinen score: computation and consistency. *J. Am. Stat. Assoc.*, pages 1–24.

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

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305.