Training Deep Energy-Based Models with f-Divergence Minimization

Lantao Yu [1] Yang Song [1] Jiaming Song [1] Stefano Ermon [1]

Abstract

Deep energy-based models (EBMs) are very flexible in distribution parametrization but computationally challenging because of the intractable partition function. They are typically trained via maximum likelihood, using contrastive divergence to approximate the gradient of the KL divergence between data and model distribution. While KL divergence has many desirable properties, other f-divergences have shown advantages in training implicit density generative models such as generative adversarial networks. In this paper, we propose a general variational framework termed f-EBM to train EBMs using any desired f-divergence. We introduce a corresponding optimization algorithm and prove its local convergence property with non-linear dynamical systems theory. Experimental results demonstrate the superiority of f-EBM over contrastive divergence, as well as the benefits of training EBMs using f-divergences other than KL.

1. Introduction

Learning deep generative models that can approximate complex distributions over high-dimensional data is an important problem in machine learning, with many applications such as image, speech, natural language generation (Radford et al., 2015; Oord et al., 2016a; Yu et al., 2017) and imitation learning (Ho & Ermon, 2016). To this end, two major branches of generative models have been widely studied: tractable density and implicit density generative models. To enable the use of maximum likelihood training, tractable density models have to use specialized architectures to build a normalized probability model. These include autoregressive models (Larochelle & Murray, 2011; Germain et al., 2015; Oord et al., 2016b; Van den Oord et al., 2016), flow-based models (Dinh et al., 2014; 2016; Kingma & Dhariwal, 2018) and sum-product networks (Poon & Domingos, 2011). However, such a normalization requirement can hinder the flexibility and expressiveness. For example, flow-based models rely on invertible transformations with tractable Jacobian determinant, while sum-product networks rely on special graph structures (with sums and products as internal nodes) to obtain tractable density. To overcome the limitations caused by the constraint of specifying a normalized explicit density, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) proposed to learn implicit density generative models within a minimax framework, where the generative model is a direct differentiable mapping from noise space to sample space. So far many variants of GANs have been proposed in order to minimize various discrepancy measures (Nowozin et al., 2016; Arjovsky et al., 2017). In particular, f-GANs (Nowozin et al., 2016) minimize a variational representation of f-divergence between data and model distribution, which is a general family of divergences for measuring the discrepancies between two distributions. Although appealing, a main drawback of GANs is the lack of explicit density, which is the central goal of density estimation and plays an indispensable role in applications such as anomaly detection, image denoising and inpainting.

Deep energy-based models (EBMs) are promising to combine the best of both worlds, in the sense that EBMs allow both extremely flexible distribution parametrization and the access to an (unnormalized) explicit density. However, because of the intractable partition function (an integral over the sample space), so far within the family of f-divergences, only KL divergence proved to be tractable to optimize with methods such as contrastive divergence (Hinton, 2002), doubly dual embedding (Dai et al., 2018) and adversarial dynamics embedding (Dai et al., 2019). Since the progress of generative modeling research was mainly driven by the study on the properties and tractable optimization of various discrepancy measures, there is an urgent need to investigate the possibility of training EBMs with other f-divergences. Specifically, the choice of discrepancy measures embodies our preferences and has a significant influence on the learned model distribution (see Figure 1 for illustration of the effect of some popular f-divergences on training EBMs in a model misspecification scenario). For example, it has been found that minimizing KL (Maximum Likelihood Estimation) is

[1] Department of Computer Science, Stanford University, Stanford, CA 94305 USA. Correspondence to: Lantao Yu <lantaoyu@cs.stanford.edu>, Stefano Ermon <ermon@cs.stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

not directly correlated with sample quality (Theis et al., 2015), and in many application scenarios we need more flexibility in trading off mode collapse vs. mode coverage.

In this paper, we propose a new variational framework termed f-EBM to enable the use of any f-divergence for training EBMs. Our framework also naturally produces a density ratio estimation, which can be used for importance sampling and bias correction (Grover et al., 2019). For example, we show that in conjunction with rejection sampling, the learned density ratio can be used to recover the data distribution even in a model misspecification scenario (see Figure 2 for illustration), which traditional methods for minimizing KL divergence such as contrastive divergence cannot do. Furthermore, with the theory of non-linear dynamical systems (Hassan, 1996), we establish a rigorous proof on the local convergence property of the proposed single-step gradient optimization algorithm for f-EBM. Finally, experimental results demonstrate the benefits of using various f-divergences to train EBMs, and more importantly, with some members in the f-divergence family as the training objectives (*e.g.*, Jensen-Shannon, Squared Hellinger and Reverse KL), we are able to achieve significant sample quality improvement compared to contrastive divergence method on some commonly used image datasets.

2. Preliminaries

2.1. The f-Divergence Family

Let P and Q denote two probability distributions with density functions p and q with respect to a base measure $\mathrm{d}\boldsymbol{x}$ on domain $\mathcal{X}$. Suppose P is absolutely continuous with respect to Q, denoted as $P \ll Q$ (*i.e.*, the Radon-Nikodym derivative $\mathrm{d}P/\mathrm{d}Q$ exists). For any convex, lower-semicontinuous function $f : [0, +\infty) \to \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence between P and Q is defined as:

$$D_f(P\|Q) := \int_{\mathcal{X}} q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} = \mathbb{E}_{q(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)\right].$$

Many divergences are special cases obtained by choosing a suitable *generator function* f. For example, $f(u) = u \log u$ and $f(u) = -\log u$ correspond to forward KL and reverse KL respectively (see Table 5 in (Nowozin et al., 2016) for more examples). In the rest of this paper, we will consider the generator function f to be strictly convex and continuously differentiable, and we will use f' to denote the derivative of f.

Definition 1 (Fenchel Duality). *For any convex, lower-semicontinuous function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ defined over a Banach space $\mathcal{X}$, the Fenchel conjugate function of f, f^* is defined over the dual space $\mathcal{X}^*$ as:*

$$f^*(\boldsymbol{x}^*) := \sup_{\boldsymbol{x} \in \mathcal{X}} \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle - f(\boldsymbol{x}), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the duality paring between $\mathcal{X}$ and $\mathcal{X}^*$. For a finite dimensional space $\mathbb{R}^m$, the dual space is also $\mathbb{R}^m$ and the duality paring is the usual vector inner product.

The function f^* is also convex and lower-semicontinuous and more importantly, we have $f^{**} = f$, which means we can represent f through its conjugate function as:

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{x}^* \in \mathcal{X}^*} \langle \boldsymbol{x}^*, \boldsymbol{x} \rangle - f^*(\boldsymbol{x}^*). \tag{2}$$

Based on Fenchel duality, Nguyen et al. (2010) proposed a general variational representation of f-divergences:

Definition 2. *Define the subdifferential $\partial f(\boldsymbol{x})$ of a convex function $f : \mathbb{R}^m \to \mathbb{R}$ at a point $\boldsymbol{x} \in \mathbb{R}^m$ as the set: $\partial f(\boldsymbol{x}) := \{\boldsymbol{z} \in \mathbb{R}^m | f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x} \rangle, \forall \boldsymbol{y} \in \mathbb{R}^m\}$. When f is differentiable at $\boldsymbol{x}$, $\partial f(\boldsymbol{x}) = \{\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\}$.*

Lemma 1 (Nguyen et al. (2010)). *Let P and Q be two probability measures over the Borel σ-algebra on domain $\mathcal{X}$ with densities p, q and $P \ll Q$. For any class of functions $\mathcal{T} = \{T : \mathcal{X} \to \mathbb{R}\}$ such that the subdifferential $\partial f(p/q)$ contains an element of $\mathcal{T}$, we have*

$$D_f(P\|Q) = \sup_{T \in \mathcal{T}} \mathbb{E}_{p(\boldsymbol{x})}[T(\boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{x})}[f^*(T(\boldsymbol{x}))], \tag{3}$$

where the supreme is attained at $T^*(\boldsymbol{x}) = f'\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)$.

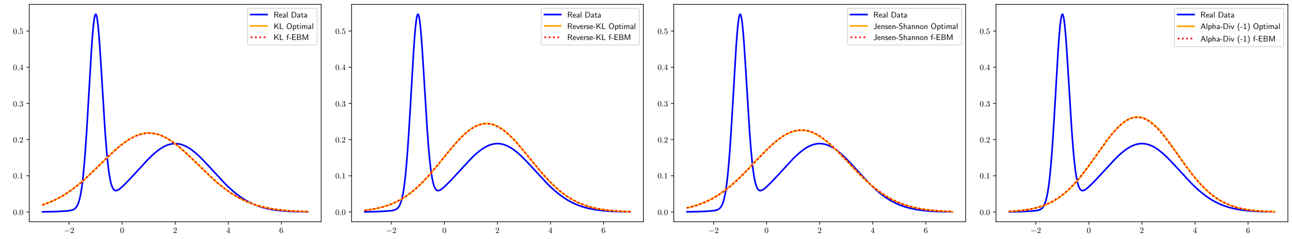
2.2. Energy-Based Generative Modeling

Suppose we are given a dataset consisting of *i.i.d.* samples $\{\boldsymbol{x}_i\}_{i=1}^N$ from an unknown data distribution $p(\boldsymbol{x})$, defined over a sample space $\mathcal{X} \subset \mathbb{R}^m$. Our goal is to find a parametric approximation $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ to the data distribution, such that we can perform downstream inferences (*e.g.*, density evaluation and sampling). Specifically, we are interested in learning an (unnormalized) energy function $E_{\boldsymbol{\theta}}(\boldsymbol{x})$ that defines the following normalized distribution:

$$q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))/Z_{\boldsymbol{\theta}}, \tag{4}$$

where $Z_{\boldsymbol{\theta}} = \int_{\mathcal{X}} \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$ is the normalization constant (also called the partition function). In this paper, unless otherwise stated, we will always use p to denote the data distribution and $q_{\boldsymbol{\theta}}$ to denote the model distribution. Moreover, we will assume $\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$ is integrable over sample space $\mathcal{X}$, *i.e.*, $Z_{\boldsymbol{\theta}}$ is finite. Because of the absence of the normalization requirement, energy-based distribution representation offers extreme modeling flexibility in the sense that we can use almost any model architecture that outputs a real number given an input as the energy function.

While flexible, inference in EBMs is challenging because of the partition function, which is generally intractable to compute exactly. For example, it is non-trivial to sample from an EBM, which usually requires Markov Chain Monte



(a) KL (b) Reverse KL (c) Jensen-Shannon (d) α-Divergence ($\alpha = -1$)

Figure 1. The influences of different f-divergences on the training of EBMs. The blue solid line represents the real data distribution; the orange solid line represents the optimal model distribution under a certain discrepancy measure obtained by directly solving the minimization problem with the real data density; the red dashed line represents the model distribution learned by f-EBM.

Carlo (MCMC) (Robert & Casella, 2013) techniques. As a gradient based MCMC method, Langevin dynamics (Neal et al., 2011; Welling & Teh, 2011) defines an efficient iterative sampling process, which asymptotically can produce samples from an energy-based distribution:

$$\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{t-1}) + \sqrt{\epsilon}\,\boldsymbol{z}_t, \tag{5}$$

where $\boldsymbol{z}_t \sim \mathcal{N}(0, I)$. The distribution of $\tilde{\boldsymbol{x}}_T$ converges to the model distribution $q_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$ when $\epsilon \to 0$ and $T \to \infty$ under some regularity conditions (Welling & Teh, 2011). Although few bounds on the mixing time are known, in practice researchers found that using a relatively small ϵ and finite T suffices to produce good samples and so far various scalable techniques have been developed for the purpose of sampling from an EBM efficiently. For example, Du & Mordatch (2019) proposed to use Langevin dynamics together with a sample replay buffer to reduce mixing time and improve sample diversity, which is a sampling strategy we employ in the experiments.

2.3. Tackling Intractable Partition Function with Contrastive Divergence

The predominant approach to training explicit density generative models is to minimize the KL divergence between the (empirical) data distribution and model distribution, which is a specific member within the f-divergence family. Minimizing KL is equivalent to the following maximum likelihood estimation (MLE) objective:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{MLE}}(\boldsymbol{\theta}; p) = \min_{\boldsymbol{\theta}} -\mathbb{E}_{p(\boldsymbol{x})}\left[\log q_{\boldsymbol{\theta}}(\boldsymbol{x})\right] \tag{6}$$

Lemma 2. *Given a $\boldsymbol{\theta}$-parametrized energy-based distribution $q_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$. Suppose both $\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$ and its partial derivative $\nabla_{\boldsymbol{\theta}} \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$ are continuous w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{x}$. With Leibniz integral rule, $\forall \boldsymbol{x} \in \mathcal{X}$, we have:*

$$\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) + \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]. \tag{7}$$

Proof. See (Turner, 2005) for derivations. □

Because of the intractable partition function, we do not have access to the normalized density $q_{\boldsymbol{\theta}}$ and cannot directly minimize the MLE objective. Fortunately, we can use Lemma 2 to estimate the gradient of $\mathcal{L}_{\mathrm{MLE}}(\boldsymbol{\theta}; p)$, , which gives rise to the contrastive divergence method (Hinton, 2002):

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{MLE}}(\boldsymbol{\theta}; p) &= -\mathbb{E}_{p(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{x}) \\ &= \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]. \end{aligned} \tag{8}$$

Intuitively this amounts to decreasing the energies of "positive" samples from p and increasing the energies of "negative samples" from $q_{\boldsymbol{\theta}}$. As mentioned in Section 2.2, MCMC techniques such as Langevin dynamics are needed in order to sample from $q_{\boldsymbol{\theta}}$, which gives rise to the following surrogate gradient estimation:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CD}-K}(\boldsymbol{\theta}; p) = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}^K(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})],$$

where $q_{\boldsymbol{\theta}}^K$ denotes the distribution after K steps of MCMC transitions from an initial distribution (typically data distribution or uniform distribution), and Equation (8) corresponds to $\mathcal{L}_{\mathrm{CD}-\infty}$.

3. Method

In this section, we consider a more general training objective, which is minimizing the f-divergence between the (empirical) data distribution and a $\boldsymbol{\theta}$-parametrized energy-based distribution:

$$\min_{\boldsymbol{\theta}} D_f(p \| q_{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[f\left(\frac{p(\boldsymbol{x})}{q_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)\right] \tag{9}$$

3.1. Challenges of Training EBMs with f-Divergences

3.1.1. Challenges of The Primal Form

The reason for the predominance of KL divergence for training EBMs is the convenience and tractability of gradient estimation. Specifically, as discussed in Section 2.3, the gradient of KL divergence reduces to the expected gradient of the log-likelihood, which can be estimated using Lemma 2. However, this is not generally applicable to other divergences, and as far as we know, within the f-divergence family, KL divergence is the only one that permits such a convenient form. For example, reverse KL is another member in the f-divergence family with a simple form:

$$D_{\mathrm{RevKL}}(p\|q_{\boldsymbol{\theta}}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\log\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{p(\boldsymbol{x})}\right]. \tag{10}$$

The gradient can be calculated as:

$$\begin{aligned}&\nabla_{\boldsymbol{\theta}} D_{\mathrm{RevKL}}(p\|q_{\boldsymbol{\theta}})\\ &= \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}\left[\nabla_{\boldsymbol{\theta}}\log q_{\boldsymbol{\theta}}(\boldsymbol{x})\left(\log\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{p(\boldsymbol{x})}+1\right)\right].\end{aligned} \tag{11}$$

Although we can still use Equation (7) in Lemma 2 to evaluate the gradient of the log-likelihood term ($\nabla_{\boldsymbol{\theta}}\log q_{\boldsymbol{\theta}}(\boldsymbol{x})$) in Equation (11), it is still infeasible to estimate because we do not know the density ratio $q_{\boldsymbol{\theta}}(\boldsymbol{x})/p(\boldsymbol{x})$ (recall that we only have samples from p and $q_{\boldsymbol{\theta}}$).

3.1.2. Challenges of The Dual Form

The variational characterization from Lemma 1 provides us a way to estimate an f-divergence only using samples from p and $q_{\boldsymbol{\theta}}$. Therefore we can instead minimize the variational representation of f-divergence by solving the following minimax problem:

$$\min_{\boldsymbol{\theta}}\max_{\boldsymbol{\omega}} \mathbb{E}_{p(\boldsymbol{x})}[T_{\boldsymbol{\omega}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))], \tag{12}$$

where $T_{\boldsymbol{\omega}} : \mathcal{X} \to \mathbb{R}$ is the variational function. When $q_{\boldsymbol{\theta}}$ is an implicit density generative model defined by a fixed noise distribution $\pi(\boldsymbol{z})$ and a direct differentiable mapping $G_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathcal{X}$ from noise space to sample space, Equation (12) corresponds to the f-GAN framework (Nowozin et al., 2016), where the gradient *w.r.t.* $\boldsymbol{\theta}$ can be conveniently estimated using reparametrization:

$$-\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] = -\mathbb{E}_{\pi(\boldsymbol{z})}[\nabla_{\boldsymbol{\theta}} f^*(T_{\boldsymbol{\omega}}(G_{\boldsymbol{\theta}}(\boldsymbol{z})))] \tag{13}$$

However, it is challenging to apply this procedure to EBMs. First of all, it is generally hard to find such a deterministic mapping $G_{\boldsymbol{\theta}}$ to produce exact samples from a given EBM. Consequently, we have to use some approximate sampling process (such as Langevin dynamics in Equation (5)) as a surrogate. Differentiating through this sampling process (a sequence of Markov chain transitions), as required for computing Equation (13), is computationally expensive. For example, we often need hundreds of steps in Langevin dynamics to produce a single sample[1]. Therefore, the same number of backward steps is required for gradient backpropagation, where each backward step further involves computing Hessian matrices whose sizes are proportional to the parameter and data dimension. This will lead to hundreds of times more memory consumption compared to using Langevin dynamics only for producing samples. A more detailed discussion is provided in Appendix E.2.

When reparameterization is not possible, an alternative approach is to use the "log derivative trick", similar to the REINFORCE algorithm from reinforcement learning (Williams, 1992; Sutton et al., 2000). Specifically, when the generative model is an EBM, the following theorem provides an unbiased estimation of the gradient:

Theorem 1. *For a $\boldsymbol{\theta}$-parametrized energy-based distribution $q_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$, under Assumption 1 in Appendix A, the gradient of the variational representation of f-divergence (Equation (12)) w.r.t. $\boldsymbol{\theta}$ can be written as:*

$$\begin{aligned}-\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] = \mathbb{E}_{q_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\cdot f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] -\\ \mathbb{E}_{q_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]\cdot\mathbb{E}_{q_{\boldsymbol{\theta}}}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))]\end{aligned}$$

When we can obtain i.i.d. samples from $q_{\boldsymbol{\theta}}$, we can get an unbiased estimation of the gradient.

Proof. See Appendix A. □

Theorem 1 provides us a possible way to train EBMs with any f-divergence. Based on Equation (12), we can alternatively train $T_{\boldsymbol{\omega}}$ using samples from p and $q_{\boldsymbol{\theta}}$, and train $q_{\boldsymbol{\theta}}$ using the gradient estimator in Theorem 1.

However, in practice we found that this approach performs poorly, especially for high-dimensional data. Intuitively, this method resembles REINFORCE in the sense that $f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))$ specifies an adaptive reward function to guide the training of a stochastic policy $q_{\boldsymbol{\theta}}$ through trial-and-error. Unlike contrastive divergence, here the energy function is never directly evaluated on the training data in the optimization process, since the expectation over $p(\boldsymbol{x})$ in Equation (12) only involves the variational function $T_{\boldsymbol{\omega}}$. Thus we cannot directly decrease the energies of the training data and increase the energies of the generated data (*i.e.*, to make training data more likely). Therefore the entire supervision signal comes from $f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))$. Taking image domain as an example, starting from a random initialization of the parameters $\boldsymbol{\theta}$, all the generated samples obtained by MCMC sampling are random noise. Since it is unlikely for an EBM to generate a realistic image by random exploration, the reward signals from $f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))$ are always uninformative. More precisely, although the gradient estimator is theoretically unbiased, the variance is extremely large and it would require a prohibitive number of samples to work. Consequently, in practice EBMs trained by this approach failed to generate reasonable images, as shown

[1] Note that when using reparametrization for gradient estimation, we cannot use a sample replay buffer to reduce the number of MCMC transition steps, because the initial distribution of the Markov chain $\pi(\boldsymbol{z})$ cannot depend on the model parameters $\boldsymbol{\theta}$.

in Appendix G.1. Even starting from a pretrained model learned by contrastive divergence, we empirically found that this approach still cannot provide effective training.

3.2. f-EBM

Inspired by the analysis of the challenges in Section 3.1, we propose to minimize a new variational representation of f-divergence between the data distribution and an energy-based distribution, which can mitigate the issues discussed above and induce an effective training scheme for EBMs.

Theorem 2. *Let P and Q be probability measures over the Borel σ-algebra on domain $\mathcal{X}$ with densities p, q and $P \ll Q$. Additionally, let q be an energy-based distribution, $q(\boldsymbol{x}) = \exp(-E(\boldsymbol{x}))/Z_q$. For any class of functions $\mathcal{H} = \{H : \mathcal{X} \to \mathbb{R}\}$ such that $\log(p \cdot Z_q) \in \mathcal{H}$ and the expectations in the following equation are finite, we have:*

$$D_f(P\|Q) = \sup_{H \in \mathcal{H}} \mathbb{E}_{p(\boldsymbol{x})}[f'(\exp(H(\boldsymbol{x}) + E(\boldsymbol{x})))] - \mathbb{E}_{q(\boldsymbol{x})}[f^*(f'(\exp(H(\boldsymbol{x}) + E(\boldsymbol{x}))))]$$

where the supreme is attained at $H^*(\boldsymbol{x}) = \log(p(\boldsymbol{x}) \cdot Z_q)$.

Proof. See Appendix B. □

As before, let us parametrize the energy function and variational function as $E_{\boldsymbol{\theta}}$ and $H_{\boldsymbol{\omega}}$ respectively. Based on the new variational representation of f-divergence in Theorem 2, we propose the following f-EBM framework:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\omega}} \mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}; p) \\ = &\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\omega}} \mathbb{E}_{p(\boldsymbol{x})}[f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})))] - \\ &\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))] \end{aligned} \tag{14}$$

For a fixed generative model $q_{\boldsymbol{\theta}}$, from Theorem 2, we know that given enough capacity and training time (*i.e.*, in the non-parametric limit), the optimal solution of the inner maximization problem is:

$$H_{\boldsymbol{\omega}^*}(\boldsymbol{x}) = \log(p(\boldsymbol{x}) \cdot Z_{\boldsymbol{\theta}}) \tag{15}$$

Consequently, the density ratio $p(\boldsymbol{x})/q_{\boldsymbol{\theta}}(\boldsymbol{x})$ between data distribution and model distribution can be recovered by[2]:

$$\begin{aligned} &\exp(H_{\boldsymbol{\omega}^*}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \\ = &\exp(\log(p(\boldsymbol{x})Z_{\boldsymbol{\theta}}))/\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x})) = p(\boldsymbol{x})/q_{\boldsymbol{\theta}}(\boldsymbol{x}) \end{aligned} \tag{16}$$

Plugging the optimal form of $H_{\boldsymbol{\omega}}$ into Equation (14), from Theorem 2, we know that the minimax problem in Equation (14) can be reformulated as:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}^*; p) = \min_{\boldsymbol{\theta}} D_f(p\|q_{\boldsymbol{\theta}}). \tag{17}$$

Hence, the global optimum of the minimax game is achieved if and only if $q_{\boldsymbol{\theta}} = p$ and $H_{\boldsymbol{\omega}} = \log(p \cdot Z_{\boldsymbol{\theta}})$.

In this framework, the maximization over $\boldsymbol{\omega}$ closes the gap between the variational lower bound and the true value of f-divergence, while the minimization over $\boldsymbol{\theta}$ improves the generative model based on the estimated value of the f-divergence. Different from the baseline method described in Section 3.1 (which is based on the original variational representation in Lemma 1), here both terms in the new variational representation of f-divergence (Equation (14)) have non-zero gradients *w.r.t.* $\boldsymbol{\theta}$. In other words, as in contrastive divergence, we are able to evaluate the energy function on both real data from p and synthetic data from $q_{\boldsymbol{\theta}}$. Therefore, the energy function can directly receive the supervision signal from the real data, which bypasses the random exploration problem of the previous formulation.

3.3. Tractable Optimization of f-EBM

To optimize with respect to $\boldsymbol{\omega}$ in Equation (14), we can simply use *i.i.d.* samples from p and $q_{\boldsymbol{\theta}}$ to obtain an unbiased estimation of $\mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}; p)$ and then update $\boldsymbol{\omega}$ with the gradient of the estimated objective using gradient ascent, since the expectations are taken with respect to distributions that do not depend on $\boldsymbol{\omega}$.

However, it is non-trivial to optimize $\boldsymbol{\theta}$ because it is hard to backpropagate gradients through the MCMC sampling process (as discussed in Section 3.1.2) and in the second term of Equation (14), both the sampling distribution and the function within the expectation depend on $\boldsymbol{\theta}$. We derive an unbiased gradient estimator for optimizing $\boldsymbol{\theta}$ in the following theorem.

Theorem 3. *For a fixed $\boldsymbol{\omega}$, under Assumptions 2 and 3 in Appendix C, the gradient of $\mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}; p)$ with respect to $\boldsymbol{\theta}$ can be written as:*

$$\begin{aligned} &\nabla_{\boldsymbol{\theta}} \mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}; p) = \\ &\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} f'(\exp(f_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(x)))] + \\ &\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})] - \\ &\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \cdot \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})] \end{aligned} \tag{18}$$

where $F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x}) := f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))$. When we can obtain i.i.d. samples from p and $q_{\boldsymbol{\theta}}$, we can get an unbiased estimation of the gradient.

Proof. See Appendix C. □

Like adversarial training for implicit density generative models (Goodfellow et al., 2014; Nowozin et al., 2016), in practice, optimizing $H_{\boldsymbol{\omega}}$ to completion in every inner loop of training is computationally prohibitive and would result in overfitting given finite samples. Motivated by the success of single-step alternative gradient updates for training generative adversarial networks, we propose a practical f-EBM

[2] Different from other variational f-divergence minimization frameworks such as f-GAN, in f-EBM, the density ratio is recovered using both the energy function and the variational function.

Algorithm 1 Single-Step f-EBM. See code implementation in Appendix E.1.

```
1: Input: Empirical data distribution $p_{\text{data}}$.
2: Initialize energy function $E_{\boldsymbol{\theta}}$ and variational function $H_{\boldsymbol{\omega}}$.
3: repeat
4:     Draw a minibatch of samples $\mathcal{D}_p$ from $p_{\text{data}}$.
5:     Draw a minibatch of samples $\mathcal{D}_q$ from $q_{\boldsymbol{\theta}}$ (e.g., using Langevin dynamics with a sample replay buffer).
6:     Estimate $\mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}; p)$ with $\mathcal{D}_p$ and $\mathcal{D}_q$.
7:     Perform SGD over $\boldsymbol{\omega}$ with $\nabla_{\boldsymbol{\omega}} \hat{\mathcal{L}}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega})$.
8:     Estimate Equation (18) with $\mathcal{D}_p$ and $\mathcal{D}_q$.
9:     Perform SGD over $\boldsymbol{\theta}$ with $-\nabla_{\boldsymbol{\theta}} \hat{\mathcal{L}}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega})$
10: until Convergence
11: Output: Learned energy-based model $q_{\boldsymbol{\theta}} \propto \exp(-E_{\boldsymbol{\theta}})$ and density ratio estimator $\exp(H_{\boldsymbol{\omega}} + E_{\boldsymbol{\theta}})$.
```

algorithm, presented in Algorithm 1. In Appendix E.1, we also provide a simple PyTorch (Paszke et al., 2019) implementation of Steps 6-9 of Algorithm 1.

As in the training of GANs, although using single-step gradient updates is not generally guaranteed to solve the minimax problem, we empirically observed that such a practical approach enjoys good distribution matching performance and convergence properties (see Appendix F.3 for visualization of the optimization trajectories in a simple setting). This motivates us to conduct a theoretical study on the local convergence property of single-step f-EBM in the next section, and we leave the study of global convergence to future work.

4. Theoretical Analysis

In this section, we conduct a theoretical study on the local convergence property of f-EBM, which shows that under proper conditions, the single-step f-EBM algorithm (simultaneous gradient updates and alternative gradient updates) is locally exponentially stable around good equilibrium points. For readability, we defer rigorous statements of assumptions, theorems and proofs to Appendix D.

The main technical tool we use is the non-linear dynamical systems theory (Hassan, 1996), which have been used to establish the local convergence property of GANs (Nagarajan & Kolter, 2017; Mescheder et al., 2018). Specifically, the linearization theorem (Theorem 4 in Appendix D.1) states that the local convergence property can be analyzed by examining the spectrum of the Jacobian of the dynamical system $\dot{\boldsymbol{\phi}} = v(\boldsymbol{\phi})$: if the Jacobian $\boldsymbol{J} = \partial v(\boldsymbol{\phi}) / \partial \boldsymbol{\phi}|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}$ evaluated at an equilibrium point $\boldsymbol{\phi}^*$ is a Hurwitz matrix (*i.e.*, all the eigenvalues of $\boldsymbol{J}$ have strictly negative real parts), then the system will converge to the equilibrium with a linear convergence rate within some neighborhood of $\boldsymbol{\phi}^*$. In other words, the equilibrium is locally exponentially stable (Definition 3 in Appendix D.1).

To begin with, we first derive the Jacobian of the following differential equations[3]:

$$\begin{pmatrix} \dot{\boldsymbol{\theta}} \\ \dot{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} -\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) \\ \nabla_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) \end{pmatrix} \tag{19}$$

where $V(\boldsymbol{\theta}, \boldsymbol{\omega})$ is a generalized definition of the minimax objective in Equation (14) (see Equation (39) in Appendix D.2).

Theorem 4. *For the dynamical system defined in Equation (39) and the updates defined in Equation (19), under Assumption 4 in Appendix D.2, the Jacobian at an equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ is:*

$$\boldsymbol{J} = \begin{bmatrix} -f''(1)\boldsymbol{K}_{EE} & f''(1)\boldsymbol{K}_{EH} \\ -f''(1)\boldsymbol{K}_{EH}^{\top} & -f''(1)\boldsymbol{K}_{HH} \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{K}_{EE} :=& (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}}^{\top} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \\ & 2\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}^{\top} E_{\boldsymbol{\theta}}(\boldsymbol{x})])|_{\boldsymbol{\theta}^*} \\ \boldsymbol{K}_{EH} :=& (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}}^{\top} H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} \\ \boldsymbol{K}_{HH} :=& (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla_{\boldsymbol{\omega}}^{\top} H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{\boldsymbol{\omega}^*} \end{aligned}$$

Now, we present the main theoretical result:

Theorem 5. *The dynamical system defined in Equation (39) and the updates defined in Equation (19) is locally exponentially stable with respect to an equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ when the Assumptions 4, 5, 6 hold for $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$. Let $\overline{\lambda}(\cdot)$ and $\underline{\lambda}(\cdot)$ denote the largest and smallest eigenvalues of a non-zero positive semi-definite matrix. The rate of convergence is governed only by the eigenvalues λ of the Jacobian $\boldsymbol{J}$ of the system at the equilibrium point, with a strictly negative real part upper bounded as:*

- *When* $\mathrm{Im}(\lambda) = 0$,

$$\mathrm{Re}(\lambda) < \frac{-f''(1)\underline{\lambda}(\boldsymbol{K}_{HH})\underline{\lambda}(\boldsymbol{K}_{EH}\boldsymbol{K}_{EH}^{\top})}{\underline{\lambda}(\boldsymbol{K}_{HH})\overline{\lambda}(\boldsymbol{K}_{EE}, \boldsymbol{K}_{HH}) + \underline{\lambda}(\boldsymbol{K}_{EH}\boldsymbol{K}_{EH}^{\top})} < 0$$

where $\overline{\lambda}(\boldsymbol{K}_{EE}, \boldsymbol{K}_{HH}) := \max(\overline{\lambda}(\boldsymbol{K}_{EE}), \overline{\lambda}(\boldsymbol{K}_{HH}))$

- *When* $\mathrm{Im}(\lambda) \neq 0$,

$$\mathrm{Re}(\lambda) \leq -\frac{f''(1)}{2}(\underline{\lambda}(\boldsymbol{K}_{EE}) + \underline{\lambda}(\boldsymbol{K}_{HH})) < 0$$

Note that these theoretical results hold for any f-divergence with strictly convex and continuously differentiable f. All the proofs for this section can be found in Appendix D.

[3] Following (Nagarajan & Kolter, 2017), we focus on the analysis of continuous time ordinary differential equations, which implies similar results for discrete time updates when the learning rate is sufficiently small.

5. Related Work

As a generalization to maximum likelihood estimation (MLE), methods based on variational f-divergence minimization have been proposed for training implicit generative models (Nowozin et al., 2016; Mohamed & Lakshminarayanan, 2016) and latent variable models (Zhang et al., 2018). Although using general f-divergences to train deep EBMs is novel, researchers have developed related techniques for learning unnormalized statistical models. Hinton (2002) proposed contrastive divergence (CD) to enable maximum likelihood training for EBMs. Inspired by persistent CD (Tieleman, 2008) which propogates Markov chains throughout training, recently Du & Mordatch (2019) proposed techniques to scale CD to high-dimensional data domains, while Nijkamp et al. (2019) proposed to learn non-convergent non-persistent short-run MCMC as an efficient alternative to CD for maximum likelihood training.

Based on the primal-dual view of MLE, doubly dual embedding (Dai et al., 2018) and adversarial dynamics embedding (Dai et al., 2019) were proposed, which introduce a dual sampler to avoid the computation of the partition function. Li et al. (2019) proposed a black-box algorithm to perform maximum likelihood training for Markov random field (MRF), which employs two variational distributions to approximately infer the latent variables and estimate the partition function of an MRF. Noise contrastive estimation (NCE) (Gutmann & Hyvärinen, 2012) train EBMs by performing non-linear logistic regression to classify real data and artificial data from a noise distribution. Theoretically, Riou-Durand & Chopin (2018) proved that when the number of artificial data points approaches infinity, NCE is asymptotically equivalent to MLE. Similar to EBMs, diffusion probabilistic models proposed by (Sohl-Dickstein et al., 2015) define the model distribution as a result of a parametric diffusion process starting from a simple known distribution, which is also trained using MLE.

In this work, to generalize maximum likelihood training of EBMs, we propose a new variational framework, which enables us to train EBMs using any desired f-divergence.

6. Experiments

6.1. Fitting Univariate Mixture of Gaussians

Setup. To understand the effects of different divergences for training EBMs, inspired by (Minka et al., 2005; Nowozin et al., 2016), we first investigate a model misspecification scenario. Specifically, we use an EBM with a quadratic energy function (corresponding to a Gaussian distribution):

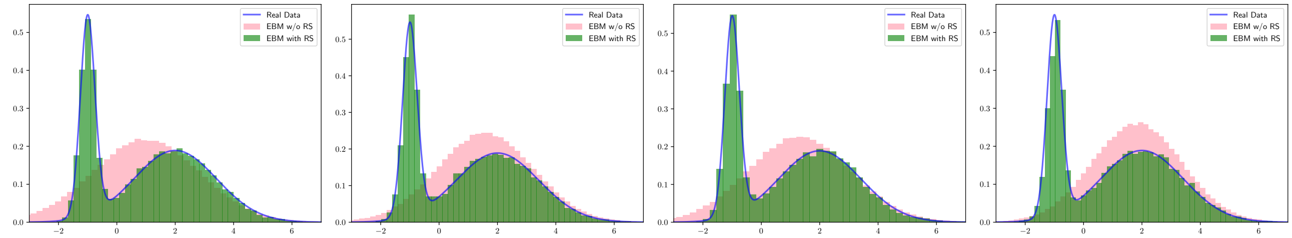
$$E_{\mu,\sigma}(x) = \frac{(x-\mu)^2}{2\sigma^2},\ q_{\mu,\sigma}(x) \propto \exp(-E_{\mu,\sigma}(x)) \quad (20)$$

to fit a mixture of Gaussians (see Figure 1 for illustration). Here μ and σ are two trainable parameters. Although for energy function in Equation (20), the partition function is actually tractable ($Z_{\mu,\sigma} = \sqrt{2\pi\sigma^2}$), for experimental purposes, we will treat it as a general EBM and do not leverage this tractability. For the implementation of the variational function $H_{\boldsymbol{\omega}}$, we use a fully-connected neural network with two hidden layers (each having 64 hidden units) and `tanh` activation functions. To optimize the f-EBM objective, we use single-step gradient updates (Algorithm 1) to alternatively train the energy function and the variational function, with a learning rate of 0.01 and a batch size of 1000.

Parameter Learning Results. Because we know the real data density, we can numerically solve the f-divergence minimization problem and get the ground-truth solution (μ^*, σ^*) induced by the chosen f-divergence. Then, we test whether single-step f-EBM algorithm can learn good parameters $(\hat{\mu}, \hat{\sigma})$ that are close to the desired solution. As shown in Figure 1, and Table 5 in Appendix F.1, single step f-EBM is capable of learning model distributions that closely match the desired solutions for various f-divergences. Furthermore, although all f-divergences are valid objectives (since they are minimized if and only if the model distribution exactly matches the data distribution), in practice the choice of discrepancy measures has a significant influence on the learned distribution because of model misspecification.

Density Ratio Estimation & Bias Correction. As shown in Equation (16), our framework additionally produces a density ratio estimator $\exp(H_{\boldsymbol{\omega}} + E_{\boldsymbol{\theta}})$. First, we visualize the density ratio estimation results in Figure 3 in Appendix F.2, from which we can see that the estimated density ratio is accurate in most areas of the data support, except in the low-density regions where very few training data points come from. Furthermore, as discussed in (Grover et al., 2019), when the learned generative model produces biased statistics relative to the real data distribution, it is possible to correct the bias using estimated density ratio in conjunction with importance sampling or rejection sampling. With the goal of sampling from real data distribution p, we employ rejection sampling (Halton, 1970; Azadi et al., 2018), where we accept samples from the EBM with a probability proportional to the density ratio. As shown in Figure 2, using the learned density ratio, we are able to recover the real data distribution even in such a model misspecification scenario.

Convergence Results. From the optimization trajectories shown in Figure 4 in Appendix F.3, we empirically found that, using single-step alternative gradient updates, f-EBM is able to converge to the desired solution starting from different random initializations.



(a) KL (b) Reverse KL (c) Jensen-Shannon (d) α-Divergence ($\alpha = -1$)

Figure 2. The effect of rejection sampling using learned density ratio ($\exp(H_{\boldsymbol{\omega}}(x) + E_{\boldsymbol{\theta}}(x))$). The blue line represents the desired real data distribution. The green and pink histograms represent the generated samples with and without rejection sampling respectively.

6.2. Modeling Natural Images

In this section, we demonstrate that f-EBMs can produce high-quality images for various choices of f-divergences[4]. Furthermore, we show that energy functions trained by f-EBMs can learn meaningful image representations for image denoising and inpainting tasks.

Setup. We conduct experiments with two commonly used image datasets, CelebA (Liu et al., 2015) and CIFAR-10 (Krizhevsky et al., 2009). Since the performance is sensitive to the model architectures, for fair comparisons, we use the same architecture and training hyper-parameters for f-EBMs and the contrastive divergence baseline (Du & Mordatch, 2019). See Appendix G.9 for more details.

Image Generation. For qualitative evaluation, we show uncurated samples from f-EBMs for CIFAR-10 and CelebA in Appendix G.2 and G.7. These high quality samples demonstrate that the f-EBM framework is effective for training EBMs under a variety of discrepancy measures. For quantitative evaluation, we report Inception (Salimans et al., 2016) and FID (Heusel et al., 2017) scores for CIFAR-10 in Table 1. For CelebA dataset, since different preprocessing may lead to numbers that are not directly comparable, and Heusel et al. (2017) show that Inception score is not a valid metric for CelebA, we only report the FID score for EBMs used in our experiments in Table 2. Table 1 and 2 suggest that we can achieve a significant sample quality improvement with f-EBMs using f-divergences such as Jensen-Shannon, Squared Hellinger and Reverse KL. Moreover, the performance of f-EBM with KL divergence is similar to contrastive divergence, whose underlying discrepancy measure is also KL. To illustrate the Langevin dynamics sampling process, we show how samples evolve from random noise in Appendix G.6 and G.8. For generalization, we show nearest neighbor images of our samples in the training set in Appendix G.5, from which we know that our model is not simply memorizing training images.

Image Denoising & Inpainting. In Appendix G.3 and G.4, we show that energy functions learned by f-EBMs can be used for image inpainting and denoising tasks, where we use Langevin dynamics with the gradients of the learned energy functions to refine the images.

Table 1. Inception and FID scores for CIFAR-10 conditional generation. We compare with results reported by DCGAN (Radford et al., 2015; Wang & Liu, 2016), Improved GAN (Salimans et al., 2016), Fisher GAN (Mroueh & Sercu, 2017), ACGAN (Odena et al., 2017), WGAN-GP (Gulrajani et al., 2017), SNGAN (Miyato et al., 2018), Contrastive Divergence (Du & Mordatch, 2019).

	Method	Inception	FID
GANs	DCGAN	6.58	-
	Improved GAN	8.09	-
	Fisher GAN	8.16	-
	ACGAN	8.25	-
	WGAN-GP	8.42	-
	SNGAN	8.60	17.50
EBMs	Contrastive Divergence (KL)	8.30	37.90
	f-EBM (KL)	8.11 ± .06	37.36
	f-EBM (Reverse KL)	8.49 ± .09	33.25
	f-EBM (Squared Hellinger)	8.57 ± .08	32.19
	f-EBM (Jensen Shannon)	**8.61** ± .06	**30.86**

Table 2. FID for CelebA (32×32) unconditional generation.

	Method	FID
EBMs	Contrastive Divergence (KL)	35.97
	f-EBM (KL)	37.41
	f-EBM (Reverse KL)	26.75
	f-EBM (Jensen Shannon)	26.53
	f-EBM (Squared Hellinger)	**23.67**

7. Discussion and Future Work

In this paper, based on a new variational representation of f-divergences, we propose a general framework termed f-EBM, which enables us to train deep energy-based models using various descrepancy measures within the f-divergence family. We demonstrate the effectiveness of f-EBM both theoretically and empirically. Experimental results on

[4] Our implementation of f-EBM can be found at: https://github.com/ermongroup/f-EBM

CIFAR-10 and CelebA datasets show that our framework can achieve significant sample quality improvement compared to the predominant contrastive divergence method. For future works, there are some other possible ways that may be potentially useful for training EBMs with f-divergences. First, to directly tackle the computational issues of the gradient reparametrization method in Section 3.1.2, we may redefine the generative process as the combination of a proposal distribution and MCMC sampling so that we can jointly train the proposal model and the energy function, and we only need to backpropagate through a few MCMC steps (Lawson et al., 2019). We may also define the generative process as MCMC sampling in the latent space together with a decoder that transforms latent codes to samples (Girolami & Calderhead, 2011; Kumar et al., 2019), such that the dimension of the latent space is much smaller than the sample space and computing Hessians is more efficient. Since these methods rely on careful design of additional components such as a proposal or a decoder, we leave them as interesting avenues for future works.

Acknowledgments

Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. This research was also supported by NSF (#1651565, #1522054, #1733686), ONR (N00014-19-1-2145), AFOSR (FA9550-19-1-0024).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., and Odena, A. Discriminator rejection sampling. *arXiv preprint arXiv:1810.06758*, 2018.

Dai, B., Dai, H., Gretton, A., Song, L., Schuurmans, D., and He, N. Kernel exponential family estimation via doubly dual embedding. *arXiv preprint arXiv:1811.02228*, 2018.

Dai, B., Liu, Z., Dai, H., He, N., Gretton, A., Song, L., and Schuurmans, D. Exponential family estimation via adversarial dynamics embedding. *arXiv preprint arXiv:1904.12083*, 2019.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pp. 881–889, 2015.

Girolami, M. and Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Grover, A., Song, J., Agarwal, A., Tran, K., Kapoor, A., Horvitz, E., and Ermon, S. Bias correction of learned generative models using likelihood-free importance weighting. *arXiv preprint arXiv:1906.09531*, 2019.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361, 2012.

Halton, J. H. A retrospective and prospective survey of the monte carlo method. *Siam review*, 12(1):1–63, 1970.

Hassan, K. Khalil, nonlinear systems. *Prentice-Hall, Inc., New Jersey*, 1996.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

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

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pp. 4565–4573, 2016.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

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

Kumar, R., Goyal, A., Courville, A., and Bengio, Y. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.

Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–37, 2011.

Lawson, J., Tucker, G., Dai, B., and Ranganath, R. Energy-inspired models: Learning with sampler-induced distributions. In *Advances in Neural Information Processing Systems*, pp. 8499–8511, 2019.

Li, C., Du, C., Xu, K., Welling, M., Zhu, J., and Zhang, B. To relieve your headache of training an mrf, take advil. *arXiv preprint arXiv:1901.08400*, 2019.

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

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

Minka, T. et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Mohamed, S. and Lakshminarayanan, B. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Mroueh, Y. and Sercu, T. Fisher gan. In *Advances in Neural Information Processing Systems*, pp. 2513–2523, 2017.

Nagarajan, V. and Kolter, J. Z. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems*, pp. 5585–5595, 2017.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

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

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pp. 271–279, 2016.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2642–2651. JMLR. org, 2017.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016a.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016b.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Poon, H. and Domingos, P. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 689–690. IEEE, 2011.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Riou-Durand, L. and Chopin, N. Noise contrastive estimation: asymptotics, comparison with mc-mle. *arXiv preprint arXiv:1801.10381*, 2018.

Robert, C. and Casella, G. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Tieleman, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071, 2008.

Turner, R. Cd notes. 2005.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pp. 4790–4798, 2016.

Wang, D. and Liu, Q. Learning to draw samples: With application to amortized mle for generative adversarial learning. *arXiv preprint arXiv:1611.01722*, 2016.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Zhang, M., Bird, T., Habib, R., Xu, T., and Barber, D. Training generative latent models by variational f-divergence minimization. 2018.