On Energy-Based Models with Overparametrized Shallow Neural Networks

Carles Domingo-Enrich [1] **Alberto Bietti** [2] **Eric Vanden-Eijnden** [1] **Joan Bruna** [1 2]

Abstract

Energy-based models (EBMs) are a simple yet powerful framework for generative modeling. They are based on a trainable energy function which defines an associated Gibbs measure, and they can be trained and sampled from via well-established statistical tools, such as MCMC. Neural networks may be used as energy function approximators, providing both a rich class of expressive models as well as a flexible device to incorporate data structure. In this work we focus on shallow neural networks. Building from the incipient theory of overparametrized neural networks, we show that models trained in the so-called 'active' regime provide a statistical advantage over their associated 'lazy' or kernel regime, leading to improved adaptivity to hidden low-dimensional structure in the data distribution, as already observed in supervised learning. Our study covers both maximum likelihood and Stein Discrepancy estimators, and we validate our theoretical results with numerical experiments on synthetic data.

1. Introduction

A central problem in machine learning is to learn generative models of a distribution through its samples. Such models may be needed simply as a modeling tool in order to discover properties of the data, or as a way to generate new samples that are similar to the training samples. Generative models come in various flavors. In some cases very few assumptions are made on the distribution and one simply tries to learn generator models in a black-box fashion (Goodfellow et al., 2014; Kingma & Welling, 2013), while other approaches make more precise assumptions on the form of the data distribution. In this paper, we focus on the latter approach, by considering Gibbs measures defined through an *energy function* f, with a density proportional to $\exp\{-f(x)\}$. Such *energy-based models* (EBMs) originate in statistical physics (Ruelle, 1969), and have become a fundamental modeling tool in statistics and machine learning (Wainwright & Jordan, 2008; Ranzato et al., 2007; LeCun et al., 2006; Xie et al., 2016; 2017; Du & Mordatch, 2019; Song & Kingma, 2021). If data is assumed to come from such a model, the learning algorithms then attempt to estimate the energy function f. The resulting learned model can then be used to obtain new samples, typically through Markov Chain Monte Carlo (MCMC) techniques.

In this paper, we study the statistical problem of learning such EBMs from data, in a non-parametric setting defined by a function class $\mathcal{F}$, and with possibly arbitrary target energy functions. If we only assume a simple Lipschitz property on the energy, learning such models will generally suffer from the curse of dimensionality (von Luxburg & Bousquet, 2004), in the sense that an exponential number of samples in the dimension is needed to find a good model. However, one may hope to achieve better guarantees when additional structure is present in the energy function.

An important source of structure comes from energy functions which capture local rather than global interactions between input features, such as those in Local Markov Random Fields or Ising models. Such energies can be expressed as linear combinations of potential functions depending only on low-dimensional projections, and are therefore amenable to efficient approximation by considering classes $\mathcal{F}$ given by shallow neural networks endowed with a sparsity-promoting norm (Bach, 2017a). Analogously to the supervised regime (Bach, 2017a; Chizat & Bach, 2020), learning in such *variation-norm* spaces $\mathcal{F} = \mathcal{F}_1$ admits a convex formulation in the overparametrized limit, whose corresponding class of Gibbs measures $\{\nu(dx) \propto \exp\{-f(dx)\},\ f \in \mathcal{F}_1\}$ is the natural infinite-dimensional extension of exponential families (Wainwright & Jordan, 2008). Our main contribution is to show that such EBMs lead to a well-posed learning setup with strong statistical guarantees, breaking the curse of dimensionality.

These statistical guarantees can be combined with qualitative optimization guarantees in this overparamerised limit under an appropriate 'active' or 'mean-field' scaling (Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2019). As it is also

[1] Courant Institute of Mathematical Sciences, New York University [2] Center for Data Science, New York University. Correspondence to: Carles Domingo-Enrich <cd2754@nyu.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

the case for supervised learning, the benefits of variation-norm spaces $\mathcal{F}_1$ contrast with their RKHS counterparts $\mathcal{F}_2$, which cannot efficiently adapt to the low-dimensional structure present in such structured Gibbs models.

The standard method to train EBMs is maximum likelihood estimation. One generic approach for this is to use gradient descent, where gradients may be approximated using MCMC samples from the current trained model. Such sampling procedures may be difficult in general, particularly for complex energy landscapes, thus we also consider different estimators based on un-normalized measures which avoid the need of sampling. We focus here on approaches based on minimizing Stein discrepancies (Gorham & Mackey, 2015; Liu & Wang, 2016), which have recently been found to be useful in deep generative models (Grathwohl et al., 2020), though we note that alternative approaches may be used, such as score matching (Hyvärinen, 2005; Song & Kingma, 2021; Song & Ermon, 2019; Block et al., 2020).

Our main focus is to study the resulting estimators when using gradient-based optimization over infinitely-wide neural networks in different regimes, showing the statistical benefits of the 'feature learning' regime when the target models have low-dimensional structure, thus extending the analogous results for supervised least-squares (Bach, 2017a) and logistic (Chizat & Bach, 2020) regression. More precisely, we make the following contributions:

- We derive generalization bounds for the learned measures in terms of the same metrics used for training (KL divergence or Stein discrepancies). Using and extending results from the theory of overparametrized neural networks, we show that when using energies in the class $\mathcal{F}_1$ we can learn target measures with certain low-dimensional structure at a rate controlled by the intrinsic dimension rather than the ambient dimension (Corollary 1 and Corollary 2).
- We show in experiments that while $\mathcal{F}_1$ energies succeed in learning simple synthetic distributions with low-dimensional structure, $\mathcal{F}_2$ energies fail (Sec. 6).

2. Related work

A recent line of research has studied the question of how neural networks compare to kernel methods, with a focus on supervised learning problems. Bach (2017a) studies two function classes that arise from infinite-width neural networks with different norms penalties on its weights, leading to the two different spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, and shows the approximation benefits of the $\mathcal{F}_1$ space for adapting to low-dimensional structures compared to the (kernel) space $\mathcal{F}_2$, an analysis that we leverage in our work. The function space $\mathcal{F}_1$ was also studied by Ongie et al. (2019); Savarese et al. (2019); Williams et al. (2019) by focusing on the ReLU activation function. More recently, this question has gained interest after several works have shown that wide neural networks trained with gradient methods may behave like kernel methods in certain regimes (see, e.g., Jacot et al., 2018). Examples of works that compare 'active/feature learning' and 'kernel/lazy' regimes include (Chizat & Bach, 2020; Ghorbani et al., 2019; Wei et al., 2020; Woodworth et al., 2020). We are not aware of any works that study questions related to this in the context of generative models in general and EBMs in particular.

Other related work includes the Stein discrepancy literature. Although Stein's method (Stein, 1972) dates to the 1970s, it has been popular in machine learning in recent years. Gorham & Mackey (2015) introduced a computational approach to compute the Stein discrepancy in order to assess sample quality. Later, Chwialkowski et al. (2016) and Liu et al. (2016) introduced the more practical kernelized Stein discrepancy (KSD) for goodness-of-fit tests, which were also studied by Gorham & Mackey (2017). Liu & Wang (2016) introduced SVGD, which was the first method to use the KSD to obtain samples from a distribution, and Barp et al. (2019) where the first to employ KSD to train parametric generative models. More recently, Grathwohl et al. (2020) used neural networks as test functions for Stein discrepancies, which arguably yields a stronger metric, and have shown how to leverage such metrics for training EBMs. The empirical success of their method provides an additional motivation for our theoretical study of the $\mathcal{F}_1$ Stein Discrepancy (Subsec. 4.2).

Finally, another notable paper close in spirit to our goal is (Block et al., 2020), which provides a detailed theoretical analysis of a score-matching generative model using Denoising Autoencoders followed by Langevin diffusion. While their work makes generally weaker assumptions and also includes a non-asymptotic analysis of the sampling algorithm, the resulting rates are unsuprisingly cursed by dimension. Our focus is on the statistical aspects which allow faster rates, leaving the quantitative computational aspects aside.

3. Setting

In this section, we present the setup of our work, recalling basic properties of EBMs, maximum likelihood estimators, Stein discrepancies, and functional spaces arising from infinite-width shallow neural networks.

Notation. If V is a normed vector space, we use $\mathcal{B}_V(\beta)$ to denote the closed ball of V of radius β, and $\mathcal{B}_V := \mathcal{B}_V(1)$ for the unit ball. If K denotes a subset of the Euclidean space, $\mathcal{P}(K)$ is the set of Borel probability measures, $\mathcal{M}(K)$ is the space of signed Radon measures and $\mathcal{M}^+(K)$ is the space of (non-negative) Radon measures. For $\nu_1, \nu_2 \in \mathcal{P}(K)$, we define the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(\nu_1 || \nu_2) := \int_K \log(\frac{d\nu_1}{d\nu_2}(x)) d\nu_1(x)$ when ν_1 is absolutely continuous with respect to ν_2, and $+\infty$ otherwise, and the cross-entropy $H(\nu_1, \nu_2) := -\int_K \log(\frac{d\nu_2}{d\tau}(x)) d\nu_1(x)$, where $\frac{d\nu_2}{d\tau}(x)$ is the Radon-Nikodym derivative w.r.t. the uniform probability measure τ of K, and the differential entropy $H(\nu_1) := -\int_K \log(\frac{d\nu_1}{d\tau}(x)) d\nu_1(x)$. If γ is a signed measure over K, then $|\gamma|_{\text{TV}}$ is the total variation (TV) norm of γ. $\mathbb{S}^d$ is the d-dimensional hypersphere, and for functions $f : \mathbb{S}^d \to \mathbb{R}$, ∇f denotes the Riemannian gradient of f. We use $\sigma(\langle \theta, x \rangle) = \max\{0, \langle \theta, x \rangle\}$ to denote a ReLU with parameter θ.

3.1. Generative energy-based models

If $\mathcal{F}$ is a class of functions (or energies) mapping a measurable set $K \subseteq \mathbb{R}^{d+1}$ to $\mathbb{R}$, for any $f \in \mathcal{F}$ we can define the probability measure ν_f as a Gibbs measure with density:

$$\frac{d\nu_f}{d\tau}(x) := \frac{e^{-f(x)}}{Z_f}, \text{ with } Z_f := \int_K e^{-f(y)} d\tau(y) \ ,$$

where $\frac{d\nu_f}{d\tau}(x)$ is the Radon-Nikodym derivative w.r.t to the uniform probability measure over K, denoted τ, and Z_f is the partition function.

Given samples $\{x_i\}_{i=1}^n$ from a target measure ν, training an EBM consists in selecting the best ν_f with energy $f \in \mathcal{F}$ according to a given criterion. A natural estimator $\hat{f}$ for the energy is the **maximum likelihood** estimator (MLE), i.e., $\hat{f} = \text{argmax}_{f \in \mathcal{F}} \prod_{i=1}^n \frac{d\nu_f}{d\tau}(x_i)$, or equivalently, the one that minimizes the cross-entropy with the samples:

$$\begin{aligned}\hat{f} &= \underset{f \in \mathcal{F}}{\text{argmin}}\, H(\nu_n, \nu_f) = \underset{f \in \mathcal{F}}{\text{argmin}} -\frac{1}{n} \sum_{i=1}^n \log\left(\frac{d\nu_f}{d\tau}(x_i)\right) \\ &= \underset{f \in \mathcal{F}}{\text{argmin}}\, \frac{1}{n} \sum_{i=1}^n f(x_i) + \log Z_f.\end{aligned} \tag{1}$$

The estimated distribution is simply $\nu_{\hat{f}}$, and samples can be obtained by the MCMC algorithm of choice.

An alternative estimator is the one that arises from minimizing the **Stein discrepancy** (SD) corresponding to a function class $\mathcal{H}$. If $\mathcal{H}$ is a class of functions from K to $\mathbb{R}^{d+1}$, the Stein discrepancy (Gorham & Mackey, 2015; Liu et al., 2016) for $\mathcal{H}$ is a non-symmetric functional defined on pairs of probability measures over K as

$$\text{SD}_{\mathcal{H}}(\nu_1, \nu_2) = \sup_{h \in \mathcal{H}} \mathbb{E}_{\nu_1}[\text{Tr}(\mathcal{A}_{\nu_2} h(x))], \tag{2}$$

where $\mathcal{A}_\nu : K \to \mathbb{R}^{(d+1)\times(d+1)}$ is the Stein operator. In order to leverage approximation properties on the sphere, we will consider functions h defined on $K = \mathbb{S}^d$. In this case, the Stein operator is defined by $\mathcal{A}_\nu h(x) := (s_\nu(x) - d \cdot x)h(x)^\top + \nabla h(x)$ (see Lemma 5), where $s_\nu(x) = \nabla \log(\frac{d\nu}{d\tau}(x))$ is named the score function. The term $d \cdot x$ is important for the spherical case in order to have $\text{SD}_{\mathcal{H}}(\nu, \nu) = 0$, while it does not appear when considering $K = \mathbb{R}^d$. The Stein discrepancy estimator is

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \text{SD}_{\mathcal{H}}(\nu_n, \nu_f). \tag{3}$$

If $\mathcal{H} = \mathcal{B}_{\mathcal{H}_0^{d+1}} = \{(h_i)_{i=1}^{d+1} \in \mathcal{H}_0^{d+1} \mid \sum_{i=1}^{d+1} \|h_i\|_{\mathcal{H}_0}^2 \leq 1\}$ for some reproducing kernel Hilbert space (RKHS) $\mathcal{H}_0$ with kernel k with continuous second order partial derivatives, there exists a closed form for the problem (2) and the corresponding object is known as **kernelized Stein discrepancy** (KSD) (Liu et al., 2016; Gorham & Mackey, 2017). For $K = \mathbb{S}^d$, the KSD takes the following form (Lemma 6):

$$\text{KSD}(\nu_1, \nu_2) = \text{SD}^2_{\mathcal{B}_{\mathcal{H}_0^{d+1}}}(\nu_1, \nu_2) = \mathbb{E}_{x,x' \sim \nu_1}[u_{\nu_2}(x, x')], \tag{4}$$

where $u_\nu(x, x') = (s_\nu(x) - d \cdot x)^\top (s_\nu(x') - d \cdot x')k(x, x') + (s_\nu(x) - d \cdot x)^\top \nabla_{x'} k(x, x') + (s_\nu(x') - d \cdot x')^\top \nabla_x k(x, x') + \text{Tr}(\nabla_{x,x'} k(x, x'))$, and we use $\tilde{u}_\nu(x, x')$ to denote the sum of the first three terms (remark that the fourth term does not depend on ν). One KSD estimator that can be used is

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \frac{1}{n^2} \sum_{i,j=1}^n \tilde{u}_{\nu_f}(x_i, x_j). \tag{5}$$

The optimization problem for this estimator is convex (Sec. 5), but it is biased. On the other hand, the estimator

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \frac{1}{n(n-1)} \sum_{i \neq j} \tilde{u}_{\nu_f}(x_i, x_j), \tag{6}$$

is unbiased, but the optimization problem is not convex.

3.2. Neural network energy classes

We are interested in the cases in which $\mathcal{F}$ is one of two classes of functions related to shallow neural networks, as studied by Bach (2017a).

Feature learning regime. $\mathcal{F}$ is the ball $\mathcal{B}_{\mathcal{F}_1}(\beta)$ of radius $\beta > 0$ of $\mathcal{F}_1$, which is the Banach space of functions $f : K \to \mathbb{R}$ such that for all $x \in K$ we have $f(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle)\, d\gamma(\theta)$, for some signed Radon measure $\gamma \in \mathcal{M}(\mathbb{S}^d)$. The norm of $\mathcal{F}_1$ is defined as $\|f\|_{\mathcal{F}_1} = \inf \left\{ |\gamma|_{\text{TV}} \mid f(\cdot) = \int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle)\, d\gamma(\theta) \right\}$.

Kernel regime. $\mathcal{F}$ is the ball $\mathcal{B}_{\mathcal{F}_2}(\beta)$ of radius $\beta > 0$ of $\mathcal{F}_2$, which is the (reproducing kernel) Hilbert space of functions $f : K \to \mathbb{R}$ such that for some absolutely continuous $\rho \in \mathcal{M}(\mathbb{S}^d)$ with $\frac{d\rho}{d\tilde{\tau}} \in \mathcal{L}^2(\mathbb{S}^d)$ (where $\tilde{\tau}$ the uniform probability measure over $\mathbb{S}^d$), we have that for all $x \in K$, $f(x) =$

$\int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle)\, d\rho(\theta)$. The norm of $\mathcal{F}_2$ is defined as $\|f\|_{\mathcal{F}_2}^2 = \inf \left\{ \int_{\mathbb{S}^d} |h(\theta)|^2 \, d\tilde{\tau}(\theta) \mid f(\cdot) = \int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle) h(\theta)\, d\tilde{\tau}(\theta) \right\}$. As an RKHS, the kernel of $\mathcal{F}_2$ is $k(x,y) = \int_{\mathbb{S}^d} \sigma(\langle x, \theta\rangle)\sigma(\langle y, \theta \rangle)\, d\tilde{\tau}(\theta)$.

Remark that since $\int |h(\theta)| d\tilde{\tau}(\theta) \leq (\int |h(\theta)|^2 \, d\tilde{\tau}(\theta))^{1/2}$ by the Cauchy-Schwarz inequality, we have $\mathcal{F}_2 \subset \mathcal{F}_1$ and $\mathcal{B}_{\mathcal{F}_2} \subset \mathcal{B}_{\mathcal{F}_1}$. The TV norm in $\mathcal{F}_1$ acts as a sparsity-promoting penalty, which encourages the selection of few well-chosen neurons and may lead to favorable adaptivity properties when the target has a low-dimensional structure. In particular, (Bach, 2017a) shows that single ReLU units belong to $\mathcal{F}_1$ but not to $\mathcal{F}_2$, and their L^2 approximations in $\mathcal{F}_2$ have exponentially high norm in the dimension. Ever since, several works have further studied the gaps arising between such nonlinear and linear regimes (Wei et al., 2019; Ghorbani et al., 2020; Malach et al., 2021). In App. D, we present dual characterizations of the maximum likelihood $\mathcal{F}_1$ and $\mathcal{F}_2$ EBMs as entropy maximizers under L^∞ and L^2 moment constraints (an infinite-dimensional analogue of Della Pietra et al. (1997); see also Mohri et al. (2012), Theorem 12.2).

The ball radius β acts as an inverse temperature. The low temperature regime $\beta \gg 1$ corresponds to expressive models with lower approximation error but higher statistical error: the theorems in Sec. 4 provide bounds on the two errors and the results of optimizing such bounds w.r.t. β. In the following, we will assume that the set $K \subset \mathbb{R}^{d+1}$ is compact. We note that there are two interesting choices for K: (i) for $K = \mathbb{S}^d$, we obtain neural networks without bias term; and (ii) for $K = K_0 \times \{R\}$, where $K_0 \subset \mathbb{R}^d$ with norm bounded by R, we obtain neural networks on K_0 with a bias term.

4. Statistical guarantees for shallow neural network EBMs

In this section, we present our statistical generalization bounds for various EBM estimators based on maximum likelihood and Stein discrepancies, highligting the adaptivity to low-dimensional structures that can be achieved when learning with energies in $\mathcal{F}_1$. All the proofs are in App. A.

4.1. Guarantees for maximum likelihood EBMs

The following theorem provides a bound of the KL divergence between the target probability measure and the maximum likelihood estimator in terms of a statistical error and an approximation error.

Theorem 1. *Assume that the class $\mathcal{F}$ has a (distribution-free) Rademacher complexity bound $\mathcal{R}_n(\mathcal{F}) \leq \frac{\beta C}{\sqrt{n}}$ and L^∞ norm uniformly bounded by β. Given n samples $\{x_i\}_{i=1}^n$ from the target measure ν with support in K, consider the maximum likelihood estimator (MLE) $\hat{\nu} := \nu_{\hat{f}}$, where $\hat{f}$ is the estimator defined in (1). With probability at least $1-\delta$, we have*

$$D_{KL}(\nu||\hat{\nu}) \leq \frac{4\beta C}{\sqrt{n}} + \beta\sqrt{\frac{8\log(1/\delta)}{n}} + \inf_{f \in \mathcal{F}} D_{KL}(\nu||\nu_f). \tag{7}$$

If $\frac{d\nu}{d\tau}(x) = e^{-g(x)}/\int_K e^{-g(y)} d\tau(y)$ for some $g : K \to \mathbb{R}$, i.e. $-g$ is the log-density of ν up to a constant term, then with probability at least $1-\delta$,

$$D_{KL}(\nu||\hat{\nu}) \leq \frac{4\beta C}{\sqrt{n}} + \beta\sqrt{\frac{8\log(1/\delta)}{n}} + 2 \inf_{f \in \mathcal{F}} \|g - f\|_\infty. \tag{8}$$

Equation (7) follows from using a classical argument in statistical learning theory. To obtain equation (8) we bound the last term of (7) by $2\inf_{f\in\mathcal{F}} \|g-f\|_\infty$ using Lemma 1 in App. A. We note that other metrics than L_∞ may be used for the approximation error, such as the Fisher divergence, but these will likely lead to similar guarantees under our assumptions. Making use of the bounds developed in (Bach, 2017a), Corollary 1 below applies (8) to the case in which $\mathcal{F}$ is the $\mathcal{F}_1$ ball $\mathcal{B}_{\mathcal{F}_1}(\beta)$ for some $\beta > 0$ and the energy of the target distribution is a sum of Lipschitz functions of orthogonal projection to low-dimensional subspaces.

Assumption 1. *Let $K = K_0 \times \{R\}$, where $K_0 \subseteq \{x \in \mathbb{R}^d | \|x\|_2 \leq R\}$ is compact. Suppose that the target probability measure ν is absolutely continuous w.r.t. the Borel measure over K and it satisfies $\forall x \in K_0$, $\frac{d\nu}{d\tau}(x, R) = \exp(-\sum_{j=1}^J \varphi_j(U_j x))/\int_{K_0} \exp(-\sum_{j=1}^J \varphi_j(U_j y)) d\tau$, where φ_j are (ηR^{-1})-Lipschitz continuous functions on the R-ball of $\mathbb{R}^k$ such that $\|\varphi_j\|_\infty \leq \eta$, and $U_j \in \mathbb{R}^{k\times d}$ with orthonormal rows.*

Corollary 1. *Let $\mathcal{F} = B_{\mathcal{F}_1}(\beta)$. Assume that Assumption 1 holds. Then, we can choose $\beta > 0$ such that with probability at least $1-\delta$ we have*

$$D_{KL}(\nu||\hat{\nu}) \leq \tilde{O}\left(\left(1 + \sqrt{\log(1/\delta)}\right) J\eta R^{-\frac{2}{k+3}} n^{-\frac{1}{k+3}}\right)$$

where the notation $\tilde{O}$ indicates that we overlook logarithmic factors and constants depending only on the dimension k.

Remarkably, Corollary 1 shows that for our class of target measures with low-dimensional structure, the KL divergence between ν and $\hat{\nu}$ decreases as $n^{-\frac{1}{k+3}}$. That is, the rate "breaks" the curse of dimensionality since the exponent only depends on the dimension k of the low-dimensional spaces, not to the ambient dimension d. This can be seen as an alternative, more structural approach to alleviate dimension-dependence compared to other standard assumptions such smoothness classes for density estimation (e.g., Singh et al., 2018; Tsybakov, 2008). As discussed earlier, a motivation for Assumption 1 comes from Markov Random Fields, where each φ_j corresponds to a local potential defined on a neighborhood determined by U_j. Note that the bound scales linearly with respect to the number of local potentials J. As our experiments illustrate (see Sec. 6), it is easy to construct target energies that are much better approximated in $\mathcal{F}_1$ than in $\mathcal{F}_2$. Indeed, we find that the test error tends to decrease more quickly as a function of the sample size when training both layers of shallow networks rather than just the second layer, which corresponds to controlling the $\mathcal{F}_1$ norm.

4.2. Guarantees for Stein Discrepancy EBMs

We now consider EBM estimators obtained by minimizing Stein discrepancies, and establish bounds on the Stein discrepancies between the target measure and the estimated one. As in Subsec. 4.1, we begin by providing error decompositions in terms of estimation and approximation error. The following theorem applies to the Stein discrepancy estimator when the set of test functions $\mathcal{H}$ is the unit ball of the space of $\mathcal{F}^{d+1}$ in a mixed $\mathcal{F}/\ell_2$ norm, with $\mathcal{F} = \mathcal{F}_1$ or $\mathcal{F}_2$. For $\mathcal{F}_1$, we will denote this particular setting as $\mathcal{F}_1$-Stein discrepancy, or $\mathcal{F}_1$-SD. Although $\mathcal{F}_1$-SD has not been studied before to our knowledge, the empirical work of Grathwohl et al. (2020) does use Stein discrepancies with neural network test functions, which provides practical motivation for considering such a metric.

Theorem 2. *Let $K = \mathbb{S}^d$. Assume that the class $\mathcal{F}$ is such that $\sup_{f\in\mathcal{F}}\{\|\nabla_i f\|_\infty | 1 \leq i \leq d+1\} \leq \beta C_1$. If $\mathcal{H} = \mathcal{B}_{\mathcal{F}_1^{d+1}} = \{h = (h_i)_{i=1}^{d+1} \mid h_i \in \mathcal{F}_1, \sum_{i=1}^{d+1} \|h_i\|_{\mathcal{F}_1}^2 \leq 1\}$ or $\mathcal{H} = \mathcal{B}_{\mathcal{F}_2^{d+1}} = \{h = (h_i)_{i=1}^{d+1} \mid h_i \in \mathcal{F}_2, \sum_{i=1}^{d+1} \|h_i\|_{\mathcal{F}_2}^2 \leq 1\}$, we have that for the estimator $\hat{\nu}$ defined in (3), with probability at least $1-\delta$,*

$$
\begin{aligned}
SD_{\mathcal{H}}(\nu, \hat{\nu}) \leq & \frac{4\sqrt{d+1}(\beta C_1 + C_2\sqrt{d+1} + d)}{\sqrt{n}} \\
& + 2(\beta C_1 + d + 1)\sqrt{\frac{(d+1)\log(\frac{d+1}{\delta})}{2n}} \\
& + \inf_{f\in\mathcal{F}} \mathbb{E}_\nu \left[\left\| -\nabla f(x) - \nabla \log\left(\frac{d\nu}{d\tau}(x)\right) \right\|_2 \right]
\end{aligned}
$$

where C_2 is a universal constant and ∇f denotes the Riemannian gradient of f.

Notice that unlike in Theorem 1, the statistical error terms in Theorem 2 depend on the ambient dimension d. While we do not show that this dependence is necessary, studying this question would be an interesting future direction. Remark as well the similarity of the approximation term with the term $2\inf_{f\in\mathcal{F}}\|g-f\|_\infty$ from equation (8), albeit in this case it involves the L^∞ norm of the gradients. Furthermore, note that the only assumption on the set $\mathcal{F}$ is a uniform L^∞ bound on $\mathcal{F}_1$, while Theorem 1 also requires a more restrictive Rademacher complexity bound on $\mathcal{F}$. This illustrates the fact that the Stein discrepancy is a weaker metric than the KL divergence.

In Theorem 3 we give an analogous result for the unbiased KSD estimator (6), under the following reasonable assumptions on the kernel k, which follow (Liu et al., 2016).

Assumption 2. *The kernel k has continuous second order partial derivatives, and satisfies $\int_{\mathbb{S}^d}\int_{\mathbb{S}^d} g(x)k(x,x')g(x')d\tau(x)d\tau(x') > 0$ for any non-zero function $g \in L^2(\mathbb{S}^d)$, $\sup_{x,x'\in\mathbb{S}^d} k(x,x') \leq C_2$, $\sup_{x,x'\in\mathbb{S}^d} \|\nabla_x k(x,x')\|_2 \leq C_3$.*

Theorem 3. *Let $K = \mathbb{S}^d$. Assume that the class $\mathcal{F}$ is such that $\sup_{f\in\mathcal{F}}\{\|\nabla f\|_\infty\} \leq \beta C_1$. Let KSD be the kernelized Stein discrepancy for a kernel that satisfies Assumption 2. If we take n samples $\{x_i\}_{i=1}^n$ of a target measure ν with almost everywhere differentiable log-density, and consider the unbiased KSD estimator (6), we have with probability at least $1-\delta$,*

$$
\begin{aligned}
KSD(\nu,\hat{\nu}) \leq & \frac{2}{\sqrt{\delta n}}((\beta C_1 + d)^2 C_2 + 2C_3(\beta C_1 + d)) \\
& + C_2 \inf_{f\in\mathcal{F}} \mathbb{E}_{x\sim\nu}\left[\left\| \nabla \log\left(\frac{d\nu}{d\tau}(x)\right) - \nabla f(x) \right\|^2 \right].
\end{aligned}
$$

The statistical error term in Theorem 3 is obtained using the expression of the variance of the estimator (6) (Liu et al., 2016). Note that Assumption 2 is fulfilled, for example, for the radial basis function (RBF) kernel $k(x,x') = \exp(-\|x-x'\|^2/(2\sigma^2))$ with $C_2 = 1, C_3 = 1/\sigma^2$.

Making use of Theorem 2 (for $\mathcal{F}_1$-SD) and Theorem 3 (for KSD), in Corollary 2 we obtain adaptivity results for target measures with low-dimensional structures similar to Corollary 1, also for $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(\beta)$. The class of target measures that we consider are those satisfying Assumption 3, which is similar to Assumption 1 but for $K = \mathbb{S}^d$ and with an additional Lipschitz condition on the gradient of $\nabla\varphi_j$.

Assumption 3. *Suppose that the target probability measure ν is absolutely continuous w.r.t. the Hausdorff measure over $\mathbb{S}^d$ and it satisfies $\forall x \in \mathbb{S}^d$, $\frac{d\nu}{d\tau}(x) = \exp(-\sum_{j=1}^J \varphi_j(U_j x))/\int_{K_0}\exp(-\sum_{j=1}^J \varphi_j(U_j y))d\tau(y)$, where φ_j are 1-homogeneous differentiable functions on the unit ball of $\mathbb{R}^k$ such that $\|\varphi_j\|_\infty \leq \eta$, $\sup_{x\in\mathbb{S}^d}\|\nabla\varphi_j(x)\|_2 \leq \eta$ and $\nabla\varphi_j$ is L-Lipschitz continuous, and $U_j \in \mathbb{R}^{k\times d}$ with orthonormal rows.*

Corollary 2. *Let $\mathcal{F} = B_{\mathcal{F}_1}(\beta)$. Suppose $K = \mathbb{S}^d$. Let Assumption 3 hold. (i) When $\hat{\nu}$ is the $\mathcal{F}_1$-SD estimator (2)* *and the assumptions of Theorem 2 hold, we can choose the inverse temperature $\beta > 0$ such that with probability at least $1 - \delta$ we have that $SD_{\mathcal{B}^{d+1}_{\mathcal{F}_1}}(\nu, \hat{\nu})$ is upper-bounded by*

$$\tilde{O}\left(\left(1 + \sqrt{\log(1/\delta)}\right) J(L+\eta)(\eta J)^{\frac{2}{k+1}} d^{\frac{1}{k+3}} n^{-\frac{1}{k+3}}\right)$$

where the notation $\tilde{O}$ indicates that we overlook logarithmic factors and constants depending only on the dimension. (ii) When $\hat{\nu}$ is the unbiased KSD estimator (6) and the assumptions of Theorem 3 hold, $\beta > 0$ can be chosen so that with probability at least $1 - \delta$ we have that $KSD(\nu, \hat{\nu})$ is upper-bounded by

$$\tilde{O}\left(\delta^{-\frac{1}{k+3}} \left(J(L+\eta)\right)^{\frac{2(k+1)}{k+3}} (\eta J)^{\frac{4}{k+3}} n^{-\frac{1}{k+3}}\right).$$

Noticeably, the rates in Corollary 2 are also of the form $\mathcal{O}(n^{-\frac{1}{k+3}})$, which means that just as in Corollary 1, the low-dimensional structure in the target measure helps in breaking the curse of dimensionality.

Proof sketch. The main challenge in the proof of Corollary 2 is to bound the approximation terms in Theorem 2 and Theorem 3. To do so, we rely on Lemma 7 in App. A, which shows the existence of $\hat{g}$ in a ball of $\mathcal{F}_2$ such that $\sup_{x \in \mathbb{S}^d} \|\nabla \hat{g}(x) - \nabla g(x)\|_2$ has a certain bound when g is bounded and has bounded and Lipschitz gradient. Lemma 7 might be of independent interest: in particular, it can be used to obtain a similar adaptivity result for score-matching EBMs, which optimize the Fisher divergence $\mathbb{E}_{x\sim\nu}[\|\nabla \log(\frac{d\nu}{dp}(x)) - \nabla f(x)\|^2]$.

5. Algorithms

This section provides a description of the optimization algorithms used for learning $\mathcal{F}_{1/2}$-EBMs using the estimators studied in Sec. 4, namely maximum likelihood, KSD, and $\mathcal{F}_1$-SD.

5.1. Algorithms for $\mathcal{F}_1$ EBMs

We provide the algorithms for the three models using a common framework. We define the function $\Phi : \mathbb{R} \times \mathbb{R}^{d+1} \to \mathcal{F}_1$ as $\Phi(w,\theta)(x) = w\sigma(\langle\theta, x\rangle)$. Given a convex loss $R : \mathcal{F}_1 \to \mathbb{R}$, we consider the problem

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}^{d+2})} F(\mu),$$
$$F(\mu) := R\left(\int \Phi(w,\theta) d\mu\right) + \lambda \int (|w|^2 + \|\theta\|_2^2) d\mu. \tag{9}$$

for some $\lambda > 0$. It is known (e.g., Neyshabur et al., 2015) that, since $|w|^2 + \|\theta\|_2^2 \geq 2|w|\|\theta\|_2$ with equality when moduli are equal, this problem is equivalent to

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}\times\mathbb{S}^d)} R\left(\int \Phi(w,\theta) d\mu\right) + \lambda \int_{\mathbb{R}\times\mathbb{S}^d} |w| d\mu.$$

And by the definition of the $\mathcal{F}_1$ norm, this is equivalent to $\inf_{f\in\mathcal{F}_1} R(f) + \lambda\|f\|_{\mathcal{F}_1}$, which is the penalized form of $\inf_{f\in\mathcal{B}_{\mathcal{F}_1}(\beta)} R(f)$ for some $\beta > 0$. Our $\mathcal{F}_1$ EBM algorithms solve problems of the form (9) for different choices of R, or equivalently, minimize the functional R over an $\mathcal{F}_1$ ball. The functional R takes the following forms for the three models considered:

(i) Cross-entropy: We have that $R(f) = \frac{1}{n}\sum_{i=1}^n f(x_i) + \log\left(\int_K e^{-f(x)} d\tau(x)\right)$, which is convex (and differentiable) because the free energy obeys such properties (e.g., by adapting Wainwright & Jordan, 2008, Prop 3.1 to the infinite-dimensional case).

(ii) Stein discrepancy: the estimator (5) corresponds to $R(f) = \sup_{h\in\mathcal{H}} \mathbb{E}_{\nu_n}[\sum_{j=1}^{d+1} -(\nabla_j f(x) + dx_j)h_j(x) + \nabla_j h_j(x)]$, which is convex as the supremum of convex (linear) functions.

(iii) Kernelized Stein discrepancy: we have $R(f) = \frac{1}{n^2}\sum_{i,j=1}^n \tilde{u}_{\nu_f}(x_i, x_j)$, which is convex (in fact, it is quadratic in ∇f).

In order to optimize (9), we discretize measures in $\mathcal{P}(\mathbb{R}^{d+2})$ as averages of point masses $\frac{1}{m}\sum_{i=1}^m \delta_{(w^{(i)},\theta^{(i)})}$, each point mass corresponding to one neuron. Furthermore, we define the function $G : (\mathbb{R}^{d+2})^m \to \mathbb{R}$ as

$$G((w^{(i)},\theta^{(i)})_{i=1}^m) := F\left(\frac{1}{m}\sum_{i=1}^m \delta_{(w^{(i)},\theta^{(i)})}\right) \tag{10}$$
$$= R\left(\frac{1}{m}\sum_{i=1}^m \Phi(w^{(i)},\theta^{(i)})\right) + \frac{\lambda}{m}\sum_{i=1}^m (|w^{(i)}|^2 + \|\theta^{(i)}\|_2^2).$$

Then, as outlined in Algorithm 1, we use gradient descent on G to optimize the parameters of the neurons, albeit possibly with noisy estimates of the gradients.

Computing an estimate the gradient of G involves computing the gradient of $R\left(\frac{1}{m}\sum_{i=1}^m \Phi(w^{(i)},\theta^{(i)})\right)$. Denoting by $z_i = (w^{(i)},\theta^{(i)})$, $\mathbf{z} = (z_i)_{i=1}^m$ and by $\nu_{\mathbf{z}}$ the Gibbs measure corresponding to the energy $f_{\mathbf{z}} := \frac{1}{m}\sum_{i=1}^m \Phi(w^{(i)},\theta^{(i)})$, we have

(i) Cross-entropy: The gradient of $R(f_{\mathbf{z}})$ with respect to z_i takes the expression $\mathbb{E}_{\nu_n}\nabla_{z_i}\Phi(z_i)(x) - \mathbb{E}_{\nu_{\mathbf{z}}}\nabla_{z_i}\Phi(z_i)(x)$. The expectation under $\nu_{\mathbf{z}}$ is estimated using MCMC samples of the EBM. Thus, the quality of gradient estimation depends on the performance of the MCMC method of choice, which can suffer for non-convex energies and low temperatures.

Algorithm 1 Generic algorithm to train $\mathcal{F}_1$ EBMs

```
input  $m$, stepsize $s$
  Get $m$ i.i.d. samples $(w_t^{(i)}, \theta_t^{(i)})$ from $\mu_0 \in \mathcal{P}(\mathbb{R}^{d+2})$.
  for $t = 0, \dots, T-1$ do
    for $i = 1, \dots, m$ do
      Compute estimates $\hat{\nabla}_{w^{(i)}} G((w_t^{(i)}, \theta_t^{(i)})_{i=1}^m)$ and $\hat{\nabla}_{\theta^{(i)}} G((w_t^{(i)}, \theta_t^{(i)})_{i=1}^m)$.
      $w_{t+1}^{(i)} \leftarrow w_t^{(i)} - s\hat{\nabla}_{w^{(i)}} G((w_t^{(i)}, \theta_t^{(i)})_{i=1}^m)$
      $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} - s\hat{\nabla}_{\theta^{(i)}} G((w_t^{(i)}, \theta_t^{(i)})_{i=1}^m)$
    end for
  end for
output  Energy $\frac{1}{m}\sum_{i=1}^m \Phi(w_T^{(i)}, \theta_T^{(i)}) \in \mathcal{F}_1$.
```

(ii) $\mathcal{F}_1$ Stein discrepancy: The (sub)gradient of $R(f_{\mathbf{z}})$ w.r.t. z_i equals $\mathbb{E}_{\nu_n}[-\beta \sum_{j=1}^{d+1} \nabla_{z_i} \nabla_x (\Phi(z_i)(x)) h_j^\star(x)]$, in which $h_j^\star$ are respectively maximizers of $-(\beta \nabla_j f(x) + dx_j) h_j(x) + \nabla_j h_j(x)$ over $\mathcal{B}_{\mathcal{F}_1}$. The gradient estimation involves $d+1$ optimization procedures over balls of $\mathcal{F}_1$ to compute $h_j^\star$, which we solve using Algorithm 1. Thus, the algorithm operates on two timescales.

(iii) Kernelized Stein discrepancy: Using (4), the gradient of $R(f_{\mathbf{z}})$ with respect to z_i takes the expression $\mathbb{E}_{x,x' \sim \nu_n}[\nabla_{z_i} u_{\nu_{\mathbf{z}}}(x, x')]$, which can be developed into closed form. The only issue is the quadratic dependence on the number of samples.

5.2. Algorithms for $\mathcal{F}_2$ EBMs

Considering convex losses $R : \mathcal{F}_1 \to \mathbb{R}$ as in Subsec. 5.1, the penalized form of the problem $\inf_{f \in \mathcal{B}_{\mathcal{F}_2}(\beta)} R(f)$ is

$$\inf_{\|h\|_2 \leq 1} R\left(\int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle) h(\theta) d\tau(\theta) \right) + \lambda \int_{\mathbb{S}^d} h^2(\theta) d\tau(\theta).$$

To optimize this, we discretize the problem: we take m samples $(\theta^{(i)})_{i=1}^m$ of the uniform measure τ that we keep fixed, and then solve the random features problem

$$\inf_{\substack{w \in \mathbb{R}^m \\ \|w\|_2 \leq 1}} R\left(\frac{1}{m} \sum_{i=1}^m w^{(i)} \sigma(\langle \theta^{(i)}, \cdot \rangle) \right) + \frac{\lambda}{m} \sum_{i=1}^m |w^{(i)}|^2 \quad (11)$$

Remark that this objective function is equivalent to the objective function $G((w^{(i)}, \theta^{(i)})_{i=1}^m)$ in equation (10) when $(\theta^{(i)})_{i=1}^m$ are kept fixed. Thus, we can solve (11) by running Algorithm 1 without performing gradient descent updates on $(\theta^{(i)})_{i=1}^m$. That is, while for the $\mathcal{F}_1$ EBM training both the features and the weights are learned via gradient descent, for $\mathcal{F}_2$ only the weights are learned.

5.3. Qualitative convergence results

The overparametrized regime corresponds to taking a large number of neurons m. In the limit $m \to \infty$, under appropriate assumptions the empirical measure dynamics corresponding to the gradient flow of $G((w^{(i)}, \theta^{(i)})_{i=1}^m)$ converge weakly to the mean-field dynamics (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018). Leveraging a result from Chizat & Bach (2018) we argue informally that in the limit $m \to \infty, t \to \infty$, with continuous time and exact gradients, the gradient flow of G converges to the global optimum of F over $\mathcal{P}(\mathbb{R}^{d+2})$ (see more details in App. B).

In contrast with this positive qualitative result, we should mention a computational aspect that distinguishes these algorithms from their supervised learning counterparts: the Gibbs sampling required to estimate the gradient at each timestep. A notorious challenge is that for generic energies (even generic energies in $\mathcal{F}_1$), either the mixing time of MCMC algorithms is cursed by dimension (Bakry et al., 2014) or the acceptance rate is exponentially small. The analysis of the extra assumptions on the target energy and initial conditions that would avoid such curse are beyond the scope of this work, but a framework based on thermodynamic integration and replica exchange (Swendsen & Wang, 1986) would be a possible route forward.

6. Experiments

In this section, we present numerical experiments illustrating our theory on simple synthetic datasets generated by teacher models with energies $f^*(x) = \frac{1}{J}\sum_{j=1}^J w_j^* \sigma(\langle \theta_j^*, x \rangle)$, with $\theta_i^* \in \mathbb{S}^d$ for all i. The code for the experiments is in `https://github.com/CDEnrich/ebms_shallow_nn`.

Experimental setup. We generate data on the sphere $\mathbb{S}^d$ from teacher models by using a simple rejection sampling strategy, given an estimate of the minimum of f^* (which provides an estimated upper bound on the unnormalized density e^{-f^*} for rejection sampling). This minimum is estimated using gradient descent with many random restarts from uniform points on the sphere. For different numbers of training samples, we run our gradient-based algorithms in $\mathcal{F}_1$ and $\mathcal{F}_2$ with different choices of step-sizes and regularization parameters λ, using $m = 500$ neurons. We report test metrics after selecting hyperparameters on a validation set of 2000 samples. For computing gradients in maximum likelihood training, we use a simple Metropolis-Hastings algorithm with uniform proposals on the sphere. To obtain non-negative test KL divergence estimates, which are needed for the log-log plots, we sample large numbers of points uniformly on the hypersphere, and compute the KL divergence of the restriction of the EBMs to these points.

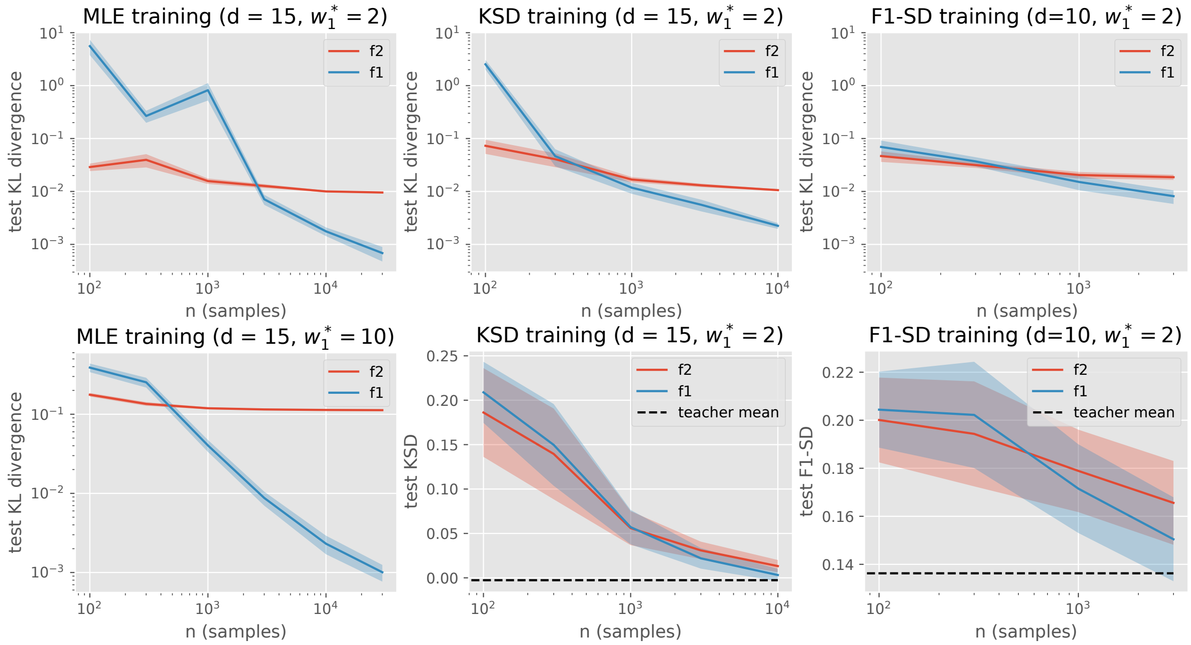

Figure 1. Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a one-neuron teacher with positive output weight. (top) Test performance measured with KL divergence estimates for $w_1^* = 2$. (bottom left) MLE on a teacher network with larger weight $w_1^* = 10$. (bottom center/right) Test KSD and $\mathcal{F}_1$-SD for models trained with the same metric with $w_1^* = 2$. For reference, the black discontinuous lines show the teacher KSD and $\mathcal{F}_1$-SD of the teacher model w.r.t. 5000 and 2000 test samples, respectively. Confidence estimates are over 10 different data samplings.

The sampling techniques that we use are effective for the toy problems considered, but more refined techniques might be needed for more complex problems in higher dimension or lower temperatures.

Learning planted neuron distributions in hyperspheres. We consider the task of learning planted neuron distributions in $d = 15$ and $d = 10$. Remark that in this setting, when $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(\beta)$ with β large enough there is no approximation error. We compare the behavior of $\mathcal{F}_1$ and $\mathcal{F}_2$ models with different estimators in Figures 1, 4 and 5 (in App. C), corresponding to models with $J = 1, 2, 4$ teacher neurons, respectively. The error bars show the average and standard deviation for 10 runs. In the three figures, the top plot in the first column represents the test KL divergence of the $\mathcal{F}_1$ and $\mathcal{F}_2$ EBMs trained with maximum likelihood for an increasing number of samples, showcasing the adaptivity of $\mathcal{F}_1$ to distributions with low-dimensional structure versus the struggle of the $\mathcal{F}_2$ model. In Figures 1 and 4 the bottom plot in the first column shows the same information for a teacher with the same structure but different values for the output weights. The separation between the $\mathcal{F}_1$ and the $\mathcal{F}_2$ models increases with higher teacher models weights.

In the three figures, the plots in the second column show the test KL divergence and test KSD, respectively, for EBMs trained with KSD (RBF kernel with $\sigma^2 = 1$). We observe that we can train EBMs successfully by optimizing the KSD; even though maximum likelihood training is directly optimizing the KL divergence, the test KL divergence values we obtain for the KSD-trained models are on par, or even slightly better, comparing at equal values of n. It is also worth noticing that in Figure 1, we observe a separation between $\mathcal{F}_1$ and $\mathcal{F}_2$ in the KL divergence plot, but not in the KSD plot. Although the training is successful, we infer that the KSD is too weak of a metric to tell that the $\mathcal{F}_1$ EBMs are better than $\mathcal{F}_2$ EBMs. In the three figures, the plots in the third column show the test KL divergences and test $\mathcal{F}_1$-SD for EBMs trained with $\mathcal{F}_1$-SD. The error bars are wider due to the two timescale algorithm used for $\mathcal{F}_1$-SD, which seems to introduce more variability. While the plots only go up to $n = 3000$, the test cross-entropy curves show a separation between $\mathcal{F}_1$ and $\mathcal{F}_2$ very similar to maximum likelihood training when comparing at equal values of n. App. C contains additional experiments for the cases $J = 1, 2$.

3D visualizations and time evolution in $d = 3$ ($\mathcal{F}_1$ EBM trained with MLE). Figure 2 shows a 3D visualization of the teacher and trained models, energies and densities corresponding to two teacher neurons with negative weights in $d = 3$. Since the dimension is small and the temperature is not too small, we used train and test sizes for which the incurred statistical error is negligible. Interestingly, while the $\mathcal{F}_1$ model achieves a KL divergence close to zero at the end of training (Figure 3), in Figure 2 we see that the

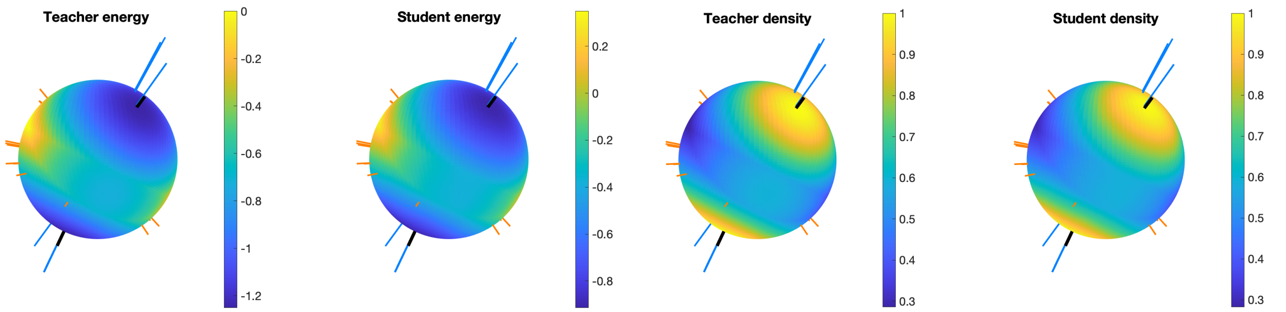

Figure 2. 3D visualization of the neuron positions, energies and densities, in $d = 3$. The teacher model has two neurons with negative weights $w_1^*, w_2^* = -2.5$, whose positions are represented by black sticks in all the images. The positions of the neurons of the trained model are represented by blue and orange sticks for negative and positive weights, resp. The two images on the left show the energies of the teacher and trained models, respectively. The energies look qualitatively very similar up to an offset of ≈ 0.3. The two images on the right show the Gibbs densities of the teacher and trained models, respectively.

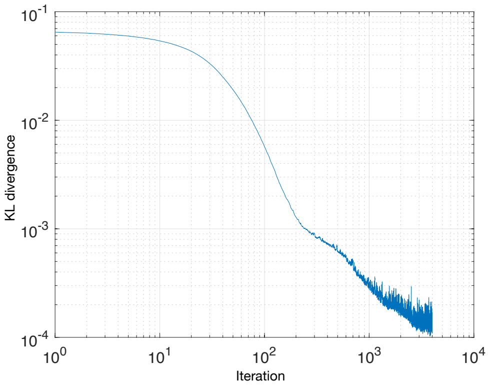

Figure 3. Log-log plot of the KL divergence between the MLE trained model and the teacher model (same as in Figure 2), versus the iteration number.

positions of the neurons of the trained model do not match the teacher neurons. In fact, there are some neurons with positive weights in the high energy region. This effect might be linked with the fact that there is a constant offset of around 0.3 between the teacher energy and the trained energy, not reflected in the Gibbs measures of the models. Figure 3 also shows that for this particular instance, the convergence is polynomial in the iteration number. A video of the training dynamics is attached in the GitHub folder.

7. Conclusions and discussion

We provide statistical error bounds for EBMs trained with KL divergence or Stein discrepancies, and show benefits of using energy models with infinite-width shallow networks in in "active" regimes in terms of adaptivity to distributions with low-dimensional structure in the energy. We empirically verify that networks in "kernel" regimes perform significantly worse in the presence of such structures, on simple teacher-student experiments.

A theoretical separation result in KL divergence or SD between $\mathcal{F}_1$ and $\mathcal{F}_2$ EBMs remains an important open question: one major difficulty for providing a lower bound on the performance for $\mathcal{F}_2$ is that L^2 (or L^∞) approximation may be not be appropriate for capturing the hardness the problem, since log-densities differing greatly in low energy regions can have arbitrarily small KL divergence. Another direction for future work is to apply the theory of shallow overparametrized neural networks to other generative models such as GANs or normalizing flows. On the computational side, in App. B we leverage existing work to state qualitative convergence results in an idealized setting of infinite width and exact gradients, but it would be interesting to develop convergence results for maximum likelihood that take the MCMC sampling into account, as done for instance by Bortoli et al. (2020) for certain exponential family models. In our setting, this would entail identifying a computationally tractable subset of $\mathcal{F}_1$ energies. A more ambitious goal is to instead move beyond the MCMC paradigm, and devise efficient sampling strategies that can operate outside the class of log-concave densities, e.g. (Gabrié et al., 2021).

Acknowledgements

We thank Marylou Gabrié for useful discussions. CD acknowledges partial support by "la Caixa" Foundation (ID 100010434), under agreement LCF/BQ/AA18/11680094. EVE acknowledges partial support from the National Science Foundation (NSF) Materials Research Science and Engineering Center Program grant DMR-1420073, NSF DMS- 1522767, and the Vannevar Bush Faculty Fellowship. JB acknowledges partial support from the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF 1845360, NSF CHS-1901091 and Samsung Electronics.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel, 2008.

Atkinson, K. and Han, W. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044. Springer, 01 2012.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017a.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, 18(1):714–751, January 2017b. ISSN 1532-4435.

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and Geometry of Markov Diffusion Operators*. Grundlehren der mathematischen Wissenschaften. Springer International Publishing, 2014. ISBN 978-3-319-00227-9.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

Bartlett, P. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2002.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Bortoli, V. D., Durmus, A., Pereyra, M., and Vidal, A. F. Efficient stochastic optimisation by unadjusted langevin monte carlo. application to maximum marginal likelihood and empirical bayesian estimation, 2020.

Borwein, J. and Zhu, Q. *Techniques of Variational Analysis*. CMS Books in Mathematics. Springer-Verlag New York, 2005.

Bourgain, J. and Lindenstrauss, J. Projection bodies. In *Geometric Aspects of Functional Analysis*, pp. 250–270. Springer, 1988.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3036–3046, 2018.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems 22*, pp. 342–350. Curran Associates, Inc., 2009.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 2606–2615. PMLR, 2016.

Della Pietra, S., Della Pietra, V., and Lafferty, J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997. doi: 10.1109/34.588021.

Du, Y. and Mordatch, I. Implicit generation and generalization in energy-based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. Adaptive monte carlo augmented with normalizing flows, 2021.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In *NeurIPS*, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods?, 2020.

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.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, volume 28, pp. 226–234. Curran Associates, Inc., 2015.

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

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

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In

Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580. Curran Associates, Inc., 2018.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, volume 21, pp. 793–800. Curran Associates, Inc., 2009.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kneser, H. Sur un theoreme fondamentale de la theorie des jeux. *C. R. Acad. Sci. Paris*, 234:2418–2420, 1952.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. 2006.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2378–2386. Curran Associates, Inc., 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR.

Malach, E., Kamath, P., Abbe, E., and Srebro, N. Quantifying the benefit of using differentiable learning over tangent kernels. *arXiv preprint arXiv:2103.01210*, 2021.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, 2012.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

Ongie, G., Willett, R., Soudry, D., and Srebro, N. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations (ICLR 2020)*, 2019.

Posner, E. C. Random coding strategies for minimum entropy. *IEEE Transations on Information Theory*, 21(4): 388–391, 1975.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. Curran Associates, Inc., 2008.

Ranzato, M., Poultney, C., Chopra, S., et al. Efficient learning of sparse representations with an energy-based model. 2007.

Rotskoff, G. M. and Vanden-Eijnden, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

Ruelle, D. *Statistical mechanics: Rigorous results*. W.A. Benjamin, 1969.

Savarese, P., Evron, I., Soudry, D., and Srebro, N. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, 2019.

Serfling, R. *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons, 2009.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. Nonparametric density estimation under adversarial losses, 2018.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

Song, Y. and Kingma, D. P. How to train your energy-based models, 2021.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pp. 583–602, 1972.

Swendsen, R. H. and Wang, J.-S. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986.

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

von Luxburg, U. and Bousquet, O. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5: 669–695, 2004.

Wainwright, M. and Jordan, M. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 01 2008.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel, 2020.

Williams, F., Trager, M., Silva, C., Panozzo, D., Zorin, D., and Bruna, J. Gradient dynamics of shallow univariate relu networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, 2020.

Xie, J., Lu, Y., Zhu, S.-C., and Wu, Y. A theory of generative convnet. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*. PMLR, 2016.

Xie, J., Zhu, S., and Wu, Y. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.