On Energy-Based Models with Overparametrized Shallow Neural Networks

Carles Domingo-Enrich[a], Alberto Bietti[b], Eric Vanden-Eijnden[a], and Joan Bruna[a,b]

[a]Courant Institute of Mathematical Sciences, New York University
[b]Center for Data Science, New York University

May 6, 2021

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 and 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 and Jordan, 2008, Ranzato et al., 2007, LeCun et al., 2006, Du and Mordatch, 2019, Song and 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 and 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 and 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 and 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 and Vanden-Eijnden, 2018, Chizat and Bach, 2018, Sirignano and Spiliopoulos, 2019]. As it is also 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 and Mackey, 2015, Liu and 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 and Kingma, 2021, Song and 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 and 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 and 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 and 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 was also studied by Gorham and Mackey [2017]. Liu and 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 $|\gamma|_{\mathrm{TV}}$ 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., $\hat{f} = \operatorname{argmax}_{f\in\mathcal{F}} \prod_{i=1}^{n} \frac{d\nu_f}{d\tau}(x_i)$, or equivalently, the one that minimizes the cross-entropy with the samples:

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

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 and Mackey, 2015, Liu et al., 2016] for $\mathcal{H}$ is a non-symmetric functional defined on pairs of probability measures over K as

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

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 $\mathrm{SD}_{\mathcal{H}}(\nu,\nu) = 0$, while it does not appear when considering $K = \mathbb{R}^d$. The Stein discrepancy estimator is

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

If $\mathcal{H} = \mathcal{B}_{\mathcal{H}_0^{d+1}} = \{(h_i)_{i=1}^{d+1} \in \mathcal{H}_0^{d+1} \mid \sum_{i=1}^{d+1} \|h_i\|^2_{\mathcal{H}_0} \le 1\}$ 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 and Mackey, 2017]. For $K = \mathbb{S}^d$, the KSD takes the following form (Lemma 6):

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

where $u_\nu(x,x') = (s_\nu(x)-d\cdot x)^\top(s_\nu(x')-d\cdot x')k(x,x') + (s_\nu(x)-d\cdot x)^\top \nabla_{x'} k(x,x') + (s_\nu(x')-d\cdot x')^\top \nabla_x k(x,x') + \mathrm{Tr}(\nabla_{x,x'} k(x,x'))$, 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

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

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

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

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 Radon measure $\gamma \in \mathcal{M}(\mathbb{S}^d)$. The norm of $\mathcal{F}_1$ is defined as $\|f\|_{\mathcal{F}_1} = \inf \left\{ |\gamma|_{\mathrm{TV}} \mid f(\cdot) = \int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle)\, d\gamma(\theta) \right\}$.

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 $\|f\|^2_{\mathcal{F}_2} = \inf \left\{ \int_{\mathbb{S}^d} |h(\theta)|^2\, d\tilde{\tau}(\theta) \mid f(\cdot) = \int_{\mathbb{S}^d} \sigma(\langle \theta, \cdot \rangle) h(\theta)\, d\tilde{\tau}(\theta) \right\}$. 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, 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 by 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. *The target probability measure* ν *is absolutely continuous w.r.t. the uniform probability measure* τ *over* K *and it satisfies* $\forall x\in K_0$, $\frac{d\nu}{d\tau}(x,R) = \exp(-\sum_{j=1}^J \varphi_j(U_j x))/\int_{K_0}\exp(-\sum_{j=1}^J\varphi_j(U_j y))d\tau(y)$, *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* $\mathcal{F}=B_{\mathcal{F}_1}(\beta)$. *Suppose* $K=K_0\times\{R\}$, *where* $K_0\subseteq\{x\in\mathbb{R}^d|\|x\|_2\leq R\}$ *is compact. 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 it satisfies that for any non-zero function* $g \in L^2(\mathbb{S}^d)$*,* $\int_{\mathbb{S}^d}\int_{\mathbb{S}^d} g(x)k(x,x')g(x')d\tau(x)d\tau(x') > 0$*, and that* $\sup_{x,x'\in\mathbb{S}^d} k(x,x') \leq C_2$*,* $\sup_{x,x'\in\mathbb{S}^d} \|\nabla_x k(x,x')\|_2 \leq C_3$*.*

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* $1-\delta$*,*

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

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. *Let* $K = \mathbb{S}^d$*. 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* $\mathcal{F} = B_{\mathcal{F}_1}(\beta)$*. 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* $\mathrm{KSD}(\nu, \hat{\nu})$ *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

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

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 and 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.

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 $t = 0, \ldots, T-1$ **do**
    for $i = 1, \ldots, m$ **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 $\frac{1}{m}\sum_{i=1}^{m} \Phi(w_T^{(i)}, \theta_T^{(i)}) \in \mathcal{F}_1$.

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.

(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

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

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 and Bach [2018], Rotskoff and Vanden-Eijnden [2018]. Leveraging a result from Chizat and 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 and 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

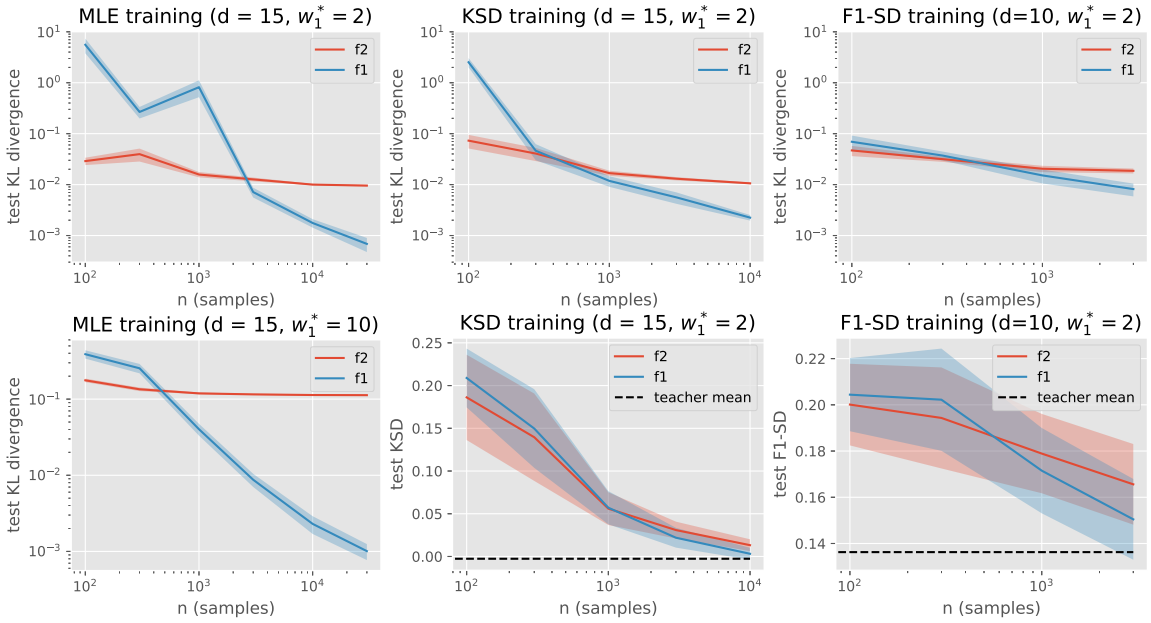

Figure 1: Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a one-neuron teacher with positive output weight. (top) Test performance measured with KL divergence estimates for $w_1^* = 2$. (bottom left) MLE on a teacher network with larger weight $w_1^* = 10$. (bottom center/right) 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.

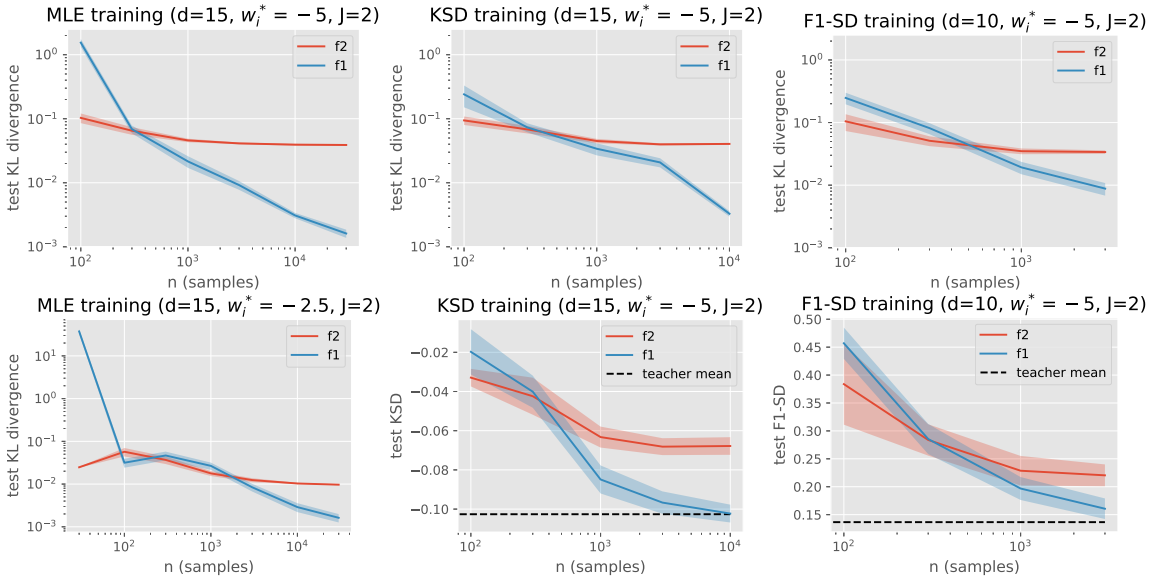

Figure 2: Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a two-neuron teacher with negative output weights. (top) Test performance measured with cross-entropy estimates with $w_1^*, w_2^* = -5$. (bottom left) MLE on a teacher network with smaller weights $w_1^*, w_2^* = -2.5$. (bottom center/right) Test KSD and $\mathcal{F}_1$-SD for models trained with the same metric, for $w_1^*, w_2^* = -5$. 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.

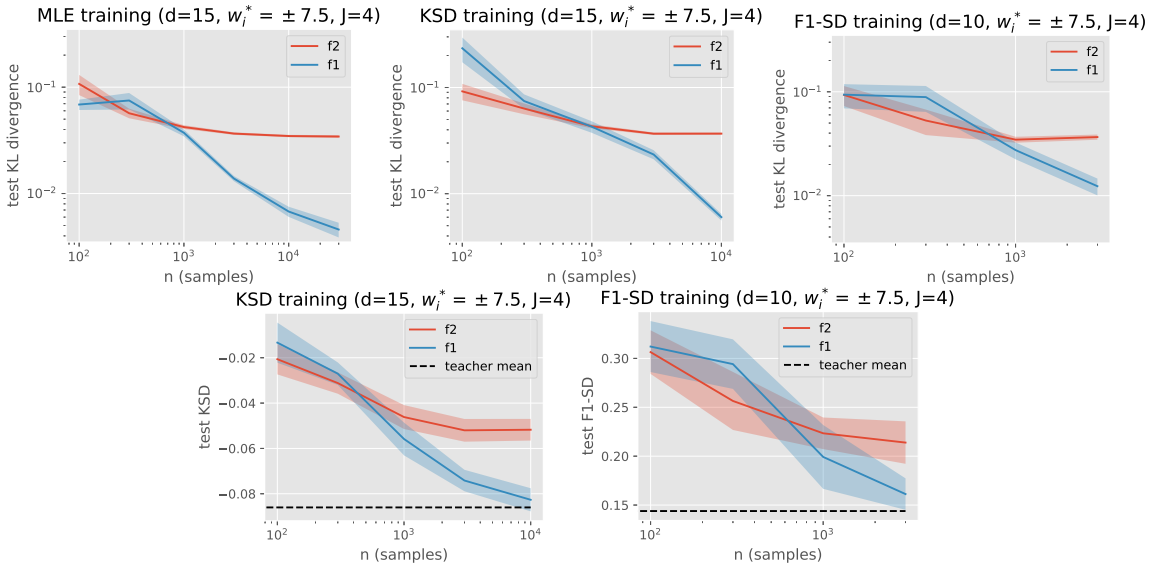

Figure 3: Test metrics obtained for MLE, KSD and $\mathcal{F}_1$-SD training on a four-neuron teacher with weights $w_1^*, w_2^* = 7.5$ and $w_3^*, w_4^* = -7.5$. (top) Test performance measured with cross-entropy estimates. (bottom) Test KSD and $\mathcal{F}_1$-SD for models trained with the same metric. 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.

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. 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, 2 and 3, 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 2 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. We observe that the separation between the $\mathcal{F}_1$ and the $\mathcal{F}_2$ models increases when the teacher models have higher 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 (with RBF kernel with $\sigma^2 = 1$). We observe that we are able to 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. It seems that in this particular instance, although the training is successful, 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

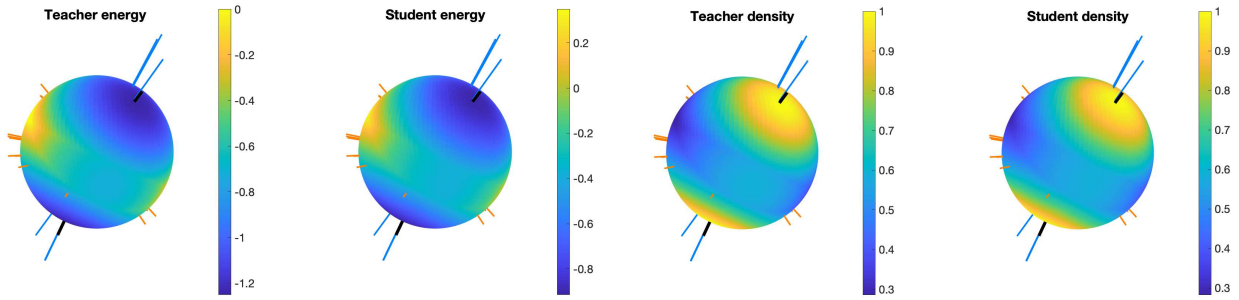

Figure 4: 3D visualization of the neuron positions, energies and densities, in $d = 3$. The teacher model has two neurons with negative weights $w_1^*, w_2^* = -2.5$, whose positions are represented by black sticks in all the images. The positions of the neurons of the trained model are represented by 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.

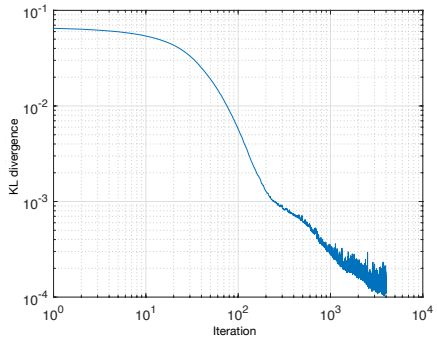

Figure 5: Log-log plot of the KL divergence between the MLE trained model and the teacher model (same as in Figure 4), versus the iteration number.

trained with $\mathcal{F}_1$-SD. Remark that 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 $J = 1, w_1^* = 10$ and $J = 2, w_i^* = -2.5$, training with KSD and $\mathcal{F}_1$-SD.

3D visualizations and time evolution in $d = 3$ ($\mathcal{F}_1$ EBM trained with MLE). Figure 4 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 statistical error due to train and test samples is negligible. Interestingly, while the $\mathcal{F}_1$ model achieves a KL divergence close to zero at the end of training (Figure 5), in Figure 4 we see that the 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. The offset is not reflected in the Gibbs measures of the models, which are invariant to constant terms.

Figure 5 also shows that for this particular instance, the convergence is polynomial in the iteration number. We attach a video of the training dynamics: `https://github.com/CDEnrich/ebms_shallow_nn/blob/main/KLfeatzunnorm1.mp4`.

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 two log-densities differing substantially 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 and long-term goal is to instead move beyond the MCMC paradigm, and devise efficient sampling strategies that can operate outside the class of log-concave densities, as for instance 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, and R. Zemel. 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, pages 3732–3747, 2020.

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

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580. Curran Associates, Inc., 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

G. M. Rotskoff and E. Vanden-Eijnden. 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.

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

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

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

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

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

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

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

C. Stein. 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*, pages 583–602, 1972.

R. H. Swendsen and J.-S. Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57: 2607–2609, Nov 1986. doi: 10.1103/PhysRevLett.57.2607. URL `https://link.aps.org/doi/10.1103/PhysRevLett.57.2607`.

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

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

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

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

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

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

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

A Proofs of Sec. 4

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, 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 $1-\delta$,

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

Proof. In the first place, remark that for all $\nu_1, \nu_2 \in \mathcal{P}(K)$ that are absolutely continuous w.r.t. p, we have $D_{KL}(\nu_1||\nu_2) = \int_K \log(\frac{d\nu_1}{d\tau}(x))d\nu_1(x) - \int_K \log(\frac{d\nu_2}{d\tau}(x))d\nu_1(x) = -H(\nu_1) + H(\nu_1,\nu_2)$, where $H(\nu_1,\nu_2)$ is the cross-entropy and $H(\nu_1)$ is the differential entropy. Hence, for all $\nu_1,\nu_2,\nu_3 \in \mathcal{P}(K)$,

$$D_{KL}(\nu_1||\nu_2) - D_{KL}(\nu_1||\nu_3) = H(\nu_1,\nu_2) - H(\nu_1,\nu_3). \tag{12}$$

Secondly, notice that for any $\nu \in \mathcal{P}(K)$ and measurable $f : K \to \mathbb{R}$,

$$\begin{aligned}\int f(x)\ d\nu(x) &= -\int \log(e^{-f(x)})\ d\nu(x) = -\int \log\left(\frac{d\nu_f}{d\tau}(x)\right)\ d\nu(x) - \log\left(\int e^{-f(x)}d\tau(x)\right)\\ &= H(\nu,\nu_f) - \log\left(\int e^{-f(x)}d\tau(x)\right),\end{aligned} \tag{13}$$

Thus, if we apply (13) on ν and its empirical version $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, we obtain that with probability at least $1-\delta$, for all $f \in \mathcal{F}$:

$$\begin{aligned}|H(\nu,\nu_f) - H(\nu_n,\nu_f)| &= \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f(x)\ d\nu(x)\right| \leq \left(2\mathcal{R}_n(\mathcal{F}) + \left(\sup_{f\in\mathcal{F}}\|f\|_\infty\right)\sqrt{\frac{2\log(1/\delta)}{n}}\right)\\ &\leq \frac{2\beta C}{\sqrt{n}} + \beta\sqrt{\frac{2\log(1/\delta)}{n}},\end{aligned} \tag{14}$$

where we have used the Rademacher generalization bound (Mohri et al. [2012], Theorem 3.3) and the Rademacher complexity bound from the assumption of the theorem.

We have

$$\begin{aligned}D_{KL}(\nu||\hat{\nu}) &= D_{KL}(\nu||\hat{\nu}) - \inf_{f\in\mathcal{F}} D_{KL}(\nu||\nu_f) + \inf_{f\in\mathcal{F}} D_{KL}(\nu||\nu_f)\\ &= \sup_{f\in\mathcal{F}}\left\{H(\nu,\hat{\nu}) - H(\nu,\nu_f)\right\} + \inf_{f\in\mathcal{F}} D_{KL}(\nu||\nu_f)\\ &\leq \sup_{f\in\mathcal{F}}\left\{H(\nu_n,\hat{\nu}) - H(\nu_n,\nu_f)\right\} + \frac{4\beta C}{\sqrt{n}} + \beta\sqrt{\frac{8\log(1/\delta)}{n}} + \inf_{f\in\mathcal{F}} D_{KL}(\nu||\nu_f)\\ &= \frac{4\beta C}{\sqrt{n}} + \beta\sqrt{\frac{8\log(1/\delta)}{n}} + \inf_{f\in\mathcal{F}} D_{KL}(\nu||\nu_f).\end{aligned}$$

This proves (7). In the second equality we have used (12). For the inequality we have used (14) twice, i.e. that $H(\nu,\hat{\nu}) = H(\nu,\nu_{\hat{f}}) \leq H(\nu_n,\nu_{\hat{f}}) + \frac{2\beta C}{\sqrt{n}} + \beta\sqrt{\frac{2\log(1/\delta)}{n}}$ and that $-H(\nu,\nu_f) \leq -H(\nu_n,\nu_f) + \frac{2\beta C}{\sqrt{n}} + \beta\sqrt{\frac{2\log(1/\delta)}{n}}$. In the last equality we have used that by the definition of $\hat{f}$, $H(\nu_n,\hat{\nu}) = H(\nu,\nu_{\beta\hat{f}}) = \min_{f\in\mathcal{F}} H(\nu,\nu_f)$.

For the proof of (8) we apply Lemma 1 into (7). □

Lemma 1. *Let $g : K \to \mathbb{R}$ be such that $\frac{d\nu}{d\tau}(x) = e^{-g(x)}/\int_K e^{-g(y)}d\tau(y)$, i.e. $-g$ is the log-density of ν up to a constant term. Then,*

$$\inf_{f\in\mathcal{F}} D_{KL}(\nu||\nu_f) \leq 2 \inf_{f\in\mathcal{F}} \|g-f\|_\infty$$

Proof. Notice that $\nu = \nu_g$. Thus, for any $f \in \mathcal{F}$,

$$\begin{aligned} D_{KL}(\nu||\nu_f) = D_{KL}(\nu_g||\nu_f) &= \int \log\left(\frac{\frac{d\nu_g}{d\tau}(x)}{\frac{d\nu_f}{d\tau}(x)}\right) d\nu_g(x) = \int \log\left(\frac{\frac{e^{-g(x)}}{\int e^{-g(y)}d\tau(y)}}{\frac{e^{-f(x)}}{\int e^{-f(y)}d\tau(y)}}\right) d\nu_g(x) \\ &= \int (f(x)-g(x))\, d\nu_g(x) - \log\left(\int e^{-g(y)}d\tau(y)\right) + \log\left(\int e^{-f(y)}d\tau(y)\right). \end{aligned} \tag{15}$$

Here, we bound

$$\int (f(x)-g(x))\, d\nu_g(x) \leq \|f-g\|_\infty,$$

and applying Lemma 2 to f and g, we obtain

$$\log\left(\int e^{-f(y)}\, d\tau(y)\right) - \log\left(\int e^{-g(y)}\, d\tau(y)\right) \leq \|f-g\|_\infty.$$

Plugging these two bounds into (15), we obtain $D_{KL}(\nu||\nu_f) \leq 2\|f-g\|_\infty$, which yields the result. □

We do not claim that the upper-bound in Lemma 1 is tight; it might be possible to provide a bound involving a weaker metric. Regardless, it suffices for our purposes.

Lemma 2. *Let $f : K \to \mathbb{R}$, $g : K \to \mathbb{R}$ be measurable functions. For some $\alpha \in [0,1]$,*

$$\log\left(\int_K e^{-f(y)}d\tau(y)\right) - \log\left(\int_K e^{-g(y)}d\tau(y)\right) = \int_K \frac{e^{-\left(\alpha f(y)+(1-\alpha)g(y)\right)}}{\int_K e^{-\left(\alpha f(x)+(1-\alpha)g(x)\right)}\, d\tau(x)} \left(f(y)-g(y)\right)\, d\tau(y)$$

Proof. We define the function

$$F(\alpha) = \log\left(\int_K e^{-\left(\alpha f(y)+(1-\alpha)g(y)\right)}\, d\tau(y)\right),$$

which has derivative

$$\frac{dF}{d\alpha}(\alpha) = \frac{-\int_K e^{-\left(\alpha f(y)+(1-\alpha)g(y)\right)}(f(y)-g(y))\, d\tau(y)}{\int_K e^{-\left(\alpha f(x)+(1-\alpha)g(x)\right)}\, d\tau(x)} = -\int_K (f(y)-g(y))p_\alpha(y)\, d\tau(y),$$

where $p_\alpha(y)$ is the density of the Gibbs probability measure corresponding to the energy $\alpha f + (1-\alpha)g$. We make use of the mean value theorem:

$$\begin{aligned} &\log\left(\int_K e^{-f(y)}d\tau(y)\right) - \log\left(\int_K e^{-g(y)}d\tau(y)\right) = F(1) - F(0) = \frac{dF}{d\alpha}(\alpha)(1-0) \\ &= -\int_K (f(y)-g(y))p_\alpha(y)\, d\tau(y). \end{aligned}$$

□

Lemma 3 (Approximation of Lipschitz functions by $\mathcal{F}_2$ balls, Proposition 6 of Bach [2017a]). *For δ greater than a constant depending only on d, for any function $f : \{x \in \mathbb{R}^d | \|x\|_2 \leq R\} \to \mathbb{R}$ such that for all x, y such that $\|x\|_2 \leq R$, $\|y\|_2 \leq R$ we have $|f(x)| \leq \eta$ and $|f(x) - f(y)| \leq \eta R^{-1} \|x - y\|_2$, there exists $h\{x \in \mathbb{R}^d | \|x\|_2 \leq R\} \times \{R\} \to \mathbb{R} \in \mathcal{F}_2$, such that $\|h\|_{\mathcal{F}_2} \leq \delta$ and*

$$\sup_{\|x\|_2 \leq R} |h(x, R) - f(x)| \leq C(d)\eta \left(\frac{R\delta}{\eta}\right)^{-2/(d+1)} \log\left(\frac{R\delta}{\eta}\right)$$

Proof. From Bach [2017a]. Notice that the factor in the bound is $\left(\frac{R\delta}{\eta}\right)^{-2/(d+1)} \log\left(\frac{R\delta}{\eta}\right)$, while in the original paper it is $\left(\frac{\delta}{\eta}\right)^{-2/(d+1)} \log\left(\frac{\delta}{\eta}\right)$. The R factor stems from the fact that we consider the neural network features to lie in $\mathbb{S}^d$, while Bach [2017a] considers them in the hypersphere of radius R^{-1}. □

Lemma 4 (Rademacher complexity bound for $\mathcal{B}_{\mathcal{F}_1}$, Section 5.1 of Bach [2017a]; Kakade et al. [2009]). *Suppose that $K \subseteq \{x \in \mathbb{R}^{d+1} | \|x\|_2 \leq R\}$. The Rademacher complexity of the function class $\mathcal{B}_{\mathcal{F}_1}$ is bounded by*

$$\mathcal{R}_n(\mathcal{B}_{\mathcal{F}_1}) \leq \frac{R}{\sqrt{n}}.$$

Corollary 1. *Let $\mathcal{F} = B_{\mathcal{F}_1}(\beta)$. Suppose $K = K_0 \times \{R\}$, where $K_0 \subseteq \{x \in \mathbb{R}^d | \|x\|_2 \leq R\}$ is compact. 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.

Proof. We will use (8) from Theorem 1. We have that $g : K_0 \times \{R\} \to \mathbb{R}$ is defined as $g(x, R) = \sum_{j=1}^J g_j(x, R) = \sum_{j=1}^J \varphi_j(U_j x, R)$.

By Lemma 3, there exists $\psi_j : \{x \in \mathbb{R}^k | \|x\|_2 \leq R\} \times R \to \mathbb{R}$ such that $\psi_j \in \mathcal{F}_2$ and $\|\psi_j\|_{\mathcal{F}_2} \leq \beta/J$, and

$$\sup_{x \in \mathbb{R}^k : \|x\|_2 \leq R} |\psi_j(x, R) - \varphi_j(x)| \leq C(k)\eta \left(\frac{R\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{R\beta}{\eta J}\right) \tag{16}$$

Hence, if we define $\tilde{g}_j : K_0 \times \{R\} \to \mathbb{R}$ as $\tilde{g}_j(x, R) := \psi_j(U_j x, R)$, we have that $\tilde{g}_j$ belongs to $\mathcal{F}_1$ by an argument similar to the one of Section 4.6 of Bach [2017a]. Namely, if we write $\psi_j(x, R) = \int_{\mathbb{S}^k} \sigma(\langle \theta, (x, R) \rangle)\, d\gamma(\theta)$ for some signed measure γ, we have

$$\tilde{g}_j(x, R) = \int_{\mathbb{S}^k} \sigma(\langle \theta, (U_j x, R) \rangle)\, d\gamma(\theta) = \int_{\mathbb{S}^k} \sigma(\langle U_j^\top \theta_{1:d}, x \rangle + \theta_{d+1} R)\, d\gamma(\theta) = \int_{\mathbb{S}^d} \sigma(\langle \theta'_{1:d}, x \rangle + \theta'_{d+1} R)\, d\gamma'(\theta'),$$

where we used the change of variable $\theta' = (U_j^\top \theta_{1:d}, \theta_{d+1})$, which maps $\mathbb{S}^k$ to $\mathbb{S}^d$. Moreover, this shows that $\tilde{g}_j$ has $\mathcal{F}_1$ norm $\|\tilde{g}_j\|_{\mathcal{F}_1} \leq \|\psi_j\|_{\mathcal{F}_2} \leq \beta/J$, which means that $\tilde{g} = \sum_{j=1}^J \tilde{g}_j \in \mathcal{F}_1$ and $\|\tilde{g}\|_{\mathcal{F}_1} \leq \beta$. Moreover,

$$\begin{aligned}
\|\tilde{g}_j - g_j\|_\infty &= \sup_{x \in K_0} |\tilde{g}_j(x, R) - g_j(x, R)| = \sup_{x \in K_0} |\psi_j(U_j x) - \varphi_j(U_j x)| \leq \sup_{x \in \mathbb{R}^k : \|x\|_2 \leq R} |\psi_j(x) - \varphi_j(x)| \\
&\leq C(k)\eta \left(\frac{R\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{R\beta}{\eta J}\right)
\end{aligned}$$

The first inequality holds because for all $x \in K$, $\|Ux\|_2 \le \|x\|_2 \le R$ by the fact that U has orthonormal rows, and the second inequality holds by (16). Thus,

$$\inf_{f \in \mathcal{B}_{\mathcal{F}_1}} \|g - f\|_\infty \le \|g - \tilde{g}\|_\infty \le \sum_{j=1}^{J} \|\tilde{g}_j - g_j\|_\infty \le C(k) J \eta \left(\frac{R\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{R\beta}{\eta J}\right) \tag{17}$$

Notice that the assumptions of Theorem 1 are fulfilled: the Rademacher complexity bound for $\mathcal{B}_{\mathcal{F}_1}$ (Lemma 4) implies that $\mathcal{R}_n(\mathcal{B}_{\mathcal{F}_1}(\beta)) \le \frac{\beta R}{\sqrt{n}}$ and it is also easy to check that $\sup_{f \in \mathcal{B}_{\mathcal{F}_1}(\beta)} \|f\|_\infty \le \beta$. Plugging (17) into (8) we obtain

$$D_{\mathrm{KL}}(\nu||\hat{\nu}) \le 4\beta \frac{\sqrt{2}R}{\sqrt{n}} + \beta \sqrt{\frac{2\log(1/\delta)}{n}} + 2C(k) J \eta \left(\frac{R\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{R\beta}{\eta J}\right).$$

If we minimize the right-hand side w.r.t. β (disregarding the log factor), we obtain that the optimal value is

$$\left(\frac{2B}{k+1}\right)^{\frac{k+1}{k+3}} \left(\frac{A}{\sqrt{n}}\right)^{\frac{2}{k+3}} + B^{\frac{k+1}{k+3}} \left(\frac{A(k+1)}{2\sqrt{n}}\right)^{\frac{2}{k+3}} \log\left(\frac{R}{\eta}\left(\frac{2B\sqrt{n}}{A(k+1)}\right)^{\frac{k+1}{k+3}}\right),$$

and the optimal β is $\left(2B\sqrt{n}/(A(k+1))\right)^{\frac{k+1}{k+3}}$, where

$$A = 4\sqrt{2}R + \sqrt{2\log(1/\delta)}, \quad B = 2C(k)(J\eta)^{\frac{k+3}{k+1}} R^{-\frac{2}{k+1}}.$$

□

Lemma 5 (Stein operator for functions on $\mathbb{S}^d$)**.** *For a probability measure ν on the sphere $\mathbb{S}^d$ with a continuous and almost everywhere differentiable density $\frac{d\nu}{d\tau}$, the Stein operator $\mathcal{A}_\nu$ is defined as*

$$(\mathcal{A}_\nu h)(x) = \left(\nabla \log\left(\frac{d\nu}{d\tau}(x)\right) - dx\right) h(x)^\top + \nabla h(x),$$

for any $h : \mathbb{S}^d \to \mathbb{R}^{d+1}$ that is continuous and almost everywhere differentiable, where ∇ denotes the Riemannian gradient. That is, for any $h : \mathbb{S}^d \to \mathbb{R}^{d+1}$ that is continuous and almost everywhere differentiable, the Stein identity holds:

$$\mathbb{E}_\nu[(\mathcal{A}_\nu h)(x)] = 0.$$

Proof. Let $h_i : \mathbb{S}^d \to \mathbb{R}$ be the i-th component of h. Notice that

$$\mathbb{E}_\nu\left[\nabla \log\left(\frac{d\nu}{d\tau}(x)\right) h_i(x) + \nabla h_i(x)\right] = \mathbb{E}_\nu\left[\nabla\left(\frac{d\nu}{d\tau}(x) h_i(x)\right) \frac{1}{\frac{d\nu}{d\tau}(x)}\right] = \int_{\mathbb{S}^d} \nabla\left(\frac{d\nu}{d\tau}(x) h_i(x)\right) d\tau(x) \tag{18}$$

Now, if we take the inner product of the right-hand side with the canonical basis vector $e_k \in \mathbb{R}^{d+1}$, we obtain

$$\begin{aligned}
&\left\langle \int_{\mathbb{S}^d} \nabla\left(\frac{d\nu}{d\tau}(x) h_i(x)\right) d\tau(x), e_k \right\rangle = \int_{\mathbb{S}^d} \left\langle (I - xx^\top) \nabla\left(\frac{d\nu}{d\tau}(x) h_i(x)\right), e_k \right\rangle d\tau(x) \\
&= \int_{\mathbb{S}^d} \left\langle \nabla\left(\frac{d\nu}{d\tau}(x) h_i(x)\right), e_k - x_k x \right\rangle d\tau(x) = -\int_{\mathbb{S}^d} \frac{d\nu}{d\tau}(x) h_i(x) \nabla \cdot (e_k - x_k x) d\tau(x),
\end{aligned}$$

where in the second equality we used that $(I - xx^\top)$ the projection matrix to the tangent space of $\mathbb{S}^d$ at x, in the third equality we used that it is symmetric, and in the last equality we used integration by parts on $\mathbb{S}^d$ ($\nabla\cdot$ denotes the Riemannian divergence).

To compute $\nabla \cdot (e_k - x_k x)$, remark that by the invariance to change of basis it is equal to the divergence of the function $g : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ defined as $x \to e_k - \frac{x_k x}{\|x\|^2}$, when restricted to $\mathbb{S}^d$. And we have

$$\nabla \cdot g(x) = \sum_{j=1}^{d+1} \partial_j g_j(x) = \sum_{j=1}^{d+1} \partial_j \left(e_{k,j} - \frac{x_k x_j}{\|x\|^2} \right) = - \left(\sum_{j=1}^{d+1} \frac{x_k}{\|x\|^2} \right) + \left(\sum_{j=1}^{d+1} \frac{2 x_k x_j^2}{\|x\|^4} \right) - \frac{x_k}{\|x\|^2}$$

For $x \in \mathbb{S}^d$, the right-hand side simplifies to $-(d+1)x_k + 2x_k - x_k = -dx_k$, which means that the right-hand side of (18) becomes

$$- \int_{\mathbb{S}^d} \frac{d\nu}{d\tau}(x) h_i(x)(-dx_k) \, d\tau(x) = d\mathbb{E}_\nu[h_i(x)x_k].$$

That means that $\mathbb{E}_\nu \left[\nabla \log \left(\frac{d\nu}{d\tau}(x) \right) h_i(x) + \nabla h_i(x) - d h_i(x) x \right] = 0$, which concludes the proof. □

Lemma 6 (Kernelized Stein discrepancy for probability measures on $\mathbb{S}^d$)**.** *For $K = \mathbb{S}^d$, and $\nu_1, \nu_2 \in \mathcal{P}(K)$ with continuous, almost everywhere differentiable log-densities, the kernelized Stein discrepancy $KSD(\nu_1, \nu_2)$ is equal to*

$$\sup_{h \in \mathcal{B}_{\mathcal{H}_0^d}} (\mathbb{E}_{\nu_1}[Tr(\mathcal{A}_{\nu_2} h(x))])^2 = \mathbb{E}_{x,x' \sim \nu_1}[(s_{\nu_2}(x) - s_{\nu_1}(x))^\top (s_{\nu_2}(x') - s_{\nu_1}(x')) k(x,x')] = \mathbb{E}_{x,x' \sim \nu_1}[u_{\nu_2}(x,x')] \tag{19}$$

where $u_\nu(x,x') = (s_\nu(x) - d \cdot x)^\top (s_\nu(x') - d \cdot x') k(x,x') + (s_\nu(x) - d \cdot x)^\top \nabla_{x'} k(x,x') + (s_\nu(x') - d \cdot x')^\top \nabla_x k(x,x') + Tr(\nabla_{x,x'} k(x,x'))$.

Proof. The argument for the first equality is from Theorem 3.8 of Liu et al. [2016], but we rewrite it with our notation. Using the Stein identity, which holds by Lemma 5, we have

$$\mathbb{E}_{\nu_1}[\mathrm{Tr}(\mathcal{A}_{\nu_2} h(x))] = \mathbb{E}_{\nu_1}[\mathrm{Tr}(\mathcal{A}_{\nu_2} h(x) - \mathcal{A}_{\nu_1} h(x))] = \mathbb{E}_{\nu_1}[(s_{\nu_2}(x) - s_{\nu_1}(x))^\top h(x)],$$

$$\sup_{h \in \mathcal{B}_{\mathcal{H}_0^d}} \mathbb{E}_{\nu_1}[(s_{\nu_2}(x) - s_{\nu_1}(x))^\top h(x)] = \sup_{h \in \mathcal{B}_{\mathcal{H}_0^d}} \int_{\mathbb{S}^d} \frac{d\nu_1}{d\tau}(x) \sum_{i=1}^{d+1} (s_{\nu_2}^{(i)}(x) - s_{\nu_1}^{(i)}(x)) h_i(x) d\tau(x)$$

$$= \sup_{h \in \mathcal{B}_{\mathcal{H}_0^d}} \sum_{i=1}^{d+1} \left\langle \int_{\mathbb{S}^d} \frac{d\nu_1}{d\tau}(x)(s_{\nu_2}^{(i)}(x) - s_{\nu_1}^{(i)}(x)) k(x,\cdot) d\tau(x), h_i(\cdot) \right\rangle_{\mathcal{H}_0} = \sqrt{\sum_{i=1}^{d+1} \left\| \int_{\mathbb{S}^d} \frac{d\nu_1}{d\tau}(x)(s_{\nu_2}^{(i)}(x) - s_{\nu_1}^{(i)}(x)) k(x,\cdot) d\tau(x) \right\|_{\mathcal{H}_0}^2}$$

$$= \sqrt{\sum_{i=1}^{d+1} \int_{\mathbb{S}^d \times \mathbb{S}^d} \frac{d\nu_1}{d\tau}(x)(s_{\nu_2}^{(i)}(x) - s_{\nu_1}^{(i)}(x)) k(x,x') \frac{d\nu_1}{d\tau}(x')(s_{\nu_2}^{(i)}(x') - s_{\nu_1}^{(i)}(x')) d\tau(x) d\tau(x')}.$$

Given the form of the Stein operator for functions on $\mathbb{S}^d$ (Lemma 5), the proof of the second equality of (19) is a straightforward analogy of the proof of Theorem 3.6 of Liu et al. [2016], which is for the Stein operator for functions on $\mathbb{R}^d$. □

Theorem 4. *Let $K = \mathbb{S}^d$. Assume that the class $\mathcal{F}$ is such that $\sup_{f \in \mathcal{F}} \{\|\nabla_i f\|_\infty | 1 \le i \le d+1\} \le \beta C_1$. Assume that $\mathcal{H} = \mathcal{B}_{\prod_{i=1}^{d+1} \mathcal{H}_i} = \{(h_i)_{i=1}^{d+1} \mid h_i \in \mathcal{H}_i, \sum_{i=1}^{d+1} \|h_i\|_{\mathcal{H}_i} \le 1\}$, where $\mathcal{H}_i$ are normed spaces of functions from $\mathbb{S}^d$ to $\mathbb{R}$. Assume that the following Rademacher complexity type bounds hold for $1 \le i \le d+1$: $\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i \in \mathcal{B}_{\mathcal{H}_i}} \frac{1}{n} \sum_{j=1}^n \sigma_j h_i(x_j)\right] \le \frac{C_2}{\sqrt{n}}$, $\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i \in \mathcal{B}_{\mathcal{H}_i}} \frac{1}{n} \sum_{j=1}^n \sigma_j \nabla_i h_i(x_j)\right] \le \frac{C_3}{\sqrt{n}}$, and that $\|h_i\|_\infty \le M, \|\nabla_i h_i\|_\infty \le M$ for all $h_i \in \mathcal{H}_i$.*

If we take n samples $\{x_i\}_{i=1}^n$ of a target measure ν with almost everywhere differentiable log-density, and consider the Stein Discrepancy estimator (SDE) $\hat{\nu} := \nu_{\hat{f}}$, where $\hat{f}$ is the estimator defined in (3), we have

that with probability at least $1-\delta$, $SD_{\mathcal{H}}(\nu,\hat{\nu})$ *is upper-bounded by*

$$\frac{4\sqrt{d+1}((\beta C_1+Rd)C_2+C_3)}{\sqrt{n}}+2M(\beta C_1+1+Rd)\sqrt{\frac{(d+1)\log((d+1)/\delta)}{2n}}+\inf_{f\in\mathcal{F}} SD_{\mathcal{H}}(\nu,\nu_f).$$

Proof. Notice that by the definition of the Stein operator,

$$\begin{aligned}\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x)) &= \mathrm{Tr}\left(\left(\nabla\log\left(\frac{d\nu_f}{d\tau}(x)\right)-dx\right)h(x)^\top+\nabla h(x)\right)=\mathrm{Tr}\left(-(\nabla f(x)+dx)h(x)^\top+\nabla h(x)\right)\\ &=\sum_{i=1}^{d+1}-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)\end{aligned}$$

Thus,

$$\begin{aligned}&\sup_{h\in\mathcal{H}}\mathbb{E}_\nu[\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x))]-\mathbb{E}_{\nu_n}[\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x))]\\ &=\sup_{h\in\mathcal{H}}\sum_{i=1}^{d+1}(\mathbb{E}_\nu[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)]-\mathbb{E}_{\nu_n}[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)])\\ &=\sup_{\substack{\sum_i|w_i|^2\le 1\\ h_i\in\mathcal{B}_{\mathcal{H}_i}}}\sum_{i=1}^{d+1}w_i(\mathbb{E}_\nu[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)]-\mathbb{E}_{\nu_n}[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)])\\ &=\sqrt{\sum_{i=1}^{d+1}\left(\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}(\mathbb{E}_\nu[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)]-\mathbb{E}_{\nu_n}[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)])\right)^2}\\ &=\sqrt{\sum_{i=1}^{d+1}\Phi_i(S_n)^2},\end{aligned}$$

where $\Phi_i(S_n)=\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}(\mathbb{E}_\nu[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)]-\mathbb{E}_{\nu_n}[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)])$. For a fixed i, we can use a classical argument based on McDiarmid's inequality (c.f. Mohri et al. [2012], Theorem 3.3) to obtain

$$\mathbb{P}\left(\Phi_i(S_n)-\mathbb{E}_{S_n'}\left[\Phi_i(S_n')\right]\ge\epsilon\right)\le\exp\left(\frac{-2\epsilon^2 n}{C_4^2}\right),$$

where $C_4=M(\beta C_1+1+Rd)$ is a uniform upper-bound on $\{\|-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)\|_\infty \mid h_i\in\mathcal{B}_{\mathcal{H}_i}\}$. Thus, using a union bound, we obtain that

$$\mathbb{P}\left(\max_{1\le i\le d+1}\left(\Phi_i(S_n)-\mathbb{E}_{S_n'}\left[\Phi_i(S_n')\right]\right)\ge\epsilon\right)\le(d+1)\exp\left(\frac{-2\epsilon^2 n}{C^2}\right),$$

and through a change of variables, that means that with probability at least $1-\delta$,

$$\begin{aligned}&\max_{1\le i\le d+1}\left(\Phi_i(S_n)-\mathbb{E}_{S_n'}\left[\Phi_i(S_n')\right]\right)\le C_4\sqrt{\frac{\log((d+1)/\delta)}{2n}}\\ &\implies \max_{1\le i\le d+1}\Phi_i(S_n)\le\max_{1\le i\le d+1}\mathbb{E}_{S_n'}\Phi_i(S_n')+C_4\sqrt{\frac{\log((d+1)/\delta)}{2n}}\\ &\implies \sqrt{\sum_{i=1}^{d+1}\Phi_i(S_n)^2}\le\sqrt{d+1}\max_{1\le i\le d+1}\Phi_i(S_n)\le\sqrt{d+1}\max_{1\le i\le d+1}\mathbb{E}_{S_n'}\Phi_i(S_n')+C_4\sqrt{\frac{(d+1)\log((d+1)/\delta)}{2n}}\end{aligned}$$

All that is left is to upper-bound $\mathbb{E}_{S_n}\Phi_i(S_n)$ for any i using Rademacher complexity bounds:

$$\begin{aligned}
&\mathbb{E}_{S_n}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\left(\mathbb{E}_\nu[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)]-\mathbb{E}_{\nu_n}[-(\nabla_i f(x)+dx_i)h_i(x)+\nabla_i h_i(x)]\right)\right]\\
&\leq \mathbb{E}_{S_n,S_n'}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n -((\nabla_i f(x_j')+dx_{j,i}')h_i(x_j')-(\nabla_i f(x_j)+dx_{j,i})h_i(x_j))+\nabla_i h_i(x_j')-\nabla_i h_i(x_j)\right]\\
&= \mathbb{E}_{\sigma,S_n,S_n'}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n \sigma_j\left(-((\nabla_i f(x_j')+dx_{j,i}')h_i(x_j')-(\nabla_i f(x_j)+dx_{j,i})h_i(x_j))+\nabla_i h_i(x_j')-\nabla_i h_i(x_j)\right)\right],
\end{aligned}$$

and this is upper-bounded by

$$\begin{aligned}
&2\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n \sigma_j\left(-(\nabla_i f(x_j)+dx_{j,i})h_i(x_j)+\nabla_i h_i(x_j)\right)\right]\\
&\leq 2\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n \sigma_j(\nabla_i f(x_j)+dx_{j,i})h_i(x_j)\right]+2\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n \sigma_j\nabla_i h_i(x_j)\right]
\end{aligned}\tag{20}$$

By Talagrand's Lemma (Mohri et al. [2012], Theorem 5.7) and the uniform L^∞ bound on $\{\nabla_i f | f\in\mathcal{F}\}$ (notice that $y\mapsto(\beta\nabla_i f(x_j)+dx_{j,i})y$ has Lipschitz constant uniformly upper-bounded by $\|\beta\nabla_i f(x_j)+dx_{j,i}\|_\infty$, which means that the assumptions of Talagrand's Lemma are fulfilled), we have

$$\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n \sigma_j(\nabla_i f(x_j)+dx_{j,i})h_i(x_j)\right]\leq(C_1\beta+Rd)\mathbb{E}_{\sigma,S_n}\left[\sup_{h_i\in\mathcal{B}_{\mathcal{H}_i}}\frac{1}{n}\sum_{j=1}^n \sigma_j h_i(x_j)\right]\leq\frac{(\beta C_1+Rd)C_2}{\sqrt{n}},$$

where we used the Rademacher complexity bound of $\mathcal{B}_{\mathcal{H}_i}$. Using the Rademacher complexity bound of $\nabla_i h_i$ as well, we conclude that the right-hand side of (20) can be upper-bounded by $\frac{2(\beta C_1+Rd)C_2+2C_3}{\sqrt{n}}$. Thus, with probability at least $1-\delta$, for all $h\in\mathcal{H}$,

$$\left|\mathbb{E}_\nu[\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x))]-\mathbb{E}_{\nu_n}[\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x))]\right|\leq\frac{2\sqrt{d+1}((\beta C_1+Rd)C_2+C_3)}{\sqrt{n}}+C_4\sqrt{\frac{(d+1)\log((d+1)/\delta)}{2n}}\tag{21}$$

We conclude the proof with an argument similar to the one of Theorem 1:

$$\begin{aligned}
&\mathrm{SD}_{\mathcal{H}}(\nu,\hat{\nu})\\
&=\mathrm{SD}_{\mathcal{H}}(\nu,\hat{\nu})-\inf_{f\in\mathcal{F}}\mathrm{SD}_{\mathcal{H}}(\nu,\nu_f)+\inf_{f\in\mathcal{F}}\mathrm{SD}_{\mathcal{H}}(\nu,\nu_f)\\
&=\sup_{f\in\mathcal{F}}\left\{\sup_{h\in\mathcal{H}}\mathbb{E}_\nu[\mathrm{Tr}(\mathcal{A}_{\hat{\nu}}h(x))]-\sup_{h\in\mathcal{H}}\mathbb{E}_\nu[\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x))]\right\}+\inf_{f\in\mathcal{F}}\mathrm{SD}_{\mathcal{H}}(\nu,\nu_f)\\
&\leq\sup_{f\in\mathcal{F}}\left\{\sup_{h\in\mathcal{H}}\mathbb{E}_{\nu_n}[\mathrm{Tr}(\mathcal{A}_{\hat{\nu}}h(x))]-\sup_{h\in\mathcal{H}}\mathbb{E}_{\nu_n}[\mathrm{Tr}(\mathcal{A}_{\nu_f}h(x))]\right\}+\frac{4\sqrt{d+1}((\beta C_1+Rd)C_2+C_3)}{\sqrt{n}}\\
&+2C_4\sqrt{\frac{\log((d+1)/\delta)}{2n}}+\inf_{f\in\mathcal{F}}\mathrm{SD}_{\mathcal{H}}(\nu,\nu_f)\\
&=\frac{4\sqrt{d+1}((\beta C_1+Rd)C_2+C_3)}{\sqrt{n}}+2M(\beta C_1+1+Rd)\sqrt{\frac{(d+1)\log((d+1)/\delta)}{2n}}+\inf_{f\in\mathcal{F}}\mathrm{SD}_{\mathcal{H}}(\nu,\nu_f).
\end{aligned}$$

In the second equality we use the definition of the Stein discrepancy (equation (2)). The inequality follows from (21) applied on ν_f and on $\hat{\nu}=\nu_{\hat{f}}$. The last equality holds because of the definition of $\hat{f}$ and the definition of C_4. □

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

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

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

Proof. Note that Lemma 5 provides the expression for the Stein operator $\mathcal{A}_\nu$ on $\mathbb{S}^d$ and shows that for any $\nu \in \mathcal{P}(\mathbb{S}^d)$ with continuous and a.e. differentiable density, the class of continuous and a.e. differentiable functions $\mathbb{S}^d \to \mathbb{R}^{d+1}$ is contained in the Stein class of ν (which by definition is the set of functions h such that the Stein identity $\mathbb{E}_\nu[A_\nu h] = 0$ holds). Using the argument of Lemma 2.3 of Liu et al. [2016], we have that for any $\nu_1, \nu_2 \in \mathcal{P}(K)$, for any h in the Stein class of ν_1 we have

$$\begin{aligned} &\mathbb{E}_{\nu_1}[\mathcal{A}_{\nu_2} h(x)] = \mathbb{E}_{\nu_1}[\mathcal{A}_{\nu_2} h(x) - \mathcal{A}_{\nu_1} h(x)] \\ &= \mathbb{E}_{\nu_1}[s_{\nu_2}(x)h(x)^\top + \nabla h(x) - dx h(x)^\top - (s_{\nu_1}(x)h(x)^\top + \nabla h(x) - dx h(x)^\top)] = \mathbb{E}_{\nu_1}[(s_{\nu_2}(x) - s_{\nu_1}(x))h(x)^\top], \end{aligned}$$

which follows from the definition of the Stein operator and the Stein identity: $\mathbb{E}_{\nu_1}[\mathcal{A}_{\nu_1} h(x)] = 0$. Thus, for any $\nu \in \mathcal{P}(K)$,

$$\begin{aligned} \mathrm{SD}_{\mathcal{B}_{\mathcal{F}_1^{d+1}}}(\nu, \nu_f) &= \sup_{h\in\mathcal{B}_{\mathcal{F}_1^{d+1}}} \mathbb{E}_\nu[\mathrm{Tr}((s_{\nu_f}(x) - s_\nu(x))h(x)^\top)] \\ &= \sup_{h\in\mathcal{B}_{\mathcal{F}_1^{d+1}}} \sum_{i=1}^{d+1} \mathbb{E}_\nu\left[\left(-\nabla_i f(x) - \nabla_i \log\left(\frac{d\nu}{d\tau}(x)\right)\right) h_i(x)\right] \\ &= \sup_{\substack{\sum_i |w_i|^2 \le 1, \\ |\gamma_i|_{\mathrm{TV}} \le 1}} \sum_{i=1}^{d+1} \mathbb{E}_\nu\left[\left(-\nabla_i f(x) - \nabla_i \log\left(\frac{d\nu}{d\tau}(x)\right)\right) w_i \int_{\mathbb{S}_d} \sigma(\langle\theta, x\rangle) d\gamma_i(\theta)\right] \\ &\le \mathbb{E}_\nu\left[\sup_{\substack{\sum_i |w_i|^2 \le 1, \\ |\gamma_i|_{\mathrm{TV}} \le 1}} \sum_{i=1}^{d+1} w_i\left(- \nabla_i f(x) - \nabla_i \log\left(\frac{d\nu}{d\tau}(x)\right)\right) \int_{\mathbb{S}_d} \sigma(\langle\theta, x\rangle) d\gamma_i(\theta)\right] \\ &= \mathbb{E}_\nu\left[\sup_{\substack{\sum_i |w_i|^2 \le 1, \\ \{\theta^{(i)}\}\subset\mathbb{S}^d}} \sum_{i=1}^{d+1} w_i\left(- \nabla_i f(x) - \nabla_i \log\left(\frac{d\nu}{d\tau}(x)\right)\right)\sigma(\langle\theta^{(i)}, x\rangle)\right] \\ &= \mathbb{E}_\nu\left[\left(\sum_{i=1}^{d+1} \sup_{\{\theta^{(i)}\}\subset\mathbb{S}^d} \left(\left(- \nabla_i f(x) - \nabla_i \log\left(\frac{d\nu}{d\tau}(x)\right)\right)\sigma(\langle\theta^{(i)}, x\rangle)\right)^2\right)^{1/2}\right] \\ &= \mathbb{E}_\nu\left[\left(\sum_{i=1}^{d+1} \left(- \nabla_i f(x) - \nabla_i \log\left(\frac{d\nu}{d\tau}(x)\right)\right)^2\right)^{1/2}\right] = \mathbb{E}_\nu\left[\left\| - \nabla f(x) - \nabla \log\left(\frac{d\nu}{d\tau}(x)\right)\right\|_2\right] \end{aligned} \tag{22}$$

Moreover, by Lemma 4:

$$\mathbb{E}_{\sigma,S_n}\left[\sup_{h\in\mathcal{B}_{\mathcal{F}_1}}\frac{1}{n}\sum_{j=1}^{n}\sigma_j h(x_j)\right]=\mathcal{R}_n(\mathcal{B}_{\mathcal{F}_1})\leq\frac{1}{\sqrt{n}}. \tag{23}$$

And

$$\begin{aligned}
&\mathbb{E}_{\sigma,S_n}\left[\sup_{h\in\mathcal{B}_{\mathcal{F}_1}}\frac{1}{n}\sum_{j=1}^{n}\sigma_j\nabla_i h(x_j)\right]=\mathbb{E}_{\sigma,S_n}\left[\sup_{|\gamma|_{\mathrm{TV}}\leq 1}\frac{1}{n}\sum_{j=1}^{n}\sigma_j\int_{\mathbb{S}^d}\nabla_i\sigma(\langle\theta,x_j\rangle)d\gamma(\theta)\right]\\
&=\mathbb{E}_{\sigma,S_n}\left[\sup_{\theta\in\mathbb{S}^d,|w|\leq 1}\frac{w}{n}\sum_{j=1}^{n}\sigma_j\mathbb{1}_{\langle\theta,x_j\rangle\geq 0}\theta_i\right]=\mathbb{E}_{\sigma,S_n}\left[\sup_{\theta\in\mathbb{S}^d}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\mathbb{1}_{\langle\theta,x_j\rangle\geq 0}\theta_i\right|\right]\\
&\leq\mathbb{E}_{\sigma,S_n}\left[\sup_{\theta\in\mathbb{S}^d}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\mathbb{1}_{\langle\theta,x_j\rangle\geq 0}\right|\right]\leq C_2\frac{\sqrt{d+1}}{\sqrt{n}},
\end{aligned} \tag{24}$$

where the last inequality follows from the Rademacher complexity bound on the hyperplane hypothesis, which is obtained through a VC dimension argument (Bach [2017a], Section 5.1; Bartlett and Mendelson [2002], Theorem 6). Moreover, $\|h\|_\infty\leq 1$ and $\|\nabla_i h\|_\infty\leq 1$ for all $h\in\mathcal{F}_1$. The proof concludes by plugging (22), (23), (24) into Theorem 4. Since $\mathcal{B}_{\mathcal{F}_2^{d+1}}\subset\mathcal{B}_{\mathcal{F}_1^{d+1}}$, all the upper-bounds of the proof hold for $\mathcal{H}=\mathcal{B}_{\mathcal{F}_2^{d+1}}$ as well. □

Theorem 5. *Let $K=\mathbb{S}^d$. Let KSD be the kernelized Stein discrepancy for a positive definite kernel k with continuous second order partial derivatives, such that for any non-zero function $g\in L^2(\mathbb{S}^d)$, $\int_{\mathbb{S}^d}\int_{\mathbb{S}^d}g(x)k(x,x')g(x')d\tau(x)d\tau(x')>0$. 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 $1-\delta$,*

$$\begin{aligned}
KSD(\nu,\hat{\nu})\leq\frac{2}{\sqrt{\delta n}}\sup_{f\in\mathcal{F}}\left(Var_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[\tilde{u}_{\nu_f}(x,x')])\right)^{1/2}\\
+\sqrt{\mathbb{E}_{x,x'\sim\nu}[k(x,x')^2]}\inf_{f\in\mathcal{F}}\mathbb{E}_{x\sim\nu}\left[\left\|\nabla\log\left(\frac{d\nu}{d\tau}(x)\right)-\beta\nabla f(x)\right\|^2\right]
\end{aligned}$$

Proof. For the kernelized Stein discrepancy estimator we can write

$$\begin{aligned}
&KSD(\nu,\hat{\nu})\\
&=KSD(\nu,\hat{\nu})-\inf_{f\in\mathcal{F}}KSD(\nu,\nu_f)+\inf_{f\in\mathcal{F}}KSD(\nu,\nu_f)\\
&=\sup_{f\in\mathcal{F}}\left\{\mathbb{E}_{x,x'\sim\nu}[u_{\hat{\nu}}(x,x')]-\mathbb{E}_{x,x'\sim\nu}[u_{\nu_f}(x,x')]\right\}+\inf_{f\in\mathcal{F}}KSD(\nu,\nu_f)\\
&=\sup_{f\in\mathcal{F}}\left\{\mathbb{E}_{x,x'\sim\nu}[\tilde{u}_{\hat{\nu}}(x,x')]-\mathbb{E}_{x,x'\sim\nu}[\tilde{u}_{\nu_f}(x,x')]\right\}+\inf_{f\in\mathcal{F}}KSD(\nu,\nu_f)\\
&\leq\sup_{f\in\mathcal{F}}\left\{\frac{1}{n(n-1)}\left(\sum_{i\neq j}\tilde{u}_{\hat{\nu}}(x_i,x_j)-\sum_{i\neq j}\tilde{u}_{\nu_f}(x_i,x_j)\right)\right\}+\frac{2}{\sqrt{\delta n}}\sup_{f\in\mathcal{F}}(\mathrm{Var}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[u_{\nu_f}(x,x')]))^{1/2}\\
&+\inf_{f\in\mathcal{F}}KSD(\nu,\nu_f)=\frac{2}{\sqrt{\delta n}}\sup_{f\in\mathcal{F}}(\mathrm{Var}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[u_{\nu_f}(x,x')]))^{1/2}+\inf_{f\in\mathcal{F}}KSD(\nu,\nu_f)
\end{aligned}$$

The third equality holds because of the definition of $\tilde{u}_\nu$ in terms of u_ν. In the first inequality we have used that for any $\tilde{\nu}$ (different from ν) with almost-everywhere differentiable log-density, $\frac{1}{n(n-1)}\sum_{i\neq j}\tilde{u}_{\tilde{\nu}}(x_i,x_j)$

has expectation $\mathbb{E}_{x,x'\sim\nu}[\tilde{u}_{\tilde{\nu}}(x,x')]$ and variance $\mathrm{Var}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[\tilde{u}_{\tilde{\nu}}(x,x')])/n$ by the theory of U-statistics (Liu et al. [2016], Theorem 4.1; Serfling [2009], Section 5.5). Thus, by Chebyshev's inequality, with probability at least $1-\delta$, we have that $\mathbb{E}_{x,x'\sim\nu}[\tilde{u}_{\tilde{\nu}}(x,x')] \leq \frac{1}{n(n-1)}\sum_{i\neq j}\tilde{u}_{\tilde{\nu}}(x_i,x_j) + \frac{1}{\sqrt{n\delta}}(\mathrm{Var}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[\tilde{u}_{\tilde{\nu}}(x,x')]))^{1/2}$.

Moreover, using the argument of Theorem 5.1 of Liu et al. [2016], by Lemma 6,

$$\begin{aligned}
KSD(\nu,\nu_f) &= \mathbb{E}_{x,x'\sim\nu}[(s_\nu(x)-s_{\nu_f}(x))^\top(s_\nu(x')-s_{\nu_f}(x'))k(x,x')] \\
&\leq \sqrt{\mathbb{E}_{x,x'\sim\nu}[k(x,x')^2]}\sqrt{\mathbb{E}_{x,x'\sim\nu}\left[\left((s_\nu(x)-s_{\nu_f}(x))^\top(s_\nu(x')-s_{\nu_f}(x'))\right)^2\right]} \\
&\leq \sqrt{\mathbb{E}_{x,x'\sim\nu}[k(x,x')^2]}\sqrt{\mathbb{E}_{x,x'\sim\nu}\left[\|s_\nu(x)-s_{\nu_f}(x)\|^2\|s_\nu(x')-s_{\nu_f}(x')\|^2\right]} \\
&= \sqrt{\mathbb{E}_{x,x'\sim\nu}[k(x,x')^2]}\ \mathbb{E}_{x\sim\nu}\left[\|s_\nu(x)-s_{\nu_f}(x)\|^2\right],
\end{aligned}$$

where $\mathbb{E}_{x\sim\nu}\left[\|s_\nu(x)-s_{\nu_f}(x)\|^2\right]$ is known as the Fisher divergence. □

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 $1-\delta$,*

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

Proof. We apply Theorem 5. We can bound

$$\begin{aligned}
\sup_{f\in\mathcal{F}}(\mathrm{Var}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[u_{\nu_f}(x,x')]))^{1/2} &\leq \sup_{f\in\mathcal{F}}(\mathbb{E}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[u_{\nu_f}(x,x')])^2)^{1/2} \leq \sup_{f\in\mathcal{F}}(\mathbb{E}_{x\sim\nu}(\mathbb{E}_{x'\sim\nu}[u_{\nu_f}(x,x')^2]))^{1/2} \\
&\leq \sup_{f\in\mathcal{F}}\sup_{x,x'\in\mathbb{S}^d}|u_{\nu_f}(x,x')| \leq ((\beta C_1+d)^2C_2+2C_3(\beta C_1+d)),
\end{aligned}$$

and $\sqrt{\mathbb{E}_{x,x'\sim\nu}[k(x,x')^2]}\leq C_2$. □

Lemma 7. *For a function $g:\mathbb{S}^d\to\mathbb{R}$, we define the partial derivative $\partial_i g:\mathbb{S}^d\to\mathbb{R}$ as the restriction to $\mathbb{S}^d$ of the partial derivative of the polynomial power series extension of g to $\mathbb{R}^n$ (i.e. the extension of a spherical harmonic to $\mathbb{R}^n$ is the polynomial whose restriction to $\mathbb{S}^d$ is equal to the spherical harmonic (Atkinson and Han [2012], Definition 2.7)). We denote by $\partial g=(\partial_i g)_{i=1}^{d+1}$ the vector of partial derivatives of g. The Riemannian gradient $\nabla g:\mathbb{S}^d\to\mathbb{R}^{d+1}$, which is intrinsic (does not depend on the extension chosen), fulfills*

$$\nabla g(x) = (\nabla_i g(x))_{i=1}^{d+1} := \left(\partial_i g(x) - \sum_{i=1}^{d+1}\partial_j g(x)x_jx_i\right)_{i=1}^{d+1}.$$

That is, $\nabla g(x)$ is the projection of $\partial g(x)$ to the tangent space of $\mathbb{S}^d$ at x.

For δ greater than a constant depending only on d, for any function $g:\mathbb{S}^d\to\mathbb{R}$ such that for all $x,y\in\mathbb{S}^d$ we have $|g(x)|\leq\eta$ and $|g(x)-g(y)|\leq\eta\|x-y\|_2$, and $\|\nabla g(x)\|_2\leq\eta$ and $\|\nabla g(x)-\nabla g(y)\|_2\leq L\|x-y\|_2$, and

g is even, there exists $\hat{g} \in \mathcal{F}_2$ such that $\|\hat{g}\|_{\mathcal{F}_2} \leq \delta$ and

$$\sup_{x \in \mathbb{S}^d} |\hat{g}(x) - g(x)| \leq C(d)\eta \left(\frac{\delta}{\eta}\right)^{-2/(d+1)} \log\left(\frac{\delta}{\eta}\right), \tag{25}$$
$$\sup_{x \in \mathbb{S}^d} \|\nabla \hat{g}(x) - \nabla g(x)\|_2 \leq C(d)(L+\eta) \left(\frac{\delta}{\eta}\right)^{-2/(d+1)} \log\left(\frac{\delta}{\eta}\right),$$

where $C(d)$ are constants depending only on the dimension d.

Proof. We will use some ideas and notation of the proof of Prop. 3 of Bach [2017a]. We can decompose $g(x) = \sum_{k \geq 0} g_k(x)$, where $g_k(x) = N(d,k) \int_{\mathbb{S}^d} g(y) P_k(\langle x, y \rangle) d\tau(y)$. g_k is the k-th spherical harmonic of g and P_k is the k-th Legendre polynomial in dimension $d+1$. Analogously, for any i between 1 and $d+1$ we can decompose $\nabla_i g(x) = \sum_{k \geq 0} (\nabla_i g)_k(x)$, where $(\nabla_i g)_k(x) = N(d,k) \int_{\mathbb{S}^d} \nabla_i g(y) P_k(\langle x, y \rangle) d\tau(y)$. Define $\widetilde{\nabla_i g} : \mathbb{R}^{d+1} \to \mathbb{R}$ to be the spherical harmonic extension of $\nabla_i g$.

Like Bach [2017a], we define $\hat{g}(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle) \hat{h}(\theta) d\tau(\theta)$, where $\hat{h}(x) = \sum_{k, \lambda_k \neq 0} \lambda_k^{-1} r^k g_k(x)$ for some $r \in (0,1)$. Equivalently, $\hat{g}(x) = \sum_{k, \lambda_k \neq 0} r^k g_k(x)$. Since g_k is a homogeneous polynomial of degree k (Atkinson and Han [2012], Definition 2.7), we have that $\hat{g}(x) = \sum_{k, \lambda_k \neq 0} g_k(rx) = g(rx)$.

With this choice of $\hat{g}$, the first equation of (25) holds by Prop. 3 of Bach [2017a].

Using this characterization of $\hat{g}$, by the chain rule we compute the Riemannian gradient

$$\nabla \hat{g}(x) = \partial \hat{g}(x) - \langle \partial \hat{g}(x), x \rangle x = \partial (g \circ (y \to ry))(x) - \langle \partial (g \circ (y \to ry))(x), x \rangle x = r \partial g(rx) - r \langle \partial g(rx), x \rangle x$$

The polynomial power series extension $\widetilde{\nabla g}$ of ∇g is by definition equal to $\nabla g(x) = \partial g(x) - \langle \partial g(x), x \rangle x = \sum_{k \geq 0} (\partial g)_k(x) - \langle (\partial g)_k(x), x \rangle x$ for $x \in \mathbb{S}^d$. Since the terms of $\sum_{k \geq 0} (\partial g)_k(x) - \langle (\partial g)_k(x), x \rangle x$ are polynomials on x, this expression is equal to the polynomial power series of $\widetilde{\nabla g}$ by uniqueness of the polynomial power series. Thus, for all $x \in \mathbb{R}^{d+1}$,

$$\widetilde{\nabla g}(x) = \sum_{k \geq 0} (\partial g)_k(x) - \langle (\partial g)_k(x), x \rangle x = \sum_{k \geq 0} \partial (g_k)(x) - \langle \partial (g_k)(x), x \rangle x = \partial g(x) - \langle \partial g(x), x \rangle x. \tag{26}$$

The second equality follows from Lemma 8, which states that $\partial(g_k) = (\partial g)_k$. Hence, by (26), we have $r \partial g(rx) - r \langle \partial g(rx), rx \rangle rx = r \widetilde{\nabla g}(rx) = r \sum_{k \geq 0} (\nabla g)_k(rx) = r \sum_{k \geq 0} r^k (\nabla g)_k(x)$. Thus, in analogy with Bach [2017a], we have

$$r \partial g(rx) - r \langle \partial g(rx), rx \rangle rx = r \sum_{k \geq 0} r^k (\nabla g)_k(x) = r \sum_{k \geq 0} r^k N(d,k) \int_{\mathbb{S}^d} \nabla g(y) P_k(\langle x, y \rangle) d\tau(y)$$
$$= r \int_{\mathbb{S}^d} \nabla g(y) \left(\sum_{k \geq 0} r^k N(d,k) P_k(\langle x, y \rangle) \right) d\tau(y) = r \int_{\mathbb{S}^d} \nabla g(y) \frac{1 - r^2}{(1 + r^2 - 2r(\langle x, y \rangle))^{(d+1)/2}} d\tau(y).$$

Hence, keeping the analogy with Bach [2017a] (and Bourgain and Lindenstrauss [1988], Equation 2.13), we

obtain that

$$\left\|\nabla g(x) - r\partial g(rx) - r\langle \partial g(rx), rx\rangle rx\right\|_2 = \left\|\int_{\mathbb{S}^d} (\nabla g(x) - r\nabla g(y)) \frac{1-r^2}{(1+r^2-2r(\langle x,y\rangle))^{(d+1)/2}} d\tau(y)\right\|_2$$

$$\leq \int_{\mathbb{S}^d} \|\nabla g(x) - r\nabla g(y)\|_2 \frac{1-r^2}{(1+r^2-2r(\langle x,y\rangle))^{(d+1)/2}} d\tau(y)$$

$$\leq \int_{\mathbb{S}^d} \|\nabla g(x) - \nabla g(y)\|_2 \frac{1-r^2}{(1+r^2-2r(\langle x,y\rangle))^{(d+1)/2}} d\tau(y) + (1-r)\int_{\mathbb{S}^d} \|\nabla g(y)\|_2 \frac{1-r^2}{(1+r^2-2r(\langle x,y\rangle))^{(d+1)/2}} d\tau(y)$$

$$\leq C_2(d)(1-r)\mathrm{Lip}(\nabla g)\int_0^1 \frac{t^d}{(1-r)^{d+1}+t^{d+1}} dt + (1-r)\int_{\mathbb{S}^d} \left(\sup_{x\in\mathbb{S}^d} \|\nabla g(x)\|_2\right)\left(\sum_{k\geq 0} r^k N(d,k) P_k(\langle x,y\rangle)\right) d\tau(y)$$

$$\leq C_3(d)\mathrm{Lip}(\nabla g)(1-r)\log\big(1/(1-r)\big) + (1-r)\left(\sup_{x\in\mathbb{S}^d} \|\nabla g(x)\|_2\right) \leq C_4(d)(1-r)(\eta + L\log(1/(1-r)))$$

In the last equality we have used that $\partial_i g$ is L-Lipschitz by assumption. And

$$\begin{aligned}\|\nabla g(x) - \nabla \hat{g}(x)\|_2^2 &= \|\nabla g(x) - r\partial g(rx) - r\langle \partial g(rx), x\rangle x\|_2^2 \\ &\leq \|\nabla g(x) - r\partial g(rx) - r\langle \partial g(rx), x\rangle x\|_2^2 + \|(1-r^2) r\langle \partial g(rx), x\rangle x\|_2^2 \qquad (27)\\ &\leq \|\nabla g(x) - r\partial g(rx) - r\langle \partial g(rx), rx\rangle rx\|_1^2 \leq (C_4(d)(1-r)(\eta + L\log(1/(1-r))))^2.\end{aligned}$$

In the second equality we have used that $\nabla g(x) - \nabla \hat{g}(x)$ is orthogonal to x (because it belongs to the tangent space at x), and the Pythagorean theorem. As in Bach [2017a], for $\delta > 0$ large enough the argument is concluded by taking $1 - r = (C_1(d)\eta/\delta)^{2/(d+1)} \in (0,1)$, which means that the (square root of the) error in the right-hand side of (27) is $C_4(d)(C_1(d)\eta/\delta)^{2/(d+1)}\left(\eta + L\log(C_1(d)\eta/\delta)^{-2/(d+1)}\right) \leq C_5(d)(L+\eta)(\delta/\eta)^{-2/(d+1)}\log(\delta/\eta)$.

Using that g is η-Lipschitz, by the argument of Bach [2017a] we have that $\|\hat{h}\|_{L^2(\mathbb{S}^d)} \leq C_1(d)\eta(1-r)^{(-d-1)/2}$, where $C_1(d)$ is a constant that depends only on d and consequently $\|\hat{g}\|_{\mathcal{F}_2} \leq C_1(d)\eta(1-r)^{(-d-1)/2}$. And for our choice of r, this bound becomes $\|\hat{g}\|_{\mathcal{F}_2} \leq C_1(d)\eta((C_1(d)\eta/\delta)^{2/(d+1)})^{(-d-1)/2} = \delta$.

□

Lemma 8. *For $g : \mathbb{S}^d \to \mathbb{R}$ with spherical harmonic decomposition $g(x) = \sum_{k\geq 0} g_k(x)$ and with partial derivative with spherical harmonic decomposition $\partial_i g(x) = \sum_{k\geq 0} (\partial_i g)_k(x)$, we have $(\partial_i g)_k(x) = \partial_i(g_k)(x)$.*

Proof. Remark that the spherical harmonics on $\mathbb{S}^d$ can be characterized as the restrictions of the homogeneous harmonic polynomials on $\mathbb{R}^{d+1}$ (Atkinson and Han [2012], Definition 2.7). k-th degree homogeneous polynomials are of sums of monomials of the form $\alpha_{i_1,\dots,i_r} x_1^{i_1}\cdot\dots\cdot x_r^{i_r}$, where $\sum_{l=1}^r i_l = k$, and harmonic polynomials are those such that $\Delta p = \sum_{i=1}^{d+1} \frac{\partial^2 p}{\partial x_i^2} = 0$. Thus, for all $k \geq 0$, g_k can be seen as the restrictions to $\mathbb{S}^d$ of homogeneous harmonic polynomials of degree k.

Notice that the i-th partial derivative of a homogeneous harmonic polynomial p of degree k is a homogeneous harmonic polynomial of degree $k-1$. That is because by commutation of partial derivatives, we have

$$\Delta(\partial_i p) = \sum_{j=1}^{d+1} \partial_{jj}\partial_i p = \sum_{j=1}^{d+1} \partial_i \partial_{jj} p = \partial_i(\Delta p) = 0.$$

Thus, $\partial_i(g_k)$ are homogeneous harmonic polynomials of degree $k-1$, which means that their restrictions to $\mathbb{S}^d$ are spherical harmonics. Since $\partial_i g(x) = \partial_i(\sum_{k\geq 0} g_k(x)) = \sum_{k\geq 0} \partial_i(g_k)(x)$ and the spherical harmonic decomposition is unique, $\partial_i(g_k)$ must be precisely the spherical harmonic components of $\partial_i g$. □

Corollary 2. *Let $\mathcal{F} = B_{\mathcal{F}_1}(\beta)$. 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 $\mathrm{KSD}(\nu, \hat{\nu})$ is upper-bounded by

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

Proof. We will use Theorem 2. Let $g : \mathbb{S}^d \to \mathbb{R}$ be defined as $g(x) = \sum_{j=1}^J g_j(x) = \sum_{j=1}^J \varphi_j(U_j x)$, where $\varphi_j : \{x \in \mathbb{R}^{k+1} | \|x\|_2 \leq 1\} \to \mathbb{R}$.

Let $\hat{\varphi}_j : \mathbb{S}^k \to \mathbb{R}$ be the restriction of φ_j to $\mathbb{S}^k$. By Lemma 7, there exists $\hat{\psi}_j : \mathbb{S}^k \to \mathbb{R}$ such that $\hat{\psi}_j \in \mathcal{F}_2$ and $\|\hat{\psi}_j\|_{\mathcal{F}_2} \leq \beta/J$, and

$$\begin{aligned}
\sup_{x\in\mathbb{S}^d} |\hat{\varphi}_j(x) - \hat{\psi}_j(x)| &\leq C(k)(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{\beta}{\eta J}\right), \\
\sup_{x\in\mathbb{S}^d} \|\nabla\hat{\varphi}_j(x) - \nabla\hat{\psi}_j(x)\|_2 &\leq C(k)(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{\beta}{\eta J}\right)
\end{aligned} \tag{28}$$

Moreover, if we denote by $\psi_j : \{x \in \mathbb{R}^{k+1} | \|x\|_2 \leq 1\} \to \mathbb{R}$ the 1-homogeneous extension of $\hat{\psi}_j$, we can write the (Euclidean) gradient of ψ_j at the point rx (with $r \in [0,1]$, $x \in \mathbb{S}^d$) in terms of the (Riemannian) gradient of $\hat{\psi}_j$ at x:

$$\nabla\psi_j(rx) = r\nabla\hat{\psi}_j(x) + \hat{\psi}_j(x)$$

Thus, by Equation 28, and renaming $C(k)$,

$$\begin{aligned}
\sup_{\|x\|_2\leq 1} \|\nabla\varphi_j(x) - \nabla\psi_j(x)\|_2 &\leq \sup_{x\in\mathbb{S}^d} \|\nabla\hat{\varphi}_j(x) - \nabla\hat{\psi}_j(x)\|_2 + \sup_{x\in\mathbb{S}^d} |\hat{\varphi}_j(x) - \hat{\psi}_j(x)| \\
&\leq C(k)(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{\beta}{\eta J}\right)
\end{aligned}$$

Hence, if we define $\tilde{g}_j : \mathbb{S}^d \to \mathbb{R}$ as $\tilde{g}_j(x) := \psi_j(U_j x)$, we check that $\tilde{g}_j$ belongs to $\mathcal{F}_1$: if $\hat{\psi}_j$ is such that $\forall x \in \mathbb{S}^d$, $\hat{\psi}_j(x) = \int_{\mathbb{S}^k} \sigma(\langle\theta, x\rangle) d\gamma(\theta)$, then $\psi_j(x) = \int_{\mathbb{S}^k} \sigma(\langle\theta, x\rangle) d\gamma(\theta)$ when $\|x\|_2 \leq 1$, and

$$\tilde{g}_j(x) = \psi_j(U_j x) = \int_{\mathbb{S}^k} \sigma(\langle\theta, U_j x\rangle) d\gamma(\theta) = \int_{\mathbb{S}^k} \sigma(\langle U_j^\top\theta, x\rangle) d\gamma(\theta) = \int_{\mathbb{S}^d} \sigma(\langle\theta', x\rangle) d\gamma'(\theta')$$

This also shows that $\tilde{g}_j$ has $\mathcal{F}_1$ norm $\|\tilde{g}_j\|_{\mathcal{F}_1} \leq \|\hat{\psi}_j\|_{\mathcal{F}_2} \leq \beta/J$, which would mean that $\tilde{g} = \sum_{j=1}^J \tilde{g}_j \in \mathcal{F}_1$ and $\|\tilde{g}\|_{\mathcal{F}_1} \leq \beta$. Moreover,

$$\begin{aligned}
\sup_{x\in\mathbb{S}^d} \|\nabla\tilde{g}_j(x) - \nabla g_j(x)\|_2 &= \sup_{x\in\mathbb{S}^d} \|\nabla(\psi_j \circ U_j)(x) - \nabla(\varphi_j \circ U_j)(x)\|_2 \\
&\leq \sup_{x\in\mathbb{S}^d} \|U_j^\top(\nabla\psi_j(U_j x) - \nabla\varphi_j(U_j x))\|_2 = \sup_{x\in\mathbb{S}^d} \|\nabla\psi_j(U_j x) - \nabla\varphi_j(U_j x)\|_2 \\
&\leq \sup_{\|y\|_2\leq 1} \|\nabla\psi_j(y) - \nabla\varphi_j(y)\|_2 \leq C(k)(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)} \log\left(\frac{\beta}{\eta J}\right)
\end{aligned}$$

The first inequality holds because the Riemannian gradient is the orthogonal projection of the Euclidean gradient of the extension, and orthogonal projections are 1-Lipschitz. The following equality holds because U_j has orthonormal rows. The second inequality holds because for all $x \in \mathbb{S}^d$, $\|U_j x\|_2 \le \|x\|_2 = 1$ by the fact that U_j has orthonormal rows, and the third inequality holds by (28).

Thus, for part (i), we have

$$\inf_{f\in\mathcal{B}_{\mathcal{F}_1}} \mathbb{E}_\nu\left[\left\| -\beta\nabla f(x) - \nabla \log\left(\frac{d\nu}{d\tau}(x)\right)\right\|_2\right] \le \inf_{f\in\mathcal{B}_{\mathcal{F}_1}} \sup_{x\in\mathbb{S}^d}\left\|\beta\nabla f(x) - \nabla\log\left(\frac{d\nu}{d\tau}(x)\right)\right\|_2$$
$$\le \sup_{x\in\mathbb{S}^d}\|\nabla\tilde{g}(x)-\nabla g(x)\|_2 \le \sum_{j=1}^{J}\sup_{x\in\mathbb{S}^d}\|\nabla\tilde{g}_j - \nabla g_j\|_2 \le C(k)J(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)}\log\left(\frac{\beta}{\eta J}\right)$$

Plugging this into Theorem 2 and using that $\sup_{f\in\mathcal{B}_{\mathcal{F}_1}}\{\|\partial_i f\|_\infty | 1\le i\le d+1\}\le 1$, we obtain

$$D_{\mathrm{KL}}(\nu||\hat{\nu}) \le \frac{4\sqrt{d+1}(\beta + C_2\sqrt{d+1}+d)}{\sqrt{n}} + 2(\beta+d+1)\sqrt{\frac{(d+1)\log(\frac{d+1}{\delta})}{2n}}$$
$$+ C(k)J(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)}\log\left(\frac{\beta}{\eta J}\right).$$

If we optimize this bound with respect to β as in the proof of Corollary 1, we obtain

$$\frac{4(C_2\sqrt{d+1}+d)}{\sqrt{n}} + 2(d+1)\sqrt{\frac{\log(\frac{d+1}{\delta})}{2n}} + \left(\frac{2B}{k+1}\right)^{\frac{k+1}{k+3}}\left(\frac{A}{\sqrt{n}}\right)^{\frac{2}{k+3}}$$
$$+ B^{\frac{k+1}{k+3}}\left(\frac{A(k+1)}{2\sqrt{n}}\right)^{\frac{2}{k+3}}\log\left(\frac{1}{\eta J}\left(\frac{2B\sqrt{n}}{A(k+1)}\right)^{\frac{k+1}{k+3}}\right),$$

and the optimal β is $\left(2B\sqrt{n}/(A(k+1))\right)^{\frac{k+1}{k+3}}$, where

$$A = 4\sqrt{d+1} + \sqrt{2(d+1)\log((d+1)/\delta)}, \quad B = C(k)J(L+\eta)(\eta J)^{\frac{2}{k+1}}.$$

For part (ii), we plug

$$\inf_{f\in\mathcal{B}_{\mathcal{F}_1}} \mathbb{E}_\nu\left[\left\| -\beta\nabla f(x) - \nabla \log\left(\frac{d\nu}{d\tau}(x)\right)\right\|_2^2\right] \le \inf_{f\in\mathcal{B}_{\mathcal{F}_1}} \sup_{x\in\mathbb{S}^d}\left\|\beta\nabla f(x) - \nabla\log\left(\frac{d\nu}{d\tau}(x)\right)\right\|_2^2$$
$$\le \sup_{x\in\mathbb{S}^d}\|\nabla\tilde{g}(x)-\nabla g(x)\|_2^2 \le \left(\sum_{j=1}^{J}\sup_{x\in\mathbb{S}^d}\|\nabla\tilde{g}_j-\nabla g_j\|_2\right)^2 \le \left(C(k)J(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)}\log\left(\frac{\beta}{\eta J}\right)\right)^2$$

into Theorem 3, and we obtain (using the notation of Theorem 3) that with probability at least $1-\delta$,

$$KSD(\nu,\hat{\nu}) \le \frac{2}{\sqrt{\delta n}}((\beta C_1+d)^2C_2 + 2C_3(\beta C_1+d)+C_4) + C_2\left(C(k)J(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)}\log\left(\frac{\beta}{\eta J}\right)\right)^2$$
$$\le \frac{2}{\sqrt{\delta n}}\left((\beta C_1+d+C_3)C_2 + \left|2C_3(d-C_2)+C_4-\frac{C_3^2}{d^2}C_2\right|\right)^2$$
$$+ C_2\left(C(k)J(L+\eta)\left(\frac{\beta}{\eta J}\right)^{-2/(k+1)}\log\left(\frac{\beta}{\eta J}\right)\right)^2 = \frac{(A\beta+B)^2}{\sqrt{n}} + D^2\beta^{-4/(k+1)}\log\left(\frac{\beta}{\eta J}\right)^2,$$

where A, B, D are defined appropriately. If we set β to minimize $\frac{A\beta+B}{n^{1/4}} + D\beta^{-2/(k+1)}$, we obtain $\beta = \left(\frac{2n^{1/4}D}{A(k+1)}\right)^{\frac{k+1}{k+3}}$, and the right-hand side becomes

$$\left(A^{\frac{2}{k+3}}\left(\frac{2D}{k+1}\right)^{\frac{k+1}{k+3}} n^{-\frac{1}{2(k+3)}} + \frac{B}{n^{1/4}}\right)^2 + D^2\left(\frac{A(k+1)}{2D}\right)^{\frac{4}{k+3}} n^{-\frac{1}{k+3}} \log\left(\frac{1}{\eta J}\left(\frac{2n^{1/4}D}{A(k+1)}\right)^{\frac{k+1}{k+3}}\right)^2.$$

□

B Qualitative convergence results

B.1 $\mathcal{F}_1$ EBMs dynamics

For a (Fréchet-) differentiable functional $F : \mathcal{P}(\mathbb{R}^{d+2}) \to \mathbb{R}$, the Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ of F is the generalization of gradient flows to the metric space $\mathcal{P}(\mathbb{R}^{d+2})$ endowed with the Wasserstein distance $W_2^2(\mu_1, \mu_2) := \inf_{\pi\in\Pi(\mu_1,\mu_2)} \int_{\mathbb{R}^{d+2}\times\mathbb{R}^{d+2}} \|x - y\|_2^2 \, d\pi(x, y)$ [Ambrosio et al., 2008]. One characterization of Wasserstein gradient flows is as the pushforward $\mu_t = (\Phi_t)_{\#}\mu_0$ of the initial measure μ_0 by the evolution operator Φ_t which maps initial conditions (w_0, θ_0) to the solution at time t of the ODE:

$$\frac{d(w, \theta)}{dt} = -\nabla\left(\frac{\delta}{\delta\mu}F(\mu_t)\right)(w, \theta),$$

where $\frac{\delta}{\delta\mu}F(\mu) : \mathbb{R}^{d+2} \to \mathbb{R}$ is the Fréchet differential or first variation of F at μ.

For any $m > 0$, we define the m-particle gradient flow $t \to u_m(t) = ((w_t^{(i)}, \theta_t^{(i)}))_{i=1}^m$ as the solution of the ODE

$$\frac{d(w_t^{(i)}, \theta_t^{(i)})}{dt} = -\nabla\left(\frac{\delta}{\delta\mu}F(\mu_{m,t})\right)(w_t^{(i)}, \theta_t^{(i)}),$$

where $\mu_{m,t} = \frac{1}{m}\sum_{i=1}^m \delta_{(w_t^{(i)},\theta_t^{(i)})}$. For the functional F defined in (9), we have that $\nabla\left(\frac{\delta}{\delta\mu}F(\mu_{m,t})\right)(w_t^{(i)}, \theta_t^{(i)})$ is equal to $\langle dR(\frac{1}{m}\sum_{j=1}^m \Phi(w_t^{(j)}, \theta_t^{(j)})), \nabla\Phi(w_t^{(i)}, \theta_t^{(i)})\rangle + \lambda(w_t^{(i)}, \theta_t^{(i)})$, which is equal to m times the gradient of the function $G((w^{(i)}, \theta^{(i)})_{i=1}^m) := F(\frac{1}{m}\sum_{j=1}^m \Phi(w^{(j)}, \theta^{(j)}))$ with respect to $(w^{(i)}, \theta^{(i)})$. Thus, $u_m(t)$ is simply the gradient flow of G (up to a time reparametrization).

Theorem 6. *[Chizat and Bach [2018], Thm. 3.3; informal] Let R be a convex differentiable loss defined on a Hilbert space with differential dR Lipschitz on bounded sets and bounded on sublevel sets which satisfies a technical Sard-type regularity assumption. Let $(\mu_t)_{t\geq 0}$ be a Wasserstein gradient flow corresponding to F in (9), such that the support of μ_0 is contained in $B(0, r_b)$ and separates the spheres $r_a\mathbb{S}^{d+1}$ and $r_b\mathbb{S}^{d+1}$ for some $0 < r_a < r_b$. If $(\mu_t)_t$ converges to μ_∞ in W_2, then μ_∞ is a global minimizer of F. Moreover, if $(\mu_{m,t})_{t\geq 0}$ is the empirical measure of $(u_m(t))_{t\geq 0}$ and $\mu_{m,0} \to \mu_0$ weakly, we have $\lim_{t,m\to\infty} F(\mu_{m,t}) = \lim_{m,t\to\infty} F(\mu_{m,t}) = F(\mu_\infty)$.*

Theorem 6 states that when the number m of particles (read neurons) goes to infinity, the function value of the gradient flow of the function $G((w^{(i)}, \theta^{(i)})_{i=1}^m)$ converges to a global optimum of F over $\mathcal{P}(\mathbb{R}^{d+2})$. Remark that Algorithm 1 corresponds to the gradient descent algorithm on $G((w^{(i)}, \theta^{(i)})_{i=1}^m)$ with noisy gradient estimates. Thus, in the small stepsize and exact gradient limits, the iterates of Algorithm 1 approximate the gradient flow of $G((w^{(i)}, \theta^{(i)})_{i=1}^m)$. This reasoning provides an informal justification that Algorithm 1 should have a sensible behavior in the appropriate limits.

Observation 1. *While Theorem 6 assumes that R is defined on a Hilbert space, this assumption is not convenient in our case because $R(f) = \frac{1}{n}\sum_{i=1}^n f(x_i) + \log\left(\int_K e^{-f(x)}dx\right)$ is not well defined on $L^2(\mathbb{R}^d)$, as it involves pointwise evaluations. However, following the argument of Chizat and Bach [2018], up to the technical Sard-type regularity assumption, it suffices to show that $F(\mu)$ is a convex differentiable loss with a first variation $\frac{\delta}{\delta\mu}F(\mu)$ such that*

- *The restriction of $\frac{\delta}{\delta\mu}F(\mu)$ to $(w,\theta) \in \mathbb{S}^{d+1}$ fulfills $\left\|\frac{\delta}{\delta\mu}F(\mu)(\cdot) - \frac{\delta}{\delta\mu}F(\mu')(\cdot)\right\|_{C^1(\mathbb{S}^{d+1})} \leq L\|h_2(\mu) - h_2(\mu')\|_{BL}$, where $h_2 : \mathcal{M}(\mathbb{R}^{d+2}) \to \mathcal{M}(\mathbb{S}^{d+1})$ is defined as $\int_{\mathbb{S}^{d+1}} \varphi(x)\, dh_2(\mu)(x) = \int_{\mathbb{R}^{d+2}} |y|^2\varphi(y/|y|)\, d\mu(y)$ and $\|\cdot\|_{BL}$ is the bounded Lipschitz norm.*
- *The restriction of $\frac{\delta}{\delta\mu}F(\mu)$ to $(w,\theta) \in \mathbb{S}^{d+1}$ is bounded on sublevel sets of $F(\mu)$ in L^∞ norm.*

To apply Theorem 6, we must check that the two statements in Observation 1 hold. Since for the maximum likelihood loss we have:

$$\frac{\delta}{\delta\mu}F(\mu)(w,\theta) = \frac{1}{n}\sum_{i=1}^n \Phi(w,\theta)(x_i) - \frac{\int_K \Phi(w,\theta)(x)\exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu(w',\theta')\right)\,d\tau(x)}{\int_K \exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu(w',\theta')\right)\,d\tau(x)} + \lambda(w^2 + \|\theta\|_2^2),$$

we obtain that for all $(w,\theta) \in \mathbb{R}^{d+2}$ and $\mu, \mu' \in \mathcal{P}(\mathbb{R}^{d+2})$,

$$\begin{aligned}
&\frac{\delta}{\delta\mu}F(\mu)(w,\theta) - \frac{\delta}{\delta\mu}F(\mu')(w,\theta)\\
=& -\frac{\int_K \Phi(w,\theta)(x)\exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu(w',\theta')\right)\,d\tau(x)}{\int_K \exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu(w',\theta')\right)\,d\tau(x)}\\
&+\frac{\int_K \Phi(w,\theta)(x)\exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu'(w',\theta')\right)\,d\tau(x)}{\int_K \exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu'(w',\theta')\right)\,d\tau(x)}\\
=&\frac{\int_K \Phi(w,\theta)(x)\exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu_t(w',\theta')\right)\left(\int_{\mathbb{R}^{d+2}}\Phi(w'',\theta'')(x)d(\mu-\mu')(w'',\theta'')\right)\,d\tau(x)}{\int_K \exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu_t(w',\theta')\right)\,d\tau(x)}\\
&-\frac{\int_K \Phi(w,\theta)(x)\exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu_t(w',\theta')\right)\,d\tau(x)}{\int_K \exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu_t(w',\theta')\right)\,d\tau(x)}\\
&\cdot\frac{\int_K\int_{\mathbb{R}^{d+2}}\Phi(w'',\theta'')(x)d(\mu-\mu')(w'',\theta'')\exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu_t(w',\theta')\right)\,d\tau(x)}{\int_K \exp\left(-\int_{\mathbb{R}^{d+2}}\Phi(w',\theta')(x)d\mu_t(w',\theta')\right)\,d\tau(x)}
\end{aligned}$$

where $\mu_t = t\mu + (1-t)\mu'$. For $(w,\theta) \in \mathbb{S}^{d+1}$ and for all $x \in K$, we have $|\Phi(w,\theta)(x)| = |w\sigma(\langle\theta,x\rangle)| \leq \mathrm{diam}(K)/2$ and $\|\nabla_{(w,\theta)}\Phi(w,\theta)(x)\|_2 = \|(\sigma(\langle\theta,x\rangle), w\mathbb{1}(\langle\theta,x\rangle))\|_2 \leq \sqrt{\mathrm{diam}(K)^2+1}$, which means that $\|\Phi(\cdot)(x)\|_{C_1} \leq \sqrt{\mathrm{diam}(K)^2+1}$. Moreover, for $x \in K$, $\int_{\mathbb{R}^{d+2}}\Phi(w'',\theta'')(x)d(\mu-\mu')(w'',\theta'') = \int_{\mathbb{S}^{d+1}}\Phi(w'',\theta'')(x)d(h_2(\mu) - h_2(\mu'))(w'',\theta'') \leq \mathrm{diam}(K)\|h_2(\mu) - h_2(\mu')\|_{BL}/2$. Hence,

$$\begin{aligned}
\left\|\frac{\delta}{\delta\mu}F(\mu)(\cdot) - \frac{\delta}{\delta\mu}F(\mu')(\cdot)\right\|_{C^1(\mathbb{S}^{d+1})} &\leq 2\max_{x\in K}\|\Phi(\cdot)(x)\|_{C_1(\mathbb{S}^{d+1})}\left|\int_{\mathbb{R}^{d+2}}\Phi(w'',\theta'')(x)d(\mu-\mu')(w'',\theta'')\right|\\
&\leq \sqrt{\mathrm{diam}(K)^2+1}\mathrm{diam}(K)\|h_2(\mu) - h_2(\mu')\|_{BL}
\end{aligned}$$

This shows the first point in Observation 1. For the second point, we have the following bound:

$$\left\| \frac{\delta}{\delta \mu} F(\mu)(w,\theta) \right\|_\infty \leq 2 \sup_{(w,\theta)\in\mathbb{S}^{d+1}} |\Phi(w,\theta)(x)| + \lambda$$

B.2 $\mathcal{F}_2$ EBMs dynamics

$\mathcal{F}_2$ is an RKHS with kernel $k(x,y) = \int_{\mathbb{S}^d} \sigma(\langle x,\theta\rangle)\sigma(\langle y,\theta\rangle)dp(\theta)$ and for the ReLU unit this kernel has a closed-form expression [Cho and Saul, 2009]. Thus, one approach to optimize EBMs with energies over $\mathcal{F}_2$-balls is to apply the representer theorem and to write an optimizer $f \in \mathcal{F}_2$ as $f(\cdot) = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$ for some $\alpha \in \mathbb{R}^n$, as well as $\|f\|^2_{\mathcal{F}_2} = \sum_{i=1}^n \alpha_i\alpha_j k(x_i,x_j)$. Then, f becomes a finite-dimensional linear function of α, and thus any loss F that is convex in f is also convex on α. However, this approach scales quadratically with the number of samples and in practical terms, it is quite far from the way neural networks are typically trained.

The approach that we use to optimize EBMs over $\mathcal{F}_2$-balls is to sample random features $(\theta_i)_{i=1}^m$ on $\mathbb{S}^d$ from a probability measure with density $q(\cdot)$ and consider an approximate kernel $k_m(x,y) = \frac{1}{m}\sum_{i=1}^m \frac{1}{q(\theta_i)}\sigma(\langle x,\theta_i\rangle)\sigma(\langle y,\theta_i\rangle)$ [Rahimi and Recht, 2008, Bach, 2017b]. The functions in the finite dimensional RKHS $\mathcal{H}_m$ with kernel k_m are of the form $h(x) = \sum_{i=1}^m v_i(q(\theta_i)m)^{-1/2}\sigma(\langle x,\theta_i\rangle)$ with norm $\|h\|_{\mathcal{H}_m} = \|v\|_2$, or through a change of variables, $h(x) = \frac{1}{m}\sum_{i=1}^m w_i\sigma(\langle x,\theta_i\rangle)$ with norm $\|h\|_{\mathcal{H}_m} = \|(w_i\sqrt{q(\theta_i)})_{i=1}^m\|_2/\sqrt{m}$.

Thus, learning a distribution with log-densities restricted in a ball of $\mathcal{H}_m$ reduces to learning the outer layer weights $(w_i)_{i=1}^n$. Namely, for R as in Subsec. 5.1, we optimize the loss

$$G((w_i)_{i=1}^m) := R\left(\frac{1}{m}\sum_{i=1}^m w_i\sigma(\langle\theta_i,\cdot\rangle)\right) + \frac{\lambda}{m}\sum_{i=1}^m w_i^2 q(\theta_i),$$

which is convex. The gradient flow for G (with scaled gradient $m\nabla_i G((w_j)_{j=1}^m)$) is

$$\frac{dw_i}{dt} = \left\langle dR\left(\frac{1}{m}\sum_{j=1}^m w_j\sigma(\langle\theta_j,\cdot\rangle)\right), \sigma(\langle\theta_i,\cdot\rangle)\right\rangle + 2\lambda w_i q(\theta_i), \tag{29}$$

and we can approximate it by gradient descent, which converges to the optimum $(w_i^\star)_{i=1}^m$ if the gradients are exact and the stepsize is well chosen.

The connection between learning in $\mathcal{H}_m$ balls and learning in $\mathcal{F}_2$ balls is not straightforward. Applying Proposition 2 of Bach [2017b] and making use of the eigenvalue decay of the $\mathcal{F}_2$ kernel on $\mathbb{S}^d$ [Bach, 2017a], for an appropriate choice of q we have that for all $f \in \mathcal{B}_{\mathcal{F}_2}$, there exists $\hat{f} \in \mathcal{H}_m$ with $\|\hat{f}\|_{\mathcal{H}_m} \leq 2$ such that $\|f - \hat{f}\|_{L^2(p)} \leq O\left((m/\log(m))^{-(d+3)/2}\right)$. This L^2 error bound is sufficient to produce a quantitative result for least squares regression. However, for the three losses considered in this paper we would need bounds for $\|f-\hat{f}\|_\infty$ and $\|\nabla f - \nabla\hat{f}\|_\infty$, which do not seem to be available (Bach [2017b] does provide a bound on $\|f-\hat{f}\|_\infty$, but under the assumption that kernel eigenfunctions have a common L^∞ norm bound, which does not hold for spherical harmonics in $\mathbb{S}^d$).

Nonetheless, a mean-field qualitative approach analogous to the $\mathcal{F}_1$ case is still possible (see Proposition 2.6 of Chizat and Bach [2018]). The learning objective in $\mathcal{F}_2$ can be written as

$$F(h) := R\left(\int_{\mathbb{S}^d}\sigma(\langle\theta,x\rangle)h(\theta)d\tilde{\tau}(\theta)\right) + \lambda\int_{\mathbb{S}^d} h^2(\theta)d\tilde{\tau}(\theta),$$

and the mean-field dynamics is

$$\frac{dh_t(\theta)}{dt} = -\left\langle dR\left(\int_{\mathbb{S}^d}\sigma(\langle\theta',\cdot\rangle)h_t(\theta')d\tilde{\tau}(\theta')\right), \sigma(\langle\theta,\cdot\rangle)\right\rangle - 2\lambda h_t(\theta) \tag{30}$$

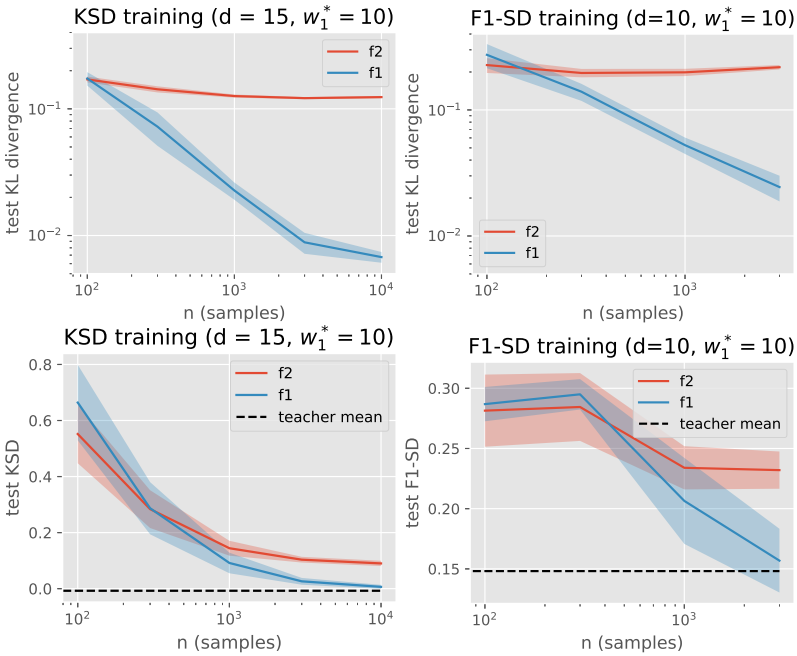

Figure 6: (Left column) Test cross entropy and test KSD for model trained with KSD at different training sample sizes n, in $d = 15$. The teacher model has one neuron of positive weight $w_1^* = 10$ and random position in the hypersphere. (Right column) Test cross entropy and test $\mathcal{F}_1$-SD for model trained with $\mathcal{F}_1$-SD at different training sample sizes n, in $d = 10$. The teacher model has one neuron of positive weight $w_1^* = 10$ and random position in the hypersphere.

If we choose $q(\cdot) = 1$, we have that (29) is the m-particle approximation of (30). Let $h^\star$ be the global minimizer of F, which is reached at a linear rate by (30) because F is strongly convex. Skipping through the details, the argument of Lemma C.15 of Chizat and Bach [2018] could be adapted to yield:

$$\lim_{t,m\to\infty} G((w_{t,i})_{i=1}^m) = \lim_{m,t\to\infty} G((w_{t,i})_{i=1}^m) = \lim_{m\to\infty} G((w_i^\star)_{i=1}^m) = F(h^\star).$$

C Additional experiments

In this section, we show plots corresponding to additional experiments. Figure 6 shows results for KSD and $\mathcal{F}_1$-SD training in the case $J = 1, w_1^* = 10$. Compared to the plots for $J = 1, w_1^* = 2$ shown in Figure 1, the separation between the $\mathcal{F}_1$ and $\mathcal{F}_2$ EBMs becomes much more apparent. Figure 7 shows results for KSD and $\mathcal{F}_1$-SD training in the case $J = 2, w_1^* = -2.5$. The separation between the $\mathcal{F}_1$ and $\mathcal{F}_2$ EBMs is smaller than in the case $J = 2, w_1^* = -5$ shown in Figure 2.

D Duality theory for $\mathcal{F}_1$ and $\mathcal{F}_2$ MLE EBMs

In this section we present the dual problems of $\min_{f\in\mathcal{F}} H(\nu_n, \nu_f)$ (i.e. problem (1)) for the cases $\mathcal{F} = \mathcal{F}_1, \mathcal{F}_2$ (Subsec. D.1), $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(\beta)$ (Subsec. D.2) and $\mathcal{F} = \mathcal{B}_{\mathcal{F}_2}(\beta)$ (Subsec. D.3). The dual problems take the form of entropy maximization under hard constraints, L^∞ and L^2 moment penalizations, respectively. The tools used involve a generalized minimax theorem and Fenchel duality, which was also used for results of the same flavor in finite dimensions (c.f. Mohri et al. [2012]). The proofs are in App. E.

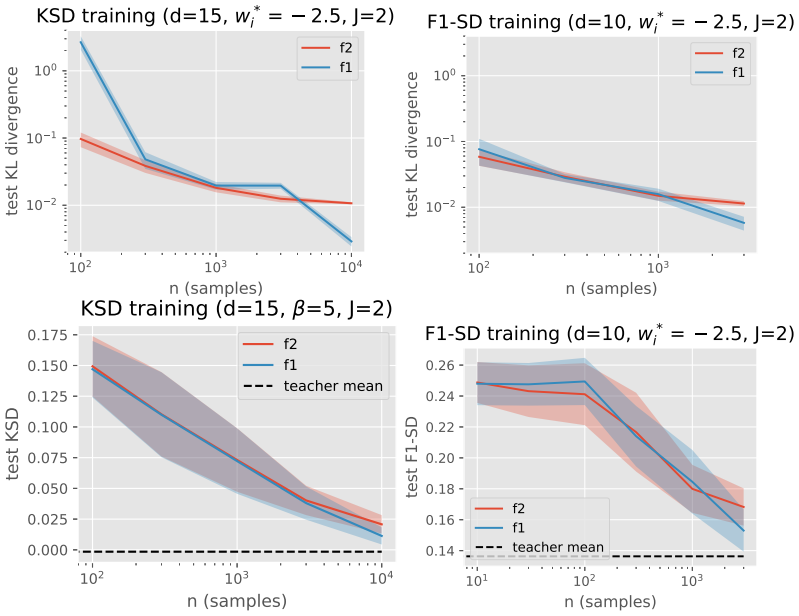

Figure 7: (Left column) Test cross entropy and test KSD for model trained with KSD at different training sample sizes n, in $d = 15$. The teacher model has two neurons of negative weights $w_1^*, w_2^* = -2.5$ and random positions in the hypersphere. (Right column) Test cross entropy and test $\mathcal{F}_1$-SD for model trained with $\mathcal{F}_1$-SD at different training sample sizes n, in $d = 10$. The teacher model has two neurons of negative weights $w_1^*, w_2^* = -2.5$ and random positions in the hypersphere.

D.1 Duality for the unconstrained problem

Consider the following entropy maximization problem under generalized moment constraints:

$$\begin{aligned} \min_{\nu \in \mathcal{P}(K)} \quad & \beta^{-1} D_{KL}(\nu||\tau) \\ \text{s.t.} \quad & \forall \theta \in \mathbb{S}^d, \ \int \sigma(\langle \theta, x \rangle) \ d\nu(x) = \frac{1}{n} \sum_{i=1}^n \sigma(\langle \theta, x_i \rangle), \end{aligned} \tag{31}$$

recalling that τ is the uniform probability measure over K and letting $\beta > 0$ be arbitrary. The constraints in this problem can be interpreted either (i) as an equality constraint in $\mathcal{C}(\mathbb{S}^d)$, i.e., the set of continuous functions on $\mathbb{S}^d$, or (ii) as an equality constraint in $L^2(\mathbb{S}^d)$, i.e., the set of square-integrable functions on $\mathbb{S}^d$. Each interpretation yields a different dual problem.

By the Riesz-Markov-Kakutani representation theorem, the set of signed Radon measures $\mathcal{M}(\mathbb{S}^d)$ can be seen as the continuous dual of $\mathcal{C}(\mathbb{S}^d)$. Hence, in the case (i), the Lagrangian for problem (31) is $L_1 : \mathcal{M}(K) \times \mathcal{M}(\mathbb{S}^d) \times \mathcal{C}(K) \times \mathbb{R} \to \mathbb{R}$ defined as $L_1(\nu, \mu, g, \lambda) = \int \log\left(\frac{d\nu}{d\tau}(x)\right) d\nu(x) + \int \left(\int \sigma(\langle \theta, x \rangle) \ d\nu(x) - \frac{1}{n} \sum_{i=1}^n \sigma(\langle \theta, x_i \rangle) \right) \ d\mu(\theta) - \int g(x) d\nu(x) + \lambda \left(\int d\nu(x) - 1 \right)$, and the dual problem is

$$\sup_{\mu \in \mathcal{M}(\mathbb{S}^d)} \ -\frac{1}{n} \sum_{i=1}^n \int \sigma(\langle \theta, x_i \rangle) \ d\mu(\theta) - \frac{1}{\beta} \log \left(\int \exp\left(-\beta \int \sigma(\langle \theta, x \rangle) \ d\mu(\theta) \right) d\tau(x) \right) \tag{32}$$

which is equivalent to the MLE problem (1) when $\mathcal{F} = \mathcal{F}_1$.

Let $\tilde{\tau}$ to denote the uniform probability measure over $\mathbb{S}^d$. In the case (ii), the Lagrangian for problem (31) is $L_2 : \mathcal{M}(K) \times L^2(\mathbb{S}^d) \times \mathcal{C}(K) \times \mathbb{R} \to \mathbb{R}$ defined as $L_2(\nu, h, g, \lambda) = \int \log\left(\frac{d\nu}{d\tau}(x)\right) d\nu(x) + \int \left(\int \sigma(\langle \theta, x \rangle) \ d\nu(x) - \frac{1}{n} \sum_{i=1}^n \sigma(\langle \theta, x_i \rangle) \right) h(\theta) \ d\tilde{\tau}(\theta) - \int g(x) d\nu(x) + \lambda \left(\int d\nu(x) - 1 \right)$, and the dual problem is

$$\sup_{h \in L^2(\mathbb{S}^d)} \ -\frac{1}{n} \sum_{i=1}^n \int \sigma(\langle \theta, x_i \rangle) h(\theta) \ d\tilde{\tau}(\theta) - \frac{1}{\beta} \log \left(\int \exp\left(-\beta \int \sigma(\langle \theta, x \rangle) h(\theta) \ d\tilde{\tau}(\theta) \right) d\tau(x) \right) \tag{33}$$

which is equivalent to the MLE problem (1) when $\mathcal{F} = \mathcal{F}_2$.

The following theorem shows that problems (31)-(32)-(33) have the same optimal value.

Theorem 7. *Strong duality holds between the entropy maximization problem (31) and each of the two dual problems (32)-(33).*

D.2 Duality for the $\mathcal{F}_1$-ball constrained problem

Using $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ to denote the empirical measure, consider the following problem, which can be seen as an L^∞-penalized version of (31):

$$\min_{\nu\in\mathcal{P}(K)} \beta^{-1} D_{KL}(\nu||\tau) + \max_{\theta\in\mathbb{S}^d} \left| \int \sigma(\langle\theta, x\rangle)\ d(\nu - \nu_n)(x) \right| \tag{34}$$

As shown in Theorem 8, the dual of this problem is a modified version of (32) in which μ is constrained to have TV norm bounded by 1:

$$\max_{\substack{\mu\in\mathcal{M}(\mathbb{S}^d)\\ |\mu|_{\mathrm{TV}}\le 1}} -\frac{1}{n}\sum_{i=1}^n \int \sigma(\langle\theta, x_i\rangle)\ d\mu(\theta) - \frac{1}{\beta}\log\left(\int \exp\left(-\beta\int \sigma(\langle\theta, x\rangle)\ d\mu(\theta)\right) d\tau(x)\right) \tag{35}$$

Remark that by the definition of $\mathcal{F}_1$, the problem (35) is equivalent to MLE problem (1) in the case $\mathcal{F} = \mathcal{B}_{\mathcal{F}_1}(\beta)$.

Theorem 8. *The problem (35) is the dual of the problem (34), and strong duality holds. Moreover, the solution $\nu^\star$ of (35) is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau}(x) = \frac{1}{Z_\beta}\exp\left(-\beta\int \sigma(\langle\theta, x\rangle)\ d\mu^\star(\theta)\right),$$

where $\mu^\star$ is a solution of (34) and Z_β is a normalization constant.

D.3 Duality for the $\mathcal{F}_2$-ball constrained problem

The following problem can be seen as an L^2-penalized version of (31):

$$\min_{\nu\in\mathcal{P}(K)} \beta^{-1} D_{KL}(\nu||\tau) + \left(\int_{\mathbb{S}^d}\left(\int \sigma(\langle\theta, x\rangle)\ d(\nu - \nu_n)(x)\right)^2 d\tilde{\tau}(\theta)\right)^{1/2} \tag{36}$$

And as shown in Theorem 9, the dual of this problem is a modified version of (33) in which h is constrained to have L^2 norm bounded by 1:

$$\max_{\substack{h\in L^2(\mathbb{S}^d)\\ \|h\|_{L^2}\le 1}} -\frac{1}{n}\sum_{i=1}^n \int \sigma(\langle\theta, x_i\rangle)h(\theta)\ d\tilde{\tau}(\theta) - \frac{1}{\beta}\log\left(\int \exp\left(-\beta\int \sigma(\langle\theta, x\rangle)h(\theta)\ d\tilde{\tau}(\theta)\right) d\tau(x)\right) \tag{37}$$

Remark that by the definition of $\mathcal{F}_2$, the problem (35) is equivalent to MLE problem (1) in the case $\mathcal{F} = \mathcal{B}_{\mathcal{F}_2}(\beta)$.

Theorem 9. *The problem (37) is the dual of the problem (36), and strong duality holds. Moreover, the solution $\nu^\star$ of (37) is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau}(x) = \frac{1}{Z_\beta}\exp\left(-\beta\int \sigma(\langle\theta, x\rangle)\ h^\star(\theta)d\tilde{\tau}(\theta)\right),$$

where $h^\star$ is a solution of (36) and Z_β is a normalization constant.

E Proofs of App. D

Theorem 10. *[Kneser [1952]] Let X be a non-empty compact convex subset of a locally convex topological vector space space E and Y a non-empty convex subset of a locally convex topological vector space space F. Let the function $f : X \times Y \to \mathbb{R}$ be such that:*

(i) For each $y \in Y$, the function $x \mapsto f(x, y)$ is upper semicontinuous and concave,

(ii) For each $x \in X$, the function $y \mapsto f(x, y)$ is convex.

Then we have

$$\sup_{x\in X} \inf_{y\in Y} f(x, y) = \inf_{y\in Y} \max_{x\in X} f(x, y).$$

Lemma 9. *The KL divergence $D_{KL}(\nu||\tau) = \int \log\left(\frac{d\nu}{d\tau}\right) d\nu$ is convex and lower semicontinuous in ν.*

Proof. See Theorem 1 of Posner [1975]. □

Observation 2. *Notice that for any functional $f : \mathcal{M}(K) \to \mathbb{R}$, we have*

$$\begin{aligned}\min_{\nu\in\mathcal{P}(K)} f(\nu) &= \min_{\nu\in\mathcal{P}(K)} f(\nu) - \int g(x)d\nu(x) + \lambda\left(\int d\nu(x) - 1\right)\\ &= \min_{\nu\in\mathcal{M}(K)} \sup_{\lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} f(\nu) - \int g(x)d\nu(x) + \lambda\left(\int d\nu(x) - 1\right).\end{aligned}$$

Theorem 7. *Strong duality holds between the entropy maximization problem (31) and each of the two dual problems (32)-(33).*

Proof. We start with (32). First, we prove that it is indeed the dual problem of (31). As stated in the main text, the problem (31) admits a Lagrangian $L_1 : \mathcal{M}(K) \times \mathcal{M}(\mathbb{S}^d) \times \mathcal{C}(K) \times \mathbb{R} \to \mathbb{R}$ defined as

$$\begin{aligned}L_1(\nu, \mu, g, \lambda) = \beta^{-1}\int \log\left(\frac{d\nu}{d\tau}(x)\right) d\nu(x) + \int\left(\int \sigma(\langle\theta, x\rangle)\, d\nu(x) - \frac{1}{n}\sum_{i=1}^{n}\sigma(\langle\theta, x_i\rangle)\right) d\mu(\theta) - \int g(x)d\nu(x)\\ + \lambda\left(\int d\nu(x) - 1\right)\end{aligned}$$

The Lagrange dual function is

$$\begin{aligned}F_1(\mu, g, \lambda) = \inf_{\nu\in\mathcal{M}(K)} L_1(\nu, \mu, g, \lambda) = -\beta^{-1}\int \exp\left(-\beta\left(\int \sigma(\langle\theta, x\rangle)\, d\mu(\theta) + g(x) - \lambda\right) - 1\right) d\tau(x)\\ - \frac{1}{n}\sum_{i=1}^{n}\int \sigma(\langle\theta, x_i\rangle)\, d\mu(\theta) - \lambda,\end{aligned} \tag{38}$$

where we have used that at the optimal ν, the first variation of L_1 w.r.t. ν must be zero:

$$\begin{aligned}0 = \frac{d}{d\nu}L_1(\nu, \mu, g, \lambda) = \beta^{-1}\log\left(\frac{d\nu}{d\tau}(x)\right) + \beta^{-1} + \int \sigma(\langle\theta, x\rangle)\, d\mu(\theta) - g(x) + \lambda,\\ \implies \frac{d\nu}{d\tau}(x) = \exp\left(-\beta\left(\int \sigma(\langle\theta, x\rangle)\, d\mu(\theta) + g(x) - \lambda\right) - 1\right).\end{aligned}$$

The Lagrange dual problem is

$$\begin{aligned}
&\sup_{\mu,\lambda,g\geq 0} F_1(\mu,g,\lambda)\\
&= \sup_{\mu,\lambda,g\geq 0} -\beta^{-1}\int \exp\left(-\beta\left(\int \sigma(\langle\theta,x\rangle)\ d\mu(\theta)+g(x)-\lambda\right)-1\right)\ d\tau(x)-\frac{1}{n}\sum_{i=1}^n\int\sigma(\langle\theta,x_i\rangle)\ d\mu(\theta)-\lambda\\
&= \sup_{\mu,\lambda} -\beta^{-1}\int \exp\left(-\beta\left(\int \sigma(\langle\theta,x\rangle)\ d\mu(\theta)-\lambda\right)-1\right)\ d\tau(x)-\frac{1}{n}\sum_{i=1}^n\int\sigma(\langle\theta,x_i\rangle)\ d\mu(\theta)-\lambda\\
&= \sup_{\mu\in\mathcal{M}(\mathbb{S}^d)} -\frac{1}{n}\sum_{i=1}^n\int\sigma(\langle\theta,x_i\rangle)\ d\mu(\theta)-\frac{1}{\beta}\log\left(\int\exp\left(-\beta\int\sigma(\langle\theta,x\rangle)\ d\mu(\theta)\right)\ d\tau(x)\right),
\end{aligned}\tag{39}$$

and the right-hand side is precisely (32). In the second equality we used that the optimal choice for g is $g=0$. In the third equality we used that the optimal λ must satisfy the first-order optimality condition:

$$\begin{aligned}
&\int \exp\left(-\beta\left(\int \sigma(\langle\theta,x\rangle)\ d\mu(\theta)-\lambda\right)-1\right)\ d\tau(x)-1=0,\\
&\implies e^{\beta\lambda}=\left(\int\exp\left(-\beta\int\sigma(\langle\theta,x\rangle)\ d\mu(\theta)-1\right)\ d\tau(x)\right)^{-1}\\
&\implies \lambda=-\frac{1}{\beta}\log\left(\int\exp\left(-\beta\int\sigma(\langle\theta,x\rangle)\ d\mu(\theta)-1\right)\ d\tau(x)\right)
\end{aligned}$$

To prove strong duality, we need to show that

$$\inf_{\nu\in\mathcal{M}(K)}\sup_{\mu\in\mathcal{M}(\mathbb{S}^d),\lambda\in\mathbb{R},g\in\mathcal{C}(K):g\geq 0} L_1(\nu,\mu,g,\lambda)=\sup_{\mu\in\mathcal{M}(\mathbb{S}^d),\lambda\in\mathbb{R},g\in\mathcal{C}(K):g\geq 0}\inf_{\nu\in\mathcal{M}(K)} L_1(\nu,\mu,g,\lambda).\tag{40}$$

If we define $\tilde{L}_1:\mathcal{P}(K)\times\mathcal{M}(\mathbb{S}^d)\to\mathbb{R}$ as $\tilde{L}_1(\nu,\mu)=L_1(\nu,\mu,0,0)$, we have that the assumptions of Theorem 10 hold for $-\tilde{L}_1$. Indeed, by Lemma 9 we have that $-\tilde{L}_1(\cdot,\mu)$ is a concave and upper semicontinuous function of ν. And by Prokhorov's theorem, $\mathcal{P}(K)$ is a compact subset of the locally convex topological vector space of signed Radon measures with the topology of weak convergence (tightness follows from the fact that K is compact). Thus,

$$\inf_{\nu\in\mathcal{P}(K)}\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\tilde{L}_1(\nu,\mu)=\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\min_{\nu\in\mathcal{P}(K)}\tilde{L}_1(\nu,\mu).\tag{41}$$

On the one hand, notice that by Observation 2,

$$\inf_{\nu\in\mathcal{M}(K)}\sup_{\mu\in\mathcal{M}(\mathbb{S}^d),\lambda\in\mathbb{R},g\in\mathcal{C}(K):g\geq 0} L_1(\nu,\mu,g,\lambda)=\inf_{\nu\in\mathcal{P}(K)}\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)} L_1(\nu,\mu,0,0)=\inf_{\nu\in\mathcal{P}(K)}\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\tilde{L}_1(\nu,\mu)\tag{42}$$

On the other hand,

$$\begin{aligned}
\sup_{\mu\in\mathcal{M}(\mathbb{S}^d),\lambda\in\mathbb{R},g\in\mathcal{C}(K):g\geq 0}\inf_{\nu\in\mathcal{M}(K)} L_1(\nu,\mu,g,\lambda)&=\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\inf_{\nu\in\mathcal{M}(K)}\sup_{\lambda\in\mathbb{R},g\in\mathcal{C}(K):g\geq 0} L_1(\nu,\mu,g,\lambda)\\
&=\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\min_{\nu\in\mathcal{P}(K)} L_1(\nu,\mu,0,0)=\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\min_{\nu\in\mathcal{P}(K)}\tilde{L}_1(\nu,\mu).
\end{aligned}\tag{43}$$

where we have used Lemma 10 in the first equality, Observation 2 in the second equality and the definition of $\tilde{L}_1$ in the third equality. Thus, the strong duality (40) follows from plugging (42) and (43) into (41).

To show that (33) is also a dual problem of (31), we consider the Lagrangian $L_2:\mathcal{M}(K)\times L^2(\mathbb{S}^d)\times\mathcal{C}(K)\times\mathbb{R}\to$

$\mathbb{R}$ defined as

$$L_2(\nu, h, g, \lambda) = \beta^{-1} \int \log\left(\frac{d\nu}{d\tau}(x)\right) d\nu(x) + \int \left(\int \sigma(\langle\theta, x\rangle)\ d\nu(x) - \frac{1}{n}\sum_{i=1}^{n} \sigma(\langle\theta, x_i\rangle)\right) h(\theta)\ d\tilde{\tau}(\theta) - \int g(x) d\nu(x)$$
$$+ \lambda\left(\int d\nu(x) - 1\right) - r\left(\int h^2(\theta)\ d\tilde{\tau}(\theta) - 1\right)$$

The reasoning to obtain the dual problem (33) is analogous. Strong duality in this case can be stated as

$$\inf_{\nu\in\mathcal{M}(K)} \sup_{h\in L^2(\mathbb{S}^d), \lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} L_2(\nu, h, g, \lambda) = \sup_{h\in L^2(\mathbb{S}^d), \lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} \inf_{\nu\in\mathcal{M}(K)} L_1(\nu, h, g, \lambda).$$

Analogously, we define $\tilde{L}_2 : \mathbb{P}(K)\times L^2(\mathbb{S}^d) \to \mathbb{R}$ as $\tilde{L}_2(\nu, h) = L_2(\nu, h, 0, 0)$, and we have that the assumptions of Theorem 10 hold for $-\tilde{L}_2$ as well, implying that $\inf_{\nu\in\mathcal{P}(K)} \sup_{h\in L^2(\mathbb{S}^d)} \tilde{L}_2(\nu, h) = \sup_{h\in L^2(\mathbb{S}^d)} \min_{\nu\in\mathcal{P}(K)} \tilde{L}_2(\nu, h)$. The concluding argument is also analogous. □

Lemma 10. *For all $\mu \in \mathcal{M}(\mathbb{S}^d)$,*

$$\sup_{\lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} \inf_{\nu\in\mathcal{M}(K)} L_1(\nu, \mu, g, \lambda) = \min_{\nu\in\mathcal{M}(K)} \sup_{\lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} L_1(\nu, \mu, g, \lambda).$$

Proof. First, notice that by (38) and (39),

$$\sup_{\lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} \inf_{\nu\in\mathcal{M}(K)} L_1(\nu, \mu, g, \lambda) = \sup_{\lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} F_1(\mu, g, \lambda)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \int \sigma(\langle\theta, x_i\rangle)\ d\mu(\theta) - \frac{1}{\beta}\log\left(\int \exp\left(-\beta \int \sigma(\langle\theta, x\rangle)\ d\mu(\theta)\right)\ d\tau(x)\right), \tag{44}$$

And by Observation 2,

$$\min_{\nu\in\mathcal{M}(K)} \sup_{\lambda\in\mathbb{R}, g\in\mathcal{C}(K): g\geq 0} L_1(\nu, \mu, g, \lambda) = \min_{\nu\in\mathcal{P}(K)} L_1(\nu, \mu, 0, 0) = \min_{\nu\in\mathcal{P}(K)} L_1(\nu, \mu, 0, 0) \tag{45}$$

If $\nu^\star \in \mathcal{P}(K)$ is a minimizer of $\min_{\nu\in\mathcal{P}(K)} L_1(\nu, \mu, 0, 0)$, it must fulfill

$$\exists C \in \mathbb{R}: \quad C = \frac{dL_1}{d\nu}(\nu^\star, \mu, 0, 0) = \beta^{-1}\log\left(\frac{d\nu}{d\tau}(x)\right) + \beta^{-1} + \int \sigma(\langle\theta, x\rangle)\ d\mu(\theta),$$
$$\implies \frac{d\nu^\star}{d\tau(x)}(x) = \exp\left(-\beta\int \sigma(\langle\theta, x\rangle)\ d\mu(\theta) + \beta C - 1\right),$$

where $-\beta C + 1 = \log\left(\int_K \exp\left(-\beta\int \sigma(\langle\theta, x\rangle)\ d\mu(\theta)\right) d\tau(x)\right)$. Hence,

$$L_1(\nu^\star, \mu, 0, 0) = \beta^{-1}\int \log\left(\frac{d\nu^\star}{d\tau(x)}(x)\right) d\nu^\star(x) + \int \left(\int \sigma(\langle\theta, x\rangle)\ d\nu^\star(x) - \frac{1}{n}\sum_{i=1}^{n} \sigma(\langle\theta, x_i\rangle)\right)\ d\mu(\theta)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \int \sigma(\langle\theta, x_i\rangle)\ d\mu(\theta) + \int \left(C - \beta^{-1}\right)\ d\nu^\star(x)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \int \sigma(\langle\theta, x_i\rangle)\ d\mu(\theta) - \frac{1}{\beta}\log\left(\int_K \exp\left(-\beta\int \sigma(\langle\theta, x\rangle)\ d\mu(\theta)\right) d\tau(x)\right)$$

If we plug this into the right-hand side of (45), we obtain exactly the right-hand side of (44), concluding the proof. □

Theorem 11. *[Fenchel strong duality; Borwein and Zhu [2005], pp. 135-137] Let X and Y be Banach spaces, $f : X \to \mathbb{R} \cup \{+\infty\}$ and $g : Y \to \mathbb{R} \cup \{+\infty\}$ be convex functions and $A : X \to Y$ be a bounded linear map. Define the Fenchel problems:*

$$p^* = \inf_{x \in X} \{f(x) + g(Ax)\}$$
$$d^* = \sup_{y^* \in Y^*} \{-f^*(A^* y^*) - g^*(-y^*)\},$$

where $f^*(x^*) = \sup_{x \in X}\{\langle x, x^* \rangle - f(x)\}$, $g^*(y^*) = \sup_{y \in Y}\{\langle y, y^* \rangle - g(y)\}$ are the convex conjugates of f, g respectively, and $A^* : Y^* \to X^*$ is the adjoint operator. Then, $p^* \geq d^*$. Moreover if f, g, and A satisfy either

1. *f and g are lower semi-continuous and $0 \in \text{core}(\text{dom}\, g - A\, \text{dom}\, f)$ where* core *is the algebraic interior and* dom h*, where h is some function, is the set $\{z : h(z) < +\infty\}$,*
2. *or $A\, \text{dom}\, f \cap \text{cont}\, g \neq \emptyset$ where* cont *are is the set of points where the function is continuous.*

Then strong duality holds, i.e. $p^* = d^*$. If $d^* \in \mathbb{R}$ then supremum is attained.

Theorem 8. *The problem (35) is the dual of the problem (34), and strong duality holds. Moreover, the solution $\nu^\star$ of (35) is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau}(x) = \frac{1}{Z_\beta} \exp\left(-\beta \int \sigma(\langle \theta, x \rangle)\, d\mu^\star(\theta)\right),$$

where $\mu^\star$ is a solution of (34) and Z_β is a normalization constant.

Proof. One way to prove Theorem 8 (and Theorem 9) would be to develop an argument based on a modification of the Lagrangian function L_1 (resp. L_2) that encodes the $\mathcal{F}_1$ restriction (resp. $\mathcal{F}_2$), and to reduce the problem once again to a min-max duality result like Theorem 10. However, this method turns out to be rather cumbersome, and we resort to an alternative approach that harnesses the power of Fenchel duality theory and yields a much faster proof. In fact, our proof structure is similar to Theorem 12.2 of Mohri et al. [2012], which focuses on the finite-dimensional case and deals with a slightly different problem. As shown by Theorem 11, the Fenchel strong duality sufficient conditions are very similar in the Euclidean and in the Banach space settings.

We will use Theorem 11 with $X = \mathcal{M}(K)$, i.e. the Banach space of signed Radon measures, and $Y = \mathcal{C}(\mathbb{S}^d)$, the Banach space of continuous functions on $\mathbb{S}^d$. Define $f : \mathcal{M}(K) \to \mathbb{R} \cup \{+\infty\}$ as

$$f(\nu) = \begin{cases} \beta^{-1} D_{KL}(\nu || \tau) & \text{if } \nu \in \mathcal{P}(K), \\ +\infty & \text{otherwise} \end{cases}$$

Define $g : \mathcal{C}(\mathbb{S}^d) \to \mathbb{R} \cup \{+\infty\}$ as

$$g(\varphi) = \max_{\theta \in \mathbb{S}^d} \left| \varphi(\theta) - \int_K \sigma(\langle \theta, x \rangle)\, d\nu_n(x) \right|,$$

and $A : \mathcal{M}(K) \to \mathcal{C}(\mathbb{S}^d)$ as $(A\nu)(\theta) = \int_K \sigma(\langle \theta, x \rangle) d\nu(x)$. Remark that A is a bounded linear operator, which implies that it has an adjoint operator. By the Riesz-Markov-Kakutani representation theorem, we have that $\mathcal{C}(\mathbb{S}^d)^* = \mathcal{M}(\mathbb{S}^d)$, which means that the adjoint of A is of the form $A^* : \mathcal{M}(\mathbb{S}^d) \to \mathcal{M}(K)^*$. By the definition of the adjoint operator, we have that for any $\mu \in \mathcal{M}(\mathbb{S}^d), \nu \in \mathcal{M}(K)$,

$$\langle A^* \mu, \nu \rangle_{\mathcal{M}(K)^*, \mathcal{M}(K)} = \langle \mu, A\nu \rangle_{\mathcal{M}(\mathbb{S}^d), \mathcal{C}(\mathbb{S}^d)} = \int_{\mathbb{S}^d} \int_K \sigma(\langle \theta, x \rangle)\, d\nu(x) d\mu(\theta) = \int_K \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle)\, d\mu(\theta) d\nu(x) \quad (46)$$

Notice that $\mathcal{C}(K) \subseteq \mathcal{M}(K)^*$ by the fact that a vector space has a natural embedding in its continuous bidual (but the continuous bidual is in general larger). Through this identification, (46) implies that we can write $A^*\mu(x) = \int_{\mathbb{S}^d} \sigma(\langle \theta, x \rangle) d\mu(\theta)$.

Our goal now is to compute the convex conjugates f^* and g^*. By the argument of Lemma B.37 of Mohri et al. [2012], which works in the infinite-dimensional case as well, the convex conjugate $f^* : \mathcal{C}(K) \to \mathbb{R} \cup \{+\infty\}$ is shown to be:

$$f^*(\psi) = \frac{1}{\beta} \log \left(\int_K \exp\left(\beta \psi(x)\right) d\tau(x) \right)$$

Remark that f^* has domain $\mathcal{M}(K)^*$, which is larger than $\mathcal{C}(K)$. However, knowing the restriction of f^* to $\mathcal{C}(K)$ will suffice for our purposes.

Moreover, $g^* : \mathcal{M}(\mathbb{S}^d) \to \mathbb{R} \cup \{+\infty\}$ fulfills:

$$\begin{aligned}
g^*(\mu) &= \sup_{\varphi \in \mathcal{C}(\mathbb{S}^d)} \left\{ \int_{\mathbb{S}^d} \varphi \, d\mu - \max_{\theta \in \mathbb{S}^d} \left| \varphi(\theta) - \int_K \sigma(\langle \theta, x \rangle) \, d\nu_n(x) \right| \right\} \\
&= \sup_{\varphi \in \mathcal{C}(\mathbb{S}^d)} \left\{ \int_{\mathbb{S}^d} \varphi \, d\mu - \sup_{\substack{\mu' \in \mathcal{M}(\mathbb{S}^d), \\ |\mu'|_{TV} \leq 1}} \int_{\mathbb{S}^d} \left(\varphi(\theta) - \int_K \sigma(\langle \theta, x \rangle) \, d\nu_n(x) \right) d\mu'(\theta) \right\} \\
&= \sup_{\varphi \in \mathcal{C}(\mathbb{S}^d)} \inf_{\substack{\mu' \in \mathcal{M}(\mathbb{S}^d), \\ |\mu'|_{TV} \leq 1}} \left\{ \int_{\mathbb{S}^d} \varphi(\theta) \, d(\mu - \mu')(\theta) + \int_{\mathbb{S}^d} \int_K \sigma(\langle \theta, x \rangle) \, d\nu_n(x) d\mu'(\theta) \right\} \\
&= \inf_{\substack{\mu' \in \mathcal{M}(\mathbb{S}^d), \\ |\mu'|_{TV} \leq 1}} \sup_{\varphi \in \mathcal{C}(\mathbb{S}^d)} \left\{ \int_{\mathbb{S}^d} \varphi(\theta) \, d(\mu - \mu')(\theta) + \int_{\mathbb{S}^d} \int_K \sigma(\langle \theta, x \rangle) \, d\nu_n(x) d\mu'(\theta) \right\} \\
&= \begin{cases} \int_{\mathbb{S}^d} \int_K \sigma(\langle \theta, x \rangle) \, d\nu_n(x) d\mu(\theta) & \text{if } |\mu|_{TV} \leq 1 \\ +\infty & \text{otherwise} \end{cases}
\end{aligned}$$

In the first equality we have used the definition of g, in the fourth equality we have used Theorem 10 (remark that $\{\mu' \in \mathcal{M}(\mathbb{S}^d) : |\mu'|_{TV} \leq 1\}$ is compact in the weak convergence topology), and in the fifth equality we have used that $\sup_{\varphi \in \mathcal{C}(\mathbb{S}^d)} \left\{ \int_{\mathbb{S}^d} \varphi(\theta) \, d(\mu - \mu')(\theta) \right\} = +\infty$ unless $\mu = \mu'$.

With these definitions, notice that problem (34) can be rewritten as $\inf_{\nu \in \mathcal{M}(K)} \{f(\nu) + g(A\nu)\}$ and problem (35) can be rewritten as $\sup_{\mu \in \mathcal{M}(\mathbb{S}^d)} \{-f^*(-A^*\mu) - g^*(\mu)\}$. Thus, strong duality between (34) and (35) follows from Fenchel strong duality, which holds by checking condition 2 of Theorem 11. We have to see that $A \operatorname{dom} f \cap \operatorname{cont} g \neq \emptyset$. Consider $\varphi(\cdot) = \int_K \sigma(\langle \cdot, x \rangle) \, d\nu(x) \in \mathcal{C}(\mathbb{S}^d)$ for some $\nu \in \mathcal{P}(K)$ absolutely continuous w.r.t. τ. Then, we have that $\varphi \in A \operatorname{dom} f$. Moreover, since g is a continuous function (in the supremum norm topology), $\operatorname{cont} g = \mathcal{C}(\mathbb{S}^d)$ and hence $\varphi \in \operatorname{cont} g$ as well, which means that the intersection is not empty.

Notice that in our case $d^* = \sup_{\mu \in \mathcal{M}(\mathbb{S}^d)} \{-f^*(-A^*\mu) - g^*(\mu)\} \in \mathbb{R}$, which by Theorem 11 implies that the supremum is attained: let $\mu^\star$ be one maximizer. We show that $p^* = \inf_{\nu \in \mathcal{M}(K)} \{f(\nu) + g(A\nu)\} = \inf_{\nu \in \mathcal{P}(K)} \{f(\nu) + g(A\nu)\}$ admits a minimizer by the direct method of the calculus of variations: notice that f and $g \circ A$ are lower semicontinuous in the topology of weak convergence (f by Lemma 9 and $g \circ A$ because it is a maximum of continuous functions, and thus its sublevel sets are closed because they are the intersection of closed sublevel sets), and $\mathcal{P}(K)$ is compact.

We now show that $\frac{d\nu^\star}{d\tau}(x) = \frac{1}{Z_\beta} \exp\left(-\beta \int \sigma(\langle \theta, x \rangle) \, d\mu^\star(\theta)\right)$, where $\nu^\star$ and $\mu^\star$ are solutions of (34) and (35), respectively, that we know exist by the previous paragraph. Recall the argument to prove Fenchel weak

duality:

$$\begin{aligned}\sup_{\mu\in\mathcal{M}(\mathbb{S}^d)}\{-f^*(-A^*\mu)-g^*(\mu)\} &= -f^*(-A^*\mu^\star)-g^*(\mu^\star)\\ &= -\sup_{\nu\in\mathcal{M}(K)}\left\{\langle -A^*\mu^\star,\nu\rangle - f(\nu)\right\} - \sup_{\varphi\in\mathcal{C}(\mathbb{S}^d)}\left\{\langle \mu^\star,\varphi\rangle - g(\varphi)\right\}\\ &\le -\sup_{\nu\in\mathcal{M}(K)}\left\{\langle -A^*\mu^\star,\nu\rangle - f(\nu)+\langle\mu^\star,A\nu\rangle - g(A\nu)\right\}\\ &= -\sup_{\nu\in\mathcal{M}(K)}\left\{\langle -f(\nu)-g(A\nu)\right\} = \inf_{\nu\in\mathcal{M}(K)}\{f(\nu)+g(A\nu)\} = f(\nu^\star)+g(A\nu^\star)\end{aligned}$$

Thus, for strong duality to hold we must have that $\nu^\star = \operatorname{argmax}_{\nu\in\mathcal{M}(K)}\left\{\langle -A^*\mu^\star,\nu\rangle - f(\nu)\right\}$, and the corresponding Euler-Lagrange condition is $\frac{d\nu^\star}{d\tau}(x) = \frac{1}{Z_\beta}\exp\left(-\beta\int\sigma(\langle\theta,x\rangle)\,d\mu^\star(\theta)\right)$. □

Theorem 9. *The problem (37) is the dual of the problem (36), and strong duality holds. Moreover, the solution $\nu^\star$ of (37) is unique and its density satisfies*

$$\frac{d\nu^\star}{d\tau}(x) = \frac{1}{Z_\beta}\exp\left(-\beta\int\sigma(\langle\theta,x\rangle)\;h^\star(\theta)d\tilde{\tau}(\theta)\right),$$

where $h^\star$ is a solution of (36) and Z_β is a normalization constant.

Proof. The proof is largely analogous to the proof of Theorem 8. We use Theorem 11 with $X=\mathcal{M}(K)$ as before, and $Y=L^2(\mathbb{S}^d)$, the Hilbert space of square-integrable functions on $\mathbb{S}^d$ under the base measure $\tilde{\tau}$, which is of course self-dual. We define f as before, and $g: L^2(\mathbb{S}^d)\to\mathbb{R}\cup\{+\infty\}$ as

$$g(\varphi) = \left(\int_{\mathbb{S}^d}\left(\varphi(\theta)-\int_K\sigma(\langle\theta,x\rangle)\,d\nu_n(x)\right)^2 d\tilde{\tau}(\theta)\right)^{1/2},$$

and consequently, $g^*: L^2(\mathbb{S}^d)\to\mathbb{R}\cup\{+\infty\}$ fulfills:

$$\begin{aligned}g^*(\psi) &= \sup_{\varphi\in L^2(\mathbb{S}^d)}\left\{\int_{\mathbb{S}^d}\varphi\psi\;d\tilde{\tau} - \left(\int_{\mathbb{S}^d}\left(\varphi(\theta)-\int_K\sigma(\langle\theta,x\rangle)\,d\nu_n(x)\right)^2 d\tilde{\tau}(\theta)\right)^{1/2}\right\}\\ &= \sup_{\varphi\in L^2(\mathbb{S}^d)}\left\{\int_{\mathbb{S}^d}\varphi\psi\;d\tilde{\tau} - \sup_{\substack{\hat{\psi}\in L^2(\mathbb{S}^d),\\ \|\hat{\psi}\|_2\le 1}}\int_{\mathbb{S}^d}\left(\varphi(\theta)-\int_K\sigma(\langle\theta,x\rangle)\,d\nu_n(x)\right)\hat{\psi}(\theta)d\tilde{\tau}(\theta)\right\}\\ &= \sup_{\varphi\in L^2(\mathbb{S}^d)}\inf_{\substack{\hat{\psi}\in L^2(\mathbb{S}^d),\\ \|\hat{\psi}\|_2\le 1}}\left\{\int_{\mathbb{S}^d}\varphi(\theta)(\psi(\theta)-\hat{\psi}(\theta))\;d\tilde{\tau}(\theta)+\int_{\mathbb{S}^d}\int_K\sigma(\langle\theta,x\rangle)\;d\nu_n(x)\;\hat{\psi}(\theta)\;d\tilde{\tau}(\theta)\right\},\end{aligned}$$

and using Theorem 10 once more, this is equal to:

$$\begin{aligned}&\inf_{\substack{\hat{\psi}\in L^2(\mathbb{S}^d),\\ \|\hat{\psi}\|_2\le 1}}\sup_{\varphi\in L^2(\mathbb{S}^d)}\left\{\int_{\mathbb{S}^d}\varphi(\theta)(\psi(\theta)-\hat{\psi}(\theta))\;d\tilde{\tau}(\theta)+\int_{\mathbb{S}^d}\int_K\sigma(\langle\theta,x\rangle)\;d\nu_n(x)\;\hat{\psi}(\theta)\;d\tilde{\tau}(\theta)\right\}\\ &= \begin{cases}\int_{\mathbb{S}^d}\int_K\sigma(\langle\theta,x\rangle)\;d\nu_n(x)\;\psi(\theta)\;d\tilde{\tau}(\theta) & \text{if } \|\psi\|_2\le 1,\\ +\infty & \text{otherwise}\end{cases}\end{aligned}$$

With these definitions, notice that problem (36) can be rewritten as $\inf_{\nu\in\mathcal{M}(K)}\{f(\nu)+g(A\nu)\}$ and problem (37) can be rewritten as $\sup_{\psi\in L^2(\mathbb{S}^d)}\{-f^*(-A^*\psi)-g^*(\psi)\}$. The rest of the proof is analogous. □