Pseudo-Spherical Contrastive Divergence

Anonymous Author(s)
Affiliation
Address
email

Abstract

Energy-based models (EBMs) offer flexible distribution parametrization. However, due to the intractable partition function, they are typically trained via contrastive divergence for maximum likelihood estimation. In this paper, we propose pseudo-spherical contrastive divergence (PS-CD) to generalize maximum likelihood learning of EBMs. PS-CD is derived from the maximization of a family of strictly proper homogeneous scoring rules, which avoids the computation of the intractable partition function and provides a generalized family of learning objectives that include contrastive divergence as a special case. Moreover, PS-CD allows us to flexibly choose various learning objectives to train EBMs without additional computational cost or variational minimax optimization. Theoretical analysis on the proposed method and experiments on both synthetic data and commonly used image datasets demonstrate the effectiveness of PS-CD and its superiority over maximum likelihood and f-EBMs. Based on a set of recently proposed indicative generative model evaluation metrics, we also provide an analysis on the modeling tradeoffs of different objectives in the PS-CD family on image generation tasks, justifying its modeling flexibility.

1 Introduction

Energy-based models (EBMs) provide a unified framework for generative and discriminative learning by capturing dependencies between random variables with an energy function. Due to the absence of the normalization constraint, EBMs offer much more flexibility in distribution parametrization and architecture design compared to properly normalized probabilistic models such as autoregressive models [51, 22], flow-based models [13, 14, 44] and sum-product networks [69]. Recently, deep EBMs have achieved considerable success in realistic image generation [16, 63, 11, 29], molecular modeling [87] and model-based planning [15], thanks to modern deep neural networks [52, 47, 34] for parametrizing expressive energy functions and improved Markov Chain Monte Carlo (MCMC) techniques [60, 72, 39, 16, 63] for efficiently sampling from EBMs.

Training EBMs consists of finding an energy function that assigns low energies to correct configurations of variables and high energies to incorrect ones [53], where a central concept is the *loss functional* that is used to measure the quality of the energy function and is minimized during training. The flexibility of EBMs does not come for free: it makes the design of loss functionals particularly challenging, as it usually involves the partition function that is generally intractable to compute. As a result, EBMs are typically trained via CD [36], which belongs to the "analysis by synthesis" scheme [30] and performs a sampling-based estimation of the gradient of KL between data distribution and energy-based distribution. Since different loss functionals will induce different solutions in practice (when the model is mis-specified and data is finite) and KL may not provide the right inductive bias [24, 88], inspired by the great success of implicit generative models [27, 64, 3], [86] proposed a variational framework to train EBMs by minimizing general f-divergences [9]. Although this framework enables us to specify various modeling preferences such as diversity/quality tradeoff, they rely on learning additional components (variational functions) within a minimax framework,

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

where the optimization is complicated by the notion of Nash equilibrium and local optimality [41] and suffers from instability and non-convergence issues [57]. Along this line, noise contrastive estimation (NCE) [33] can train EBMs with a family of loss functionals induced by different Bregman divergences. However, in practice, NCE usually relies on carefully-designed noise distribution such as context-dependent noise distribution [40] or joint learning of a flow-based noise distribution [20].

In this paper, we draw inspiration from statistical decision theory [12] and propose a novel perspective for designing loss functionals for training EBMs without involving auxiliary models or variational optimization. Specifically, to generalize maximum likelihood training of EBMs, we focus on maximizing pseudo-spherical scoring rules [73, 26], which are *strictly proper* such that the data distribution is the unique optimum and *homogeneous* such that they can be evaluated without the knowledge of the normalization constant. Under the "analysis by synthesis" scheme used in CD and f-EBM [86], we then derive a practical algorithm termed Pseudo-Spherical Contrastive Divergence (PS-CD). Different from f-EBM, PS-CD enables us to specify flexible modeling preferences without requiring additional computational cost or unstable minimax optimization. We provide a theoretical analysis on the sample complexity and convergence property of PS-CD, as well as its connections to maximum likelihood training. With experiments on both synthetic data and commonly used image datasets, we show that PS-CD achieves significant sample quality improvement over conventional maximum likelihood training and competitive performance to f-EBM without expensive variational optimization. Based on a set of recently proposed generative model evaluation metrics [59], we further demonstrate the various modeling tradeoffs enabled by PS-CD, justifying its modeling flexibility.

2 Preliminaries

2.1 Energy-Based Distribution Representation and Sampling

Given a set of *i.i.d.* samples $\{\boldsymbol{x}_i\}_{i=1}^N$ from some unknown data distribution $p(\boldsymbol{x})$ defined over the sample space $\mathcal{X} \subset \mathbb{R}^m$, the goal of generative modeling is to learn a $\boldsymbol{\theta}$-parametrized probability distribution $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ to approximate the data distribution $p(\boldsymbol{x})$. In the context of energy-based modeling, instead of directly parametrizing a properly normalized distribution, we first parametrize an unnormalized energy function $E_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$, which further defines a normalized probability density via the Boltzmann distribution:

$$q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\overline{q}_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_{\boldsymbol{\theta}}} = \frac{\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))}{Z_{\boldsymbol{\theta}}}, \tag{1}$$

where $Z_{\boldsymbol{\theta}} := \int_{\mathcal{X}} \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$ is the partition function (normalization constant). In this paper, unless otherwise stated, we will use $\overline{q}$ to denote an unnormalized density and q to denote the corresponding normalized distribution. We also assume that the exponential of the negative energy belongs to the L^1 space, $\mathcal{E} := \left\{E_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} < \infty\right\}$, *i.e.*, $Z_{\boldsymbol{\theta}}$ is finite.

Since energy-based models (EBMs) represent a probability distribution by assigning unnormalized scalar values (energies) to the data points, we can use any model architecture that outputs a bounded real number given an input to implement the energy function, which allows extreme flexibility in distribution parametrization. However, it is non-trivial to sample from an EBM, usually requiring MCMC [72] techniques. Specifically, in this work we consider using Langevin dynamics [60, 85], a gradient-based MCMC method that performs noisy gradient descent to traverse the energy landscape and arrive at the low-energy configurations:

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

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 [85]. In order to sample from an energy-based distribution efficiently, many scalable techniques have been proposed such as learning non-convergent, non-persistent, short-run MCMC [63] and using a sample replay buffer to improve mixing time and sample diversity [16]. In this work, we leverage these recent advances when we need to obtain samples from an EBM.

2.2 Maximum Likelihood Training of EBMs via Contrastive Divergence

The predominant approach to training explicit density generative models is to approximately minimize the KL divergence between the (empirical) data distribution and model distribution. Minimizing KL

divergence is equivalent to the following maximum likelihood estimation (MLE) objective:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{MLE}}(\boldsymbol{\theta}; p) = \min_{\boldsymbol{\theta}} -\mathbb{E}_{p(\boldsymbol{x})}\left[\log q_{\boldsymbol{\theta}}(\boldsymbol{x})\right] = \min_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{x})}[E_{\boldsymbol{\theta}}(\boldsymbol{x})] + \log Z_{\boldsymbol{\theta}}. \tag{3}$$

Because of the intractable partition function (an integral over the sample space), we cannot directly optimize the above MLE objective. To tackle this issue, [36] proposed contrastive divergence (CD) algorithm as a convenient way to estimate the gradient of $\mathcal{L}_{\mathrm{MLE}}(\boldsymbol{\theta}; p)$ using samples from $q_{\boldsymbol{\theta}}$:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{MLE}}(\boldsymbol{\theta}; p) = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] + \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})], \tag{4}$$

which can be interpreted as decreasing the energies of real data from p and increasing the energies of fake data generated by $q_{\boldsymbol{\theta}}$. As discussed above, evaluating Equation (4) typically relies on MCMC methods such as the Langevin dynamics sampling procedure defined in Equation (2) to produce samples from the model distribution $q_{\boldsymbol{\theta}}$, which induces a surrogate gradient estimation:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CD}-K}(\boldsymbol{\theta}; p) = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}^{K}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})], \tag{5}$$

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 (4) corresponds to $\mathcal{L}_{\mathrm{CD}-\infty}$.

2.3 Strictly Proper Scoring Rules

Stemming from statistical decision theory [12], scoring rules evaluate the quality of probabilistic forecasts by assigning numerical scores based on the predictive distributions and the events that materialize. Formally, consider a compact sample space $\mathcal{X}$. Let $\mathcal{M}$ be a space of all locally 1-integrable non-negative finite measures and $\mathcal{P}$ be a subspace consisting of all probability measures on the sample space $\mathcal{X}$. A scoring rule $S(\boldsymbol{x}, q)$ specifies the *utility* of forecasting using a probability forecast $q \in \mathcal{P}$ for a given sample $\boldsymbol{x} \in \mathcal{X}$. With slightly abused notation, we write the *expected score* of $S(\boldsymbol{x}, q)$ under a reference distribution p as:

$$S(p, q) := \mathbb{E}_{p(\boldsymbol{x})}[S(\boldsymbol{x}, q)]. \tag{6}$$

Definition 1 (Proper Scoring Rules [25])**.** *A scoring rule $S : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ is called proper relative to $\mathcal{P}$ if the corresponding expected score satisfies:*

$$\forall p, q \in \mathcal{P}, S(p, q) \leq S(p, p). \tag{7}$$

It is strictly proper if the equality holds if and only if $q = p$.

In prediction and elicitation problems, strictly proper scoring rules encourage the forecaster to make honest predictions based on their true beliefs [21]. In estimation problems, where we want to approximate a distribution p with another parametric distribution $q_{\boldsymbol{\theta}}$, strictly proper scoring rules provide attractive learning objectives:

$$\arg\max_{q_{\boldsymbol{\theta}} \in \mathcal{P}_{\boldsymbol{\theta}}} S(p, q_{\boldsymbol{\theta}}) = \arg\max_{q_{\boldsymbol{\theta}} \in \mathcal{P}_{\boldsymbol{\theta}}} \mathbb{E}_{p(\boldsymbol{x})}[S(\boldsymbol{x}, q_{\boldsymbol{\theta}})] = p \text{ (when } p \in \mathcal{P}_{\boldsymbol{\theta}}). \tag{8}$$

When a scoring rule S is strictly proper relative to $\mathcal{P}$, the associated *generalized entropy function* and *divergence function* are defined as:

$$G(p) := \sup_{q \in \mathcal{P}} S(p, q) = S(p, p), \quad D(p, q) := S(p, p) - S(p, q) \geq 0. \tag{9}$$

$G(p)$ is convex and represents the maximally achievable utility, while $D(p, q)$ is the Bregman divergence [6] associated with the convex function G and the equality holds only when $p = q$.

Next, we introduce a specific kind of scoring rules that are particularly suitable for learning unnormalized statistical models.

Definition 2 (Homogeneous Scoring Rules [66])**.** *A scoring rule is homogeneous if it satisfies (here the domain of the score function is extended to $\mathcal{X} \times \mathcal{M}$):*

$$\forall \lambda > 0, \boldsymbol{x} \in \mathcal{X}, \;\; S(\boldsymbol{x}, q) = S(\boldsymbol{x}, \lambda \cdot q). \tag{10}$$

Since scaling the model distribution q by a positive constant λ does not change the value of a homogeneous scoring rule, such homogeneity allows us to evaluate it without computing the intractable partition function of an energy-based distribution. Thus, strictly proper and homogeneous scoring rules are natural candidates for new training objectives of EBMs.

Example 1. *A notable example of scoring rules is the widely used logarithm score:* $S(\boldsymbol{x}, q) = \log q(\boldsymbol{x})$. *The associated generalized entropy is the negative Shannon entropy:* $G(p) = \mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{x})]$, *and the associated Bregman divergence is the KL divergence:* $D(p, q) = \mathbb{E}_{p(\boldsymbol{x})}[\log(p(\boldsymbol{x})/q(\boldsymbol{x}))]$. *From Definitions 1 and 2, we know that the logarithm score is strictly proper but not homogeneous. Specifically, for a* $\boldsymbol{\theta}$*-parametrized energy-based distribution* $q_{\boldsymbol{\theta}} = \overline{q}_{\boldsymbol{\theta}}/Z_{\boldsymbol{\theta}}$, *since* $S(\boldsymbol{x}, \overline{q}_{\boldsymbol{\theta}}) = S(\boldsymbol{x}, Z_{\boldsymbol{\theta}} \cdot q_{\boldsymbol{\theta}}) = S(\boldsymbol{x}, q_{\boldsymbol{\theta}}) + \log Z_{\boldsymbol{\theta}} \neq S(\boldsymbol{x}, q_{\boldsymbol{\theta}})$ *and* $\log Z_{\boldsymbol{\theta}}$ *cannot be ignored during the optimization of* $\boldsymbol{\theta}$, *we need to use tailored methods such as contrastive divergence [36] or doubly dual embedding [10] to tackle the intractable partition function.*

3 Training EBMs by Maximizing Homogeneous Scoring Rules

In this section, we derive a new principle for training EBMs from the perspective of optimizing strictly proper homogeneous scoring rules. All proofs for this section can be found in Appendix B.

3.1 Pseudo-Spherical Scoring Rule

In this section, we introduce the pseudo-spherical scoring rule, which is a representative family of strictly proper homogeneous scoring rules that have great potentials for training deep energy-based models and allow flexible and convenient specification of modeling preferences, yet have not been explored before in the context of energy-based generative modeling.

Definition 3 (Pseudo-Spherical Scoring Rule [73, 26])**.** *For* $\gamma > 0$, *the pseudo-spherical scoring rule is defined as:*

$$S(\boldsymbol{x}, q) := \frac{q(\boldsymbol{x})^{\gamma}}{(\int_{\mathcal{X}} q(\boldsymbol{y})^{\gamma+1} \mathrm{d}\boldsymbol{y})^{\frac{\gamma}{\gamma+1}}} = \frac{\overline{q}(\boldsymbol{x})^{\gamma}}{(\int_{\mathcal{X}} \overline{q}(\boldsymbol{y})^{\gamma+1} \mathrm{d}\boldsymbol{y})^{\frac{\gamma}{\gamma+1}}} = \left(\frac{\overline{q}(\boldsymbol{x})}{\|\overline{q}\|_{\gamma+1}}\right)^{\gamma} \tag{11}$$

where $\|\overline{q}\|_{\gamma+1} := \left(\int_{\mathcal{X}} \overline{q}(\boldsymbol{y})^{\gamma+1} \mathrm{d}\boldsymbol{y}\right)^{\frac{1}{\gamma+1}}$.

The expected pseudo-spherical score under a reference distribution p *is defined as:*

$$S_{ps}(p, q) := \mathbb{E}_{p(\boldsymbol{x})}[S(\boldsymbol{x}, q)] = \frac{\mathbb{E}_{p(\boldsymbol{x})}[\overline{q}(\boldsymbol{x})^{\gamma}]}{(\int_{\mathcal{X}} \overline{q}(\boldsymbol{y})^{\gamma+1} \mathrm{d}\boldsymbol{y})^{\frac{\gamma}{\gamma+1}}} \tag{12}$$

Example 2. *The classic spherical scoring rule [18] is a special case in the pseudo-spherical family, which corresponds to* $\gamma = 1$*:*

$$S(\boldsymbol{x}, q) = \frac{\overline{q}(\boldsymbol{x})}{(\int_{\mathcal{X}} \overline{q}(\boldsymbol{y})^{2} \mathrm{d}\boldsymbol{y})^{\frac{1}{2}}} = \frac{\overline{q}(\boldsymbol{x})}{\|\overline{q}\|_2} \tag{13}$$

The family of pseudo-spherical scoring rules is appealing because it introduces a different and principled way for assessing a probability forecast. For example, the spherical scoring rule has an interesting geometric interpretation. Suppose the sample space $\mathcal{X}$ contains n mutually exclusive and exhaustive outcomes. Then a probability forecast can be represented as a vector $\boldsymbol{q} = (q_1, \ldots, q_n)$. Let vector $\boldsymbol{p} = (p_1, \ldots, p_n)$ represent the oracle probability forecast. The expected spherical score can be written as:

$$S(p, q) = \mathbb{E}_{p(\boldsymbol{x})}[S(\boldsymbol{x}, q)] = \frac{\sum_i p_i q_i}{\sqrt{\sum_i q_i^2}} = \|\boldsymbol{p}\|_2 \frac{\langle \boldsymbol{p}, \boldsymbol{q} \rangle}{\|\boldsymbol{p}\|_2 \|\boldsymbol{q}\|_2} = \|\boldsymbol{p}\|_2 \cos(\angle(\boldsymbol{p}, \boldsymbol{q})) \tag{14}$$

where $\langle \boldsymbol{p}, \boldsymbol{q} \rangle$ and $\angle(\boldsymbol{p}, \boldsymbol{q})$ denote the inner product and the angle between vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ respectively. In other words, when we want to evaluate the expected spherical score of a probability forecast $\boldsymbol{q}$ under real data distribution $\boldsymbol{p}$ using samples, the angle between $\boldsymbol{p}$ and $\boldsymbol{q}$ is the sufficient statistics. Since we know that both $\boldsymbol{p}$ and $\boldsymbol{q}$ belong to the probability simplex $\mathcal{P} = \{\boldsymbol{v} | \sum_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{v}(\boldsymbol{x}) = 1 \text{ and } \forall \boldsymbol{x} \in \mathcal{X}, \boldsymbol{v}(\boldsymbol{x}) \geq 0.\}$, the expected score will be minimized if and only if the angle is zero, which implies $\boldsymbol{p} = \boldsymbol{q}$. More importantly, since all we need to do is to minimize the angle of the deviation, we are allowed to scale $\boldsymbol{q}$ by a constant. Specifically, when $\boldsymbol{q}$ is an energy-based distribution $\boldsymbol{q} = \left(\frac{\exp(-E_1)}{\sum_i \exp(-E_i)}, \ldots, \frac{\exp(-E_n)}{\sum_i \exp(-E_i)}\right)$, we can instead evaluate and minimize the angle between data distribution $\boldsymbol{p}$ and the unnormalized distribution $\overline{\boldsymbol{q}} = (\exp(-E_1), \ldots, \exp(-E_n))$, since $\angle(\boldsymbol{p}, \boldsymbol{q}) = \angle(\boldsymbol{p}, \overline{\boldsymbol{q}})$. More generally, we have the following theorem to justify the use of pseudo-spherical scoring rules for training energy-based models:

Theorem 1 ([25, 67])**.** *Pseudo-spherical scoring rule is strictly proper and homogeneous.*

As the original definition of pseudo-spherical scoring rule (Equation (11)) takes the form of a fraction, for computational considerations, in this paper we instead focus on optimizing its *composite scoring rule* (Definition 2.1 in [42]):

Definition 4 (γ-score [19])**.** *For the expected pseudo-spherical score $S_{ps}(p,q)$ defined in Equation (12) with $\gamma > 0$, the expected γ-score is defined as:*

$$S_\gamma(p,q) := \frac{1}{\gamma}\log(S_{ps}(p,q)) = \frac{1}{\gamma}\log\left(\mathbb{E}_{p(\boldsymbol{x})}[\overline{q}(\boldsymbol{x})^\gamma]\right) - \log(\|\overline{q}\|_{\gamma+1}) \tag{15}$$

Since $\frac{1}{\gamma}\log(u)$ is strictly increasing in u, $S_\gamma(p,q)$ is a strictly proper homogeneous composite score:

$$\arg\max_{q\in\mathcal{P}} S_\gamma(p,q) = \arg\max_{q\in\mathcal{P}} \frac{1}{\gamma}\log(S_{ps}(p,q)) = \arg\max_{q\in\mathcal{P}} S_{ps}(p,q) = p. \tag{16}$$

3.2 Pseudo-Spherical Contrastive Divergence

Suppose we parametrize the energy-based model distribution as $q_{\boldsymbol{\theta}} \propto \overline{q}_{\boldsymbol{\theta}} = \exp(-E_{\boldsymbol{\theta}})$ and we want to minimize the negative γ-score in Equation (15):

$$\min_{\boldsymbol{\theta}} \mathcal{L}_\gamma(\boldsymbol{\theta};p) = \min_{\boldsymbol{\theta}} -\frac{1}{\gamma}\log\left(\mathbb{E}_{p(\boldsymbol{x})}[\overline{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^\gamma]\right) + \log(\|\overline{q}_{\boldsymbol{\theta}}\|_{\gamma+1}) \tag{17}$$

In the following theorem, we derive the gradient of $\mathcal{L}_\gamma(\boldsymbol{\theta};p)$ with respect to $\boldsymbol{\theta}$:

Theorem 2. *The gradient of $\mathcal{L}_\gamma(\boldsymbol{\theta};p)$ with respect to $\boldsymbol{\theta}$ can be written as:*

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_\gamma(\boldsymbol{\theta};p) = -\frac{1}{\gamma}\nabla_{\boldsymbol{\theta}}\log\left(\mathbb{E}_{p(\boldsymbol{x})}[\exp(-\gamma E_{\boldsymbol{\theta}}(\boldsymbol{x}))]\right) - \mathbb{E}_{r_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})] \tag{18}$$

where the auxiliary distribution $r_{\boldsymbol{\theta}}$ is also an energy-based distribution defined as:

$$r_{\boldsymbol{\theta}}(\boldsymbol{x}) := \frac{\overline{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\gamma+1}}{\int_{\mathcal{X}}\overline{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\gamma+1}\mathrm{d}\boldsymbol{x}} = \frac{\exp(-(\gamma+1)E_{\boldsymbol{\theta}}(\boldsymbol{x}))}{\int_{\mathcal{X}}\exp(-(\gamma+1)E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}.$$

In App. B.1, we provide two different ways to prove the above theorem. The first one is more straightforward and directly differentiates through the term $\log(\|\overline{q}_{\boldsymbol{\theta}}\|_{\gamma+1})$. The second one leverages a variational representation of $\log(\|\overline{q}_{\boldsymbol{\theta}}\|_{\gamma+1})$, where the optimal variational distribution happens to take an analytical form of $r^*_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \overline{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{\gamma+1}$, thus avoiding the minimax optimization in other variational frameworks such as [86, 10, 11]. The main challenge in maximizing γ-score is that it is generally intractable to exactly compute the gradient of the second term in Equation (15).

During training, estimating the second term of Equation (18) requires us to obtain samples from the auxiliary distribution $r_{\boldsymbol{\theta}} \propto \exp(-(\gamma+1)E_{\boldsymbol{\theta}})$, while at test time, we want to sample from the model distribution $q_{\boldsymbol{\theta}} \propto \exp(-E_{\boldsymbol{\theta}})$ that approximates the data distribution. Due to the restrict regularity conditions on the convergence of Langevin dynamics, in practice, we found it challenging to use the iterative sampling process in Equation (2) with a fixed number of transition steps and step size to produce samples from $r_{\boldsymbol{\theta}}$ and $q_{\boldsymbol{\theta}}$ simultaneously, as the temperature $\gamma+1$ in $r_{\boldsymbol{\theta}}$ simply amounts to a linear rescaling to the energy function during training. Thus for generality, as in contrastive divergence [36, 16, 63] and f-EBM [86], we make the minimal assumption that we only have a sampling procedure to produce samples from $q_{\boldsymbol{\theta}}$ for both learning and inference procedures.

In this case, we can leverage the analytical form of $r_{\boldsymbol{\theta}}$ and *self-normalized importance sampling* [65] (which has been used to derive gradient estimators in other contexts such as importance weighted autoencoder [7, 17]) to obtain a consistent estimation of Equation (18):

Theorem 3. *Let $\boldsymbol{x}_1^+,\ldots,\boldsymbol{x}_N^+$ be i.i.d. samples from $p(\boldsymbol{x})$ and $\boldsymbol{x}_1^-,\ldots,\boldsymbol{x}_N^-$ be i.i.d. samples from $q_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))$. Define the gradient estimator as:*

$$\nabla_{\boldsymbol{\theta}}\widehat{\mathcal{L}_\gamma^N(\boldsymbol{\theta}}};p) := -\nabla_{\boldsymbol{\theta}}\frac{1}{\gamma}\log\left(\frac{1}{N}\sum_{i=1}^N\exp(-\gamma E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+))\right) - \frac{\sum_{i=1}^N\omega_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)}{\sum_{i=1}^N\omega_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)} \tag{19}$$

where the self-normalized importance weight $\omega_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-) := \overline{r}_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)/\overline{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-) = \exp(-\gamma E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-))$. Then the gradient estimator converges to the true gradient (Equation (18)) in probability, i.e., $\forall\epsilon > 0$:

$$\lim_{N\to\infty}\mathbb{P}\left(\left\|\nabla_{\boldsymbol{\theta}}\widehat{\mathcal{L}_\gamma^N(\boldsymbol{\theta}}};p) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_\gamma(\boldsymbol{\theta};p)\right\| \geq \epsilon\right) = 0.$$

Algorithm 1 Pseudo-Spherical Contrastive Divergence.

1: **Input:** Empirical data distribution p_{data}. Pseudo-spherical scoring rule hyperparameter γ.
2: Initialize energy function $E_{\boldsymbol{\theta}}$.
3: **repeat**
4: Draw a minibatch of samples $\{\boldsymbol{x}_1^+, \ldots, \boldsymbol{x}_N^+\}$ from p_{data}.
5: Draw a minibatch of samples $\{\boldsymbol{x}_1^-, \ldots, \boldsymbol{x}_N^-\}$ from $q_{\boldsymbol{\theta}} \propto \exp(-E_{\boldsymbol{\theta}})$ (*e.g.*, using Langevin dynamics with a sample replay buffer).
6: Update the energy function by stochastic gradient descent:

$$\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}_\gamma^N(\boldsymbol{\theta}}; p) = -\nabla_{\boldsymbol{\theta}} \frac{1}{\gamma} \log \left(\frac{1}{N} \sum_{i=1}^N \exp(-\gamma E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)) \right) - \frac{\sum_{i=1}^N \exp(-\gamma E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)}{\sum_{i=1}^N \exp(-\gamma E_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-))}$$

7: **until** Convergence

We summarize the pseudo-spherical contrastive divergence (PS-CD) training procedure in Algorithm 1. In Appendix A, we also provide a simple PyTorch implementation for stochastic gradient descent (SGD) with the gradient estimator in Equation (19).

3.3 Connections to Maximum Likelihood Estimation and Extension to $\gamma < 0$

From Equation (9) in Section 2.3, we know that γ-score induces the following Bregman divergence (the divergence function associated with proper composite scoring rule is analogously defined in Def. 2.1 in [42]):

$$D_\gamma(p, q_{\boldsymbol{\theta}}) = S_\gamma(p, p) - S_\gamma(p, q_{\boldsymbol{\theta}})$$

and maximizing γ-score is equivalent to minimizing $D_\gamma(p, q_{\boldsymbol{\theta}})$. In the following lemma, we show that when $\gamma \to 0$, $D_\gamma(p, q_{\boldsymbol{\theta}})$ will recover the KL divergence between p and $q_{\boldsymbol{\theta}}$, and the gradient of PS-CD will recover the gradient of CD.

Lemma 1. *When $\gamma \to 0$, we have:*

$$\lim_{\gamma \to 0} D_\gamma(p, q_{\boldsymbol{\theta}}) = D_{\text{KL}}(p \| q_{\boldsymbol{\theta}}); \quad \lim_{\gamma \to 0} \nabla_{\boldsymbol{\theta}} \mathcal{L}_\gamma(\boldsymbol{\theta}; p) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{MLE}}(\boldsymbol{\theta}; p).$$

Inspired by [83, 54] that generalize Rényi divergence beyond its definition to negative orders, we now consider the extension of γ-score with $\gamma < 0$ (although it may not be strictly proper for these γ values). The following lemma shows that maximizing such scoring rule amounts to maximizing a lower bound of logarithm score (MLE) with an additional Rényi entropy regularization.

Lemma 2. *When $-1 \leq \gamma < 0$, we have:*

$$S_\gamma(p, q) \leq \mathbb{E}_{p(\boldsymbol{x})}[\log q(\boldsymbol{x})] + \frac{\gamma}{\gamma + 1} \mathcal{H}_{\gamma+1}(q)$$

where $\mathcal{H}_{\gamma+1}(q)$ is the Rényi entropy of order $\gamma + 1$.

4 Theoretical Analysis

In this section, to gain a deeper understanding of our PS-CD algorithm and how the proposed estimator behaves, we analyze its sample complexity and convergence property under certain conditions. All the proofs for this section can be found in Appendix C.

4.1 Sample Complexity

We start with analyzing the sample complexity of the consistent gradient estimator in Equation (19), that is how fast it approaches the true gradient value. We first make the following assumption:

Assumption 1. *The energy function is bounded by K and the gradient is bounded by L:*

$$\forall \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta,\ |E_{\boldsymbol{\theta}}(\boldsymbol{x})| \leq K,\ \|\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\| \leq L.$$

This assumption is typically easy to satisfy because in practice we always use a bounded sample space (*e.g.* normalizing images to [0,1] or truncated Gaussian) to ensure stability. For example, in image modeling experiments, we use L_2 regularization on the outputs of the energy function (hence bounded energy values), as well as normalized inputs and spectral normalization [58] for the neural network that realizes the energy function (hence Lipschitz continuous with bounded gradient).

With vector Bernstein inequality [45, 31], we have the following theorem showing a sample complexity of $O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ such that the estimation error is less than ϵ with probability at least $1-\delta$:

Theorem 4. *For any constants $\epsilon > 0$ and $\delta \in (0,1)$, when the number of samples N satisfies $N \geq \frac{32L^2 e^{8\gamma K}(1+4\log(2/\delta))}{\epsilon^2}$, we have $\mathbb{P}\left(\left\|\nabla_{\boldsymbol{\theta}}\widehat{\mathcal{L}_\gamma^N}(\boldsymbol{\theta};p) - \nabla_{\boldsymbol{\theta}}\mathcal{L}_\gamma(\boldsymbol{\theta};p)\right\| \leq \epsilon\right) \geq 1-\delta$.*

4.2 Convergence

Typically, convergence of SGD are analyzed for unbiased gradient estimators, while the gradient estimator in PS-CD is asymptotically consistent but biased. Building on the sample complexity bound above and prior theoretical works for analyzing SGD [8, 23], we analyze the convergence of PS-CD. For notational convenience, we use $\mathcal{L}(\boldsymbol{\theta})$ to denote the loss function $\mathcal{L}_\gamma(\boldsymbol{\theta};p) = -S_\gamma(p, q_{\boldsymbol{\theta}})$. Besides Assumption 1, we further make the following assumption on the smoothness of $\mathcal{L}(\boldsymbol{\theta})$:

Assumption 2. *The loss function $\mathcal{L}(\boldsymbol{\theta})$ is M-smooth (with $M > 0$):*

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta,\ \|\nabla\mathcal{L}(\boldsymbol{\theta}_1) - \nabla\mathcal{L}(\boldsymbol{\theta}_2)\| \leq M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

This is a common assumption for analyzing first-order optimization methods, which is also used in [23, 8]. Also this is a relatively mild assumption since we do not require the loss function to be convex in $\boldsymbol{\theta}$. Since in non-convex optimization, the convergence criterion is typically measured by gradient norm, following [62, 23], we use $\|\nabla\mathcal{L}(\boldsymbol{\theta})\| \leq \xi$ to judge whether a solution $\boldsymbol{\theta}$ is approximately a stationary point.

Theorem 5. *For any constants $\alpha \in (0,1)$ and $\delta \in (0,1)$, suppose that the step sizes satisfy $\eta_t < 2(1-\alpha)/M$ and the sample size N_t used for estimating $\widehat{\boldsymbol{g}}_t$ is sufficiently large (satisfying Equation (36)). Let $\mathcal{L}^*$ denote the minimum value of $\mathcal{L}(\boldsymbol{\theta})$. Then with probability at least $1-\delta$, the output of Algorithm 2 (in Appendix C.2), $\widehat{\boldsymbol{\theta}}$, satisfies (constant $C := \alpha M L^2 e^{4\gamma K}$):*

$$\mathbb{E}[\|\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}})\|^2] < \frac{2(\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}^*) + 12C\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T (2(1-\alpha)\eta_t - M\eta_t^2)}$$

The above theorem implies the following corollary, which shows a typical convergence rate of $O(1/\sqrt{T})$ for non-convex optimization problems:

Corollary 1. *Under the conditions in Theorem 5 except that we use constant step sizes: $\eta_t = \min\{(1-\alpha)/M, 1/\sqrt{T}\}$ for $t = 1, \ldots, T$. Then with probability at least $1-\delta$, we have (constant $C := \alpha M L^2 e^{4\gamma K}$):*

$$\mathbb{E}[\|\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}})\|^2] < \frac{2M(\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}^*)}{(1-\alpha)^2 T} + \frac{2(\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}^*) + 12C}{(1-\alpha)\sqrt{T}}$$

In Appendix C.2, we discuss more on the strongly convex (Theorem 7) and convex cases (Theorem 8).

5 Related Work

Direct KL Minimization. Under the "analysis by synthesis" scheme, [36] proposed Contrastive Divergence (CD), which estimates the gradient of the log-partition function (arising from KL) using samples from some MCMC procedure. To improve the mixing time of MCMC, [16] proposed to employ Persistent CD and a replay buffer to store intermediate samples from Markov chains throughout training, and [63] proposed to learn non-convergent short-run MCMC. Both approaches (long-run and short-run MCMC) work well with PS-CD in our experiments. PS-CD may also benefit from recent advances on unbiased MCMC [39, 70], which we leave as interesting future work.

Fenchel Duality. By exploiting the primal-dual view of KL, recent works [10, 11, 2] proposed to cast maximum likelihood training of EBMs as minimax problems, which introduce a dual sampler

and are approximately solved by alternating gradient descent ascent updates. Along this line, to allow flexible modeling preferences, [86] proposed f-EBM to enable the use of any f-divergence to train EBMs, which also relies on minimax optimization. By contrast, in this work, we leverage the analytical form of the optimal variational distribution and self-normalized importance sampling to reach a framework that requires no adversarial training and has no additional computational cost compared to CD while allowing flexible modeling preferences. Besides convenient optimization, PS-CD and f-EBM trains EBMs with two different families of divergences (hence complementary) with KL being the only shared one, since any pseudo-spherical scoring rule corresponds to a Bregman divergence (Section 2.3) and the only member in f-divergence that is also Bregman divergence is α-divergence (with KL as special case) (Theorem 4 in [1]).

Homogeneous Scoring Rules. [81] proposed to learn unnormalized statistical models on discrete sample space by maximizing γ-score, which uses empirical data distribution ($\widehat{p}(\boldsymbol{x}) = n_{\boldsymbol{x}}/n$, where $n_{\boldsymbol{x}}$ is the number of appearance of $\boldsymbol{x}$ in the dataset and n is the total number of data) as a surrogate to the real data density p and relies on a localization trick to bypass the computation of $\|q_{\boldsymbol{\theta}}\|_{\gamma+1}$. Consequently, it is only amenable to finite discrete sample space such as natural language [49], whereas PS-CD is applicable to any unnormalized probabilistic model in continuous domains. Another popular homogeneous scoring rule is the Hyvärinen score, which gives rise to the score matching objective [38] for EBM training. However, score matching and its variants [84, 79] have difficulties in low data density regions and do not perform well in practice when training EBMs on high-dimensional datasets [77]. Moreover, since the score matching objective involves the Hessian of log-density functions that is generally expensive to compute [56], methods such as approximate propagation [43], curvature propagation [56] and sliced score matching [80] are needed to approximately compute the trace of the Hessian.

Noise Contrastive Estimation. Another principle for learning EBMs is Noise Contrastive Estimation (NCE) [33], where an EBM is learned by contrasting a prescribed noise distribution with tractable density against the unknown data distribution. Using various Bregman divergences, NCE can be generalized to a family of different loss functionals [32, 82]. However, finding an appropriate noise distribution for NCE is highly non-trivial. In practice, NCE typically works well in conjunction with a carefully-designed noise distribution such as context-dependent noise distribution [40] or joint learning of a flow-based noise distribution [20].

In this work, we focus on generalizing maximum likelihood by deriving novel training objectives for EBMs without involving auxiliary models (*e.g.*, the variational function in [86], the flow-based noise distribution in [20] and the amortized sampler in [48, 11, 28]).

6 Experiments

In this section, we demonstrate the effectiveness of PS-CD on several 1-D and 2-D synthetic datasets as well as commonly used image datasets.

Setup. The 2-D synthetic datasets include Cosine, Swiss Roll, Moon, Mixture of Gaussian, Funnel and Rings, which cover different modalities and geometries (see Figure 2 in App. D.1 for illustration). To test the practical usefulness, we use MNIST [52], CIFAR-10 [46] and CelebA [55] in our experiments for modeling natural images. Following [77], for CelebA, we first center-crop the images to 140×140, then resize them to 64×64. More experimental details about the data processing, model architectures, sampling strategies and additional experimental results can be found in App. D.

Effects of Different γ Values. To illustrate the modeling flexibility brought by PS-CD and provide insights on the effects of different γ values, we first conduct a 1-D synthetic experiment similar to the one in [86]. As shown in Figure 1, when fitting a mixture of Gaussian with a single Gaussian (*i.e.*, model mis-specification case), the family of PS-CD offers flexible tradeoffs between quality and diversity (*i.e.*, mode collapse vs. mode coverage). Although in the well-specified case these objectives induce the same optimal solution, in this example, we can see that a larger γ leads to higher entropy. More importantly, compared to f-EBM [86] that also provides similar modeling flexibility and includes CD as a special case, PS-CD

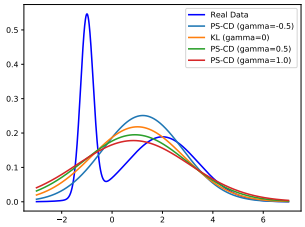

Figure 1: The effects of different γ values when fitting a mixture of Gaussian with a single Gaussian.

does not require expensive and unstable minimax optimization (no additional cost compared to CD). In App. D.2, we further visualize the objective landscapes for different γ values. As shown in Figure 3 and 4, when the model is well-specified, different objectives will induce the same optimal solution since they are strictly proper; when the model is mis-specified (corresponding to practical scenarios), different objectives will exhibit different modeling preferences.

Furthermore, to better demonstrate the modeling flexibility brought by PS-CD in high-dimensional case, we conduct experiments on CIFAR-10 using a set of indicative and reliable metrics (Density, Coverage, Precision, Recall) proposed by [59] to evaluate the effects of γ from various perspectives. Please refer to App. D.3 for experimental results (Table 3) and detailed discussions.

2-D Synthetic Data. For quantitative evaluation of the 2-D synthetic data experiments, we follow [76] and report the maximum mean discrepancy (MMD, [5]) between the generated samples and validation samples in Table 2 in App. D.1, which demonstrates that PS-CD outperforms its CD counterpart on all but the Funnel dataset. From the histograms of samples shown in Figure 2 in App. D.1, we can also have similar observations. For example, CD fails to place high densities in the center of the right mode in MoG, while PS-CD places the modes correctly.

Image Generation. In Figure 5 in App.D.4, we show MNIST, CIFAR-10 and CelebA samples produced by PS-CD (with $\gamma = 1.0$), which demonstrate that our approach can produce highly realistic images with simple model architectures. As suggested in [16], we use Fréchet Inception Distance (FID) [35] as the quantitative evaluation metric for CIFAR-10 and CelebA, as Langevin dynamics converge to local minima that artificially inflate Inception Score [74]. From Table 1[1], we can see that various members (different γ values) in the family of PS-CD can outperform CD significantly, and more surprisingly, PS-CD also shows competitive performance to the recently proposed f-EBMs, without requiring expensive minimax optimization. While our method currently does not outperform the state-of-the-art image generation methods such as improved denoising score matching [78], which relies on carefully selected noise schedule and specially designed noise-dependent score network (modified U-Net architecture, hence not directly comparable to our results), we think that our work opens up a new research direction by bridging statistical decision theory (homogeneous proper scoring rules) and deep energy-based generative modeling. Moreover, under the setting of simple model architectures and the same hyper-parameter configuration (*e.g.*, batch size, learning rate, network structure, etc.), our empirical results suggest clear superiority of PS-CD over traditional CD and recent f-EBMs.

Table 1: FID scores for CIFAR-10 conditional and CelebA unconditional image generation. We list comparisons with results reported by CD [16], Noise-Conditional Score Network (NCSN) [78] and f-EBMs [86]. $\gamma = 1.0$ corresponds to maximizing spherical scoring rule (Example 2).

Method	FID
CIFAR-10 (32×32) Conditional	
Contrastive Divergence (KL)	37.90
f-EBM (KL)	37.36
f-EBM (Reverse KL)	33.25
f-EBM (Squared Hellinger)	32.19
f-EBM (Jensen Shannon)	30.86
Pseudo-Spherical CD ($\gamma = 2.0$)	33.19
Pseudo-Spherical CD ($\gamma = 1.0$)	**29.78**
Pseudo-Spherical CD ($\gamma = 0.5$)	35.02
Pseudo-Spherical CD ($\gamma = -0.5$)	**27.95**
CelebA (64×64)	
Contrastive Divergence (KL)	26.10
NCSN (w/o denoising)	26.89
NCSN (w/ denoising)	25.30
NCSNv2 (w/o denoising)	28.86
NCSNv2 (w/ denoising)	**10.23**
Pseudo-Spherical CD ($\gamma = 1.0$)	**24.76**
Pseudo-Spherical CD ($\gamma = -0.5$)	**20.35**

7 Conclusion

From the perspective of maximizing strictly proper homogeneous scoring rules, we propose pseudo-spherical contrastive divergence (PS-CD) to generalize maximum likelihood estimation of energy-based models. Different from prior works that involve joint training of auxiliary models [86, 20, 48, 11, 28], PS-CD allows us to specify flexible modeling preferences without additional computational cost compared to contrastive divergence. We provide a theoretical analysis on the sample complexity and convergence property of the proposed method, as well as its connection to maximum likelihood. Finally, we demonstrate the effectiveness of PS-CD with extensive experiments on both synthetic data and commonly used image datasets.

[1]For CelebA dataset, we reproduced the short-run MCMC method [63] using our code base. Moreover, the f-EBM paper only reported results on CelebA 32x32 and we empirically found it is not comparable to our method in CelebA 64x64 (higher resolution), indicating better scalability of PS-CD to high-dimensional case.

References

[1] Shun-ichi Amari. Divergence function, information monotonicity and information geometry. In *Workshop on information theoretic methods in science and engineering (WITMSE)*. Citeseer, 2009.

[2] Michael Arbel, Liang Zhou, and Arthur Gretton. Kale: When energy-based learning meets adversarial training. *arXiv preprint arXiv:2003.05033*, 2020.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[5] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[6] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[7] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[8] Jie Chen and Ronny Luss. Stochastic gradient descent with biased but consistent gradient estimators. *arXiv preprint arXiv:1807.11880*, 2018.

[9] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

[10] Bo Dai, Hanjun Dai, Arthur Gretton, Le Song, Dale Schuurmans, and Niao He. Kernel exponential family estimation via doubly dual embedding. *arXiv preprint arXiv:1811.02228*, 2018.

[11] Bo Dai, Zhen Liu, Hanjun Dai, Niao He, Arthur Gretton, Le Song, and Dale Schuurmans. Exponential family estimation via adversarial dynamics embedding. *arXiv preprint arXiv:1904.12083*, 2019.

[12] A Philip Dawid. Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. *Department of Statistical Science, University College London. http://www. ucl. ac. uk/Stats/research/abs94. html, Tech. Rep*, 139, 1998.

[13] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[15] Yilun Du, Toru Lin, and Igor Mordatch. Model based planning with energy based models. *arXiv preprint arXiv:1909.06878*, 2019.

[16] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems 32*, pages 3603–3613, 2019.

[17] Axel Finke and Alexandre H Thiery. On importance-weighted autoencoders. *arXiv preprint arXiv:1907.10477*, 2019.

[18] Daniel Friedman. Effective scoring rules for probabilistic forecasts. *Management Science*, 29(4):447–454, 1983.

[19] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

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

[21] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.

[22] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015.

[23] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[24] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[25] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[26] IJ Good. Comment on "measuring information and uncertainty" by robert j. buehler. *Foundations of Statistical Inference*, pages 337–339, 1971.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[28] Will Grathwohl, Jacob Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. No mcmc for me: Amortized sampling for fast and stable training of energy-based models. *arXiv preprint arXiv:2010.04230*, 2020.

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

[30] Ulf Grenander, Michael I Miller, Michael Miller, et al. *Pattern theory: from representation to inference*. Oxford university press, 2007.

[31] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[32] Michael U Gutmann and Jun-ichiro Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 283–290, 2011.

[33] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The journal of machine learning research*, 13(1):307–361, 2012.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

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

[37] Tito Homem-de Mello. On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling. *SIAM Journal on Optimization*, 19(2):524–551, 2008.

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

[39] Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased markov chain monte carlo with couplings. *arXiv preprint arXiv:1708.03625*, 2017.

[40] Shihao Ji, SVN Vishwanathan, Nadathur Satish, Michael J Anderson, and Pradeep Dubey. Blackout: Speeding up recurrent neural network language models with very large vocabularies. *arXiv preprint arXiv:1511.06909*, 2015.

[41] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

[42] Takafumi Kanamori, Hironori Fujisawa, et al. Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli*, 20(4):2278–2304, 2014.

[43] Durk P Kingma and Yann L Cun. Regularized estimation of image statistics by score matching. In *Advances in neural information processing systems*, pages 1126–1134, 2010.

[44] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[45] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1895–1904. JMLR. org, 2017.

[46] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[48] Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.

[49] Matthieu Labeau and Shay B Cohen. Experimenting with power divergences for language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4095–4105, 2019.

[50] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[51] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37, 2011.

[52] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[54] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[55] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[56] James Martens, Ilya Sutskever, and Kevin Swersky. Estimating the hessian by back-propagating curvature. *arXiv preprint arXiv:1206.6464*, 2012.

[57] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

[58] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[59] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020.

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

[61] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[62] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[63] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. In *Advances in Neural Information Processing Systems*, pages 5233–5243, 2019.

[64] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[65] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[66] Matthew Parry, A Philip Dawid, Steffen Lauritzen, et al. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

[67] Matthew Parry et al. Linear scoring rules for probabilistic binary classification. *Electronic Journal of Statistics*, 10(1):1596–1607, 2016.

[68] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[69] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE, 2011.

[70] Yixuan Qiu, Lingsong Zhang, and Xiao Wang. Unbiased contrastive divergence algorithm for training energy-based latent variable models.

[71] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.

[72] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[73] Thornton B Roby. Belief states and the uses of evidence. *Behavioral Science*, 10(3):255–270, 1965.

[74] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[75] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[76] Jiaming Song and Stefano Ermon. Bridging the gap between f-gans and wasserstein gans. *arXiv preprint arXiv:1910.09779*, 2019.

[77] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019.

[78] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, 2020.

[79] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, page 204, 2019.

[80] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020.

[81] Takashi Takenouchi and Takafumi Kanamori. Empirical localization of homogeneous divergences on discrete sample spaces. In *Advances in Neural Information Processing Systems*, pages 820–828, 2015.

[82] Masatoshi Uehara, Takafumi Kanamori, Takashi Takenouchi, and Takeru Matsuda. A unified statistically efficient estimation framework for unnormalized models. In *International Conference on Artificial Intelligence and Statistics*, pages 809–819, 2020.

[83] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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

[85] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[86] Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. In *International Conference on Machine Learning*, pages 10957–10967. PMLR, 2020.

[87] Jun Zhang, Yaokun Lei, Yi Isaac Yang, and Yi Qin Gao. Deep learning for multi-scale molecular modeling. 2020.

[88] Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems*, pages 10792–10801, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] Due to space limit, we describe potential societal impacts here: In this paper, we focus on the theoretical aspects of unnormalized statistical models and propose a new algorithm for training deep energy-based models, which is a powerful class of generative models with numerous potential applications including model-based planning for robotics

and autonomous driving, molecular and chemical modeling, and realistic images, videos and text generation. However, powerful generative models may also have potential negative impacts. For example, researchers have shown that they can be used to create adversarial examples for attacking deployed machine learning systems or manipulating reinforcement learning agents. That being said, researchers also show that deep generative models can be employed to improve the robustness of machine learning systems and RL agents via adversarial training. Overall, it is unclear whether the net impact will be positive or negative. Moreover, like other statistical models that discover patterns from data, it treats the given dataset as independent and identically distributed samples from the desired distribution. When the dataset contains undesired bias, it will also be captured by the learned model. Thus, we should carefully control the data collection process to take account of critical aspects such as fairness.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All the assumptions, lemmas and theorems are clearly presented in the main paper and the Appendices.
   (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs for all theoretical results can be found in Appendix B and C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 6, Appendix A and D.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6 and Appendix D.5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For low-dimensional synthetic experiments, see Table 2. For large-scale image generation experiments, we follow previous works, which typically report single numbers due to expensive cost of running experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]